UNIVERSITY OF GLASGOW

Time and Frequency Domain Statistical Methods for High-frequency Time Series

by

Amira Sherif Mohamed Elayouty

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
College of Science and Engineering
School of Mathematics and Statistics

March 2017

Declaration of Authorship

I, AMIRA SHERIF MOHAMED ELAYOUTY, declare that this thesis titled, 'Time and Frequency Domain Statistical Methods for High-frequency Time Series' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

The work presented in Chapters 1 and 2 has been published in Environmental and Ecological Statistics with the title "Challenges in modeling detailed and complex environmental data sets: a case study modeling the excess partial pressure of fluvial CO_2" (September 2015). A paper based on part of this work has also been presented at the 29th International Workshop on Statistical Modelling (IWSM) in Gottingen, 2014. Part of the work in Chapters 3 and 4 has been presented at the Biennial Conference of the Research Group for Environmental Statistics of the Italian Statistical Society (GRASPA) in Bari, 2015. Part of the work presented in Chapter 5 has been presented at the 26th Annual Conference of The International Enviornmetrics Society (TIES) in Edinburgh, 2016. A manuscript based on this chapter is currently in preparation.

Signed: ________________________________

Date: ________________________________

“All roads that lead to success have to pass through hard work boulevard at some point.”

Eric Thomas

Abstract

Advances in sensor technology enable environmental monitoring programmes to record and store measurements at high-temporal resolution over long time periods. These large volumes of high-frequency data promote an increasingly comprehensive picture of many environmental processes that would not have been accessible in the past with monthly, fortnightly or even daily sampling. However, benefiting from these increasing amounts of high-frequency data presents various challenges in terms of data processing and statistical modeling using standard methods and software tools. These challenges are attributed to the large volumes of data, the persistent and long memory serial correlation in the data, the signal to noise ratio, and the complex and time-varying dynamics and inter-relationships between the different drivers of the process at different timescales.

This thesis aims at using and developing a variety of statistical methods in both the time and frequency domains to effectively explore and analyze high-frequency time series data as well as to reduce their dimensionality, with specific application to a 3 year hydrological time series. Firstly, the thesis investigates the statistical challenges of exploring, modeling and analyzing these large volumes of high-frequency time series. Thereafter, it uses and develops more advanced statistical techniques to: (i) better visualize and identify the different modes of variability and common patterns in such data, and (ii) provide a more adequate dimension reduction representation to the data, which takes into account the persistent serial dependence structure and non-stationarity in the series. Throughout the thesis, a 15-minute resolution time series of excess partial pressure of carbon dioxide (EpCO_2) obtained for a small catchment in the River Dee in Scotland has been used as an illustrative data set. Understanding the bio-geochemical and hydrological drivers of EpCO_2 is very important to the assessment of the global carbon budget.

Specifically, Chapters 1 and 2 present a range of advanced statistical approaches in both the time and frequency domains, including wavelet analysis and additive models, to visualize and explore temporal variations and relationships between variables for the River Dee data across the different timescales to investigate the statistical challenges posed by such data. In Chapter 3, a functional data analysis approach is employed to identify the common daily patterns of EpCO_2 by means of functional principal component analysis and functional cluster analysis. The techniques used in this chapter assume independent functional data. However, in numerous applications, functional observations are serially correlated over time, e.g. where each curve represents a segment of the whole time interval. In this situation, ignoring the temporal dependence may result in an inappropriate dimension reduction of the data and inefficient inference procedures.

Subsequently, the dynamic functional principal components, recently developed by Hormann et al. (2014), are considered in Chapter 4 to account for the temporal correlation using a frequency domain approach. A specific contribution of this thesis is the extension of the methodology of dynamic functional principal components to temporally dependent functional data estimated using any type of basis functions, not only orthogonal basis functions. Based on the scores of the proposed general version of dynamic functional principal components, a novel clustering approach is proposed and used to cluster the daily curves of $\mathrm{EpCO_2}$ taking into account the dependence structure in the data.

The dynamic functional principal components depend in their construction on the assumption of second-order stationarity, which is not a realistic assumption in most environmental applications. Therefore, in Chapter 5, a second specific contribution of this thesis is the development of a time-varying dynamic functional principal components which allows the components to vary smoothly over time. The performance of these smooth dynamic functional principal components is evaluated empirically using the $\mathrm{EpCO_2}$ data and using a simulation study. The simulation study compares the performance of smooth and original dynamic functional principal components under both stationary and non-stationary conditions. The smooth dynamic functional principal components have shown considerable improvement in representing non-stationary dependent functional data in smaller dimensions.

Using a bootstrap inference procedure, the smooth dynamic functional principal components have been subsequently employed to investigate whether or not the spectral density and covariance structure of the functional time series under study change over time. To account for the possible changes in the covariance structure, a clustering approach based on the proposed smooth dynamic functional principal components is suggested and the results of application are discussed.

Finally, Chapter 6 provides a summary of the work presented within this thesis, discusses the limitations and implications and proposes areas for future research.

Acknowledgements

Firstly, I would like to express my deepest gratitude to my supervisors Prof. Marian Scott and Dr. Claire Miller for their valuable guidance, continuous support and motivation throughout my Ph.D. I am extremely thankful for their precious suggestions and advices and limitless confidence and patience. I also take this opportunity to thank my external co-supervisor Prof. Susan Waldron for providing the data for this research and for her encouragement, generous explanations and constant help. Without this perfectly integrated leadership this thesis would not have been possible.

Special acknowledgments are due to the University of Glasgow, Science and Engineering College Sensor Studentship, for funding my Ph.D. and to the School of Mathematics and Statistics for providing the funds necessary to attend and participate in conferences.

My thanks extend to the whole department of Statistics at the University of Glasgow, including staff and postgraduate students. I would like to thank all the academic staff at the department for providing help and advice when needed and to all the staff in the School for answering all my questions. I am also thankful to all my colleagues and friends I have met in the department since I started my Ph.D., the Friday drinks, lunches, dinners and coffee times which were always the ultimate escape from hard work. I am extremely grateful, in particular, to Ruth and to all my office mates, Kelly, Mengyi, Guowen, Craig, Cunyi and Aisyah for their help, company, laughs and nice "wee" chats and lovely memories. Thank you, Mengyi, for all your excitements and enjoyable chats.

A special recognition is due to my teachers and colleagues at Cairo University for their support and for sharing their experience and knowledge.

Special thanks to the great and real friends I have made here in Glasgow. Thank you, Amira E., Linda, Mona, Maria and Samar for your time, advice and moral support and for being the best friends ever. Big thanks to my life time friends back home, Amira M. and Rania, for your love and cheering up phone calls that have always supported me. Many thanks to Mohammed Albadrawy for his constant care and motivation.

Last, but not least, I owe the deepest gratitude to my great family for their love, prayers and endless support and encouragement throughout all my studies. Thank you for encouraging me to follow my dream and pursue this Ph.D. and tolerating being far away from you. Thank you my little sister, Sarah, for your belief in me and for hauling me up when I hit rock bottom. Thank you my inspiring mum and dad for your endless love and caring and above all for teaching me to dream and achieve whatever I dream of with faith, hard work, determination and enthusiasm.

Contents

List of Figures

List of Tables

This thesis is dedicated to my parents for their endless love and support throughout my life, and to the loving memory of my cousin, Heba.

Chapter 1

Introduction

A sustainable environment is an essential aspect of life for all living organisms to grow and prosper. However, environmental challenges such as climate change, water and air pollution, changes in water quality and quantity, and loss of carbon in soil are increasingly threatening our planet. Studying these inherent or induced environmental changes is, therefore of great importance to society (Intergovernmental Panel on Climate Change IPCC, 2013) to understand the potential problems and propose realistic solutions to sustain environmental quality. Environmental monitoring programmes and technologies are continually being developed to enhance our ability to understand environmental systems and detect changes occurring within these systems. In the past, monitoring programmes typically involved monthly, fortnightly, weekly, and occasionally daily sampling campaigns but rarely shorter time intervals (Kirchner et al., 2004, Neal et al., 2012). However, in reality, most environmental processes are continuous in time with changes potentially occurring at sub-daily scales; and hence monitoring programmes of high temporal resolution are needed to observe and understand the significance of these rapid changes.

Sensor technology is continuously developing and as a result, the ability to record and store measurements is ever-improving (Yick et al., 2008). Accordingly, environmental monitoring programmes deploy these sensors to record hourly or sub-hourly (e.g., every minute) measurements (Moraetis et al., 2010). Sensor hydrological data recorded at short time frames over long time periods are known as “Hydrological High-Frequency Data (HHFD)” (Kirchner et al., 2004). Such HHFD allow us to address new research questions which were previously inaccessible (Kirchner et al., 2004). Jeong et al. (2006) have demonstrated that high-frequency environmental monitoring data offer new opportunities for better understanding and detection of pollution events in the coastal ocean. However, they outline many data processing and interpretation challenges, especially when the data are used to inform rapid management decision making. Using high-frequency (15-minute) sensor data of dissolved oxygen and temperature in an eastern Iowan stream, Loperfido et al. (2009) have identified diel dissolved oxygen dynamics, which were largest at the upstream station. Benefiting from the continuous monitoring programmes, Neal et al. (2013) have presented a high-frequency water quality data set comprising 2 years of 7-hourly water quality data for two streams in the Upper Severn catchment at Plynlimon in Mid-Wales, where measurements are available for 50 analytes. These data have shown complex patterns over a wide range of timescales, challenging the understanding of the catchment processes and informing advanced modeling efforts in the future. Recently, Ockendena et al. (2016) have shown that high-temporal resolution of river discharge, turbidity and nutrient data were useful to learn about the complex dynamics of nutrients and the hydrological and bio-geochemical processes of the catchments, which might be misrepresented by infrequent, low intensity sampling regimes. The high-frequency data of nitrate and total phosphorus have provided new understanding of the factors controlling their transport in a lowland stream.

It is evident that the enormous advances in sensor technology enable environmental monitoring programmes to record and store data at any arbitrary high-frequency more efficiently. This enables us to obtain an increasingly comprehensive picture of many environmental processes. However, such high-frequency data pose several statistical modeling and analysis challenges. Many of the currently available statistical methods and software tools are not designed to properly manipulate the complexity of such volumes of data (Kirchner et al., 2004), arising from the persistent correlation between observations, complex and varying dynamics over the different time scales and the inter-relationships between the drivers of the process. Therefore, new statistical methods are needed to analyze and model these large complex data sets. This thesis aims at exploring and presenting the complexities of analyzing and modeling hydrological high-frequency data using currently available advanced statistical methods, in addition to developing new statistical tools capable of reducing data dimensionality and complexity, efficiently analyzing HHFD and extracting useful information. Throughout the thesis, the statistical methodology is illustrated using a data set of high-resolution sensor-generated time series of partial pressure of carbon dioxide obtained for a small catchment located in Scotland.

1.1 Excess Partial Pressure of Carbon Dioxide ($EpCO_2$)

The increase in atmospheric carbon dioxide concentrations, resulting from the use of fossil fuels during recent years, is considered the ultimate cause for the rise in atmospheric temperature. Rivers and oceans have been identified as major sinks and sources for anthropogenic and atmospheric CO_2 (Intergovernmental Panel on Climate Change IPCC, 2001). Therefore, many research efforts have been devoted to the understanding and quantification of the aqueous carbon cycle, which plays an important role in the global carbon cycle and hence in controlling the climate on earth (Intergovernmental Panel on Climate Change IPCC, 2001). The aqueous partial pressure of carbon dioxide is a measure of the capacity for CO_2 exchange between the water and the atmosphere (Li et al., 2012). The excess partial pressure of carbon dioxide ($EpCO_2$) in surface freshwater is a dynamic representation of the interacting bio-geochemical and hydrological processes that produce, consume, and transport carbon dioxide (Waldron et al., 2007). If the river is over-saturated (an excess partial pressure, > 1), CO_2 is effluxed, representing direct linkage of terrestrial and atmospheric carbon cycles (Butman and Raymond, 2011). As surface waters are capable of degassing large amounts of CO_2 to the atmosphere (Cole et al., 1994, Li et al., 2013, Raymond et al., 1997, Richey et al., 2002, Yao et al., 2007), they have been included in the assessment of the global carbon budget (Butman and Raymond, 2011). Understanding of the temporal variability in the capacity for degassing and the drivers of such variability is of value in refining the uncertainty for the estimates of effluxed CO_2.

Many studies have examined the temporal and spatial variations of $EpCO_2$ and the mechanisms controlling these variations in some high and low order rivers and large lotic systems (Butman and Raymond, 2011, Cole et al., 1994, Dawson et al., 2009, Li et al., 2013, 2012, Raymond et al., 1997, Richey et al., 2002, Yao et al., 2007). All these studies have shown that high-order rivers are important sources of atmospheric CO_2, and that low-order rivers also contain very high concentrations of dissolved CO_2 (Butman and Raymond, 2011). For example, the Hudson River, flowing from north to south through eastern New York in the United States, is over-saturated with CO_2 with respect to the atmosphere and $EpCO_2$ exhibits a diel cycle that reaches its maximum in summer (Raymond et al., 1997). Richey et al. (2002) have concluded that the evasion of CO_2 from rivers of the central Amazon basin constitutes an important carbon loss process and that there is a pronounced seasonality in evasion linked to dry and wet seasons. However, the estimates of the effluxed CO_2 are uncertain because of the large temporal and spatial variability.

Yao et al. (2007) have studied the spatio-temporal dynamics of $EpCO_2$ in the lower reaches of Xijiang River using monthly measurements recorded at six sites along the river from April 2005 to March 2006. During the study period, the river was characterized by CO_2 over-saturation with clear seasonal and spatial variations due to the seasonal changes in the external bio-geochemical processes and internal dynamics. These processes include the transport of soil CO_2 through base flow and interflow, in-situ aquatic respiration of organic carbon, photosynthesis of aquatic plants and CO_2 evasion from water to air. The study has shown the presence of a seasonal relationship between $EpCO_2$ and water discharge, which is still incompletely understood due to the insufficient sampling frequency.

The seasonal variations in $EpCO_2$ have been investigated along an integrated river continuum within the Dee Basin located in North-East Scotland (Dawson et al., 2009). The water samples were collected weekly at 12 sites along the River Dee from October 2007 to September 2008. One-Way ANOVA and post-hoc tests were employed to assess significant differences between sites. Then, a seasonality index was constructed and t-tests were used to determine significant differences between the mean seasonal $EpCO_2$ values at each site. It is evident that $EpCO_2$ exhibits seasonal variations in the catchment biogeochemical processes and that the variability of processes controlling $EpCO_2$ are spatial and time scale dependent. Pearson correlation coefficients and multiple linear regression highlighted the non-significant relationship between water discharge and $EpCO_2$, although low $EpCO_2$ values were clearly associated with low flow. This paper provided primary insights about the spatial and intra-annual variability of the dynamics controlling $EpCO_2$ in surface waters. However, sub-daily measurements are needed to enhance the understanding of smaller-scale temporal fluctuations of free CO_2 concentrations.

More recently, daily measurements of $EpCO_2$ from July 2008 to August 2009 were used to reveal the $EpCO_2$ daily to seasonal dynamics in the upper Yangtze River basin (Li et al., 2012). During the whole study period, the river was characterized by CO_2 oversaturation with respect to the atmosphere and the $EpCO_2$ exhibited clear daily and monthly variations. The results have indicated that $EpCO_2$ is more affected by pH and partially regulated by temperature through altering the alkalinity and DIC concentrations. In contrast, $EpCO_2$ appears to be weakly correlated with water flow due to the monthly variability in $EpCO_2$ and the huge fluctuations of $EpCO_2$ in the flooding period. This weak relation could be attributed to insufficient sampling. Therefore, the authors have concluded that higher frequency sampling in space and time are required, to better understand the $EpCO_2$ dynamics, due to the spatio-temporal heterogeneity in the catchment characteristics and anthropogenic activities.

Generally, $EpCO_2$ is highly dynamic in lower-order rivers (Waldron et al., 2007) and therefore sub-daily measurements across different seasonal periods should provide sufficient detail to understand fluctuations of $EpCO_2$ concentrations at smaller timescales (Dawson et al., 2009). In this thesis, three years of (15-minute) frequency sensor-generated data are used to investigate and reveal the diurnal, seasonal and inter-annual variations of $EpCO_2$ and explain the mechanisms controlling these variations in a small-order river. This long-term hydrological high-frequency data set encompasses seasonality and varying time periods between hydrological events, but also allows many new features, including pulses and short duration events to be identified, which would not have been apparent with monthly or daily sampling.

The following section describes the study site and its geological and climatological characteristics, the sampling strategy, the sensor *in-situ* data and the calculation of excess partial pressure carbon dioxide. Next, Section 1.3 explores and presents the main features of the data and the results of the primary exploratory analysis carried out. Then, a description of different time series analysis techniques is provided in Section 1.4. Subsequently, Section 1.5 presents a time-frequency domain approach for exploring and analyzing high-frequency time series with application to the available EpCO_2 data. Finally, the objectives and structure of the thesis are presented in Sections 1.6 and 1.7, respectively.

1.2 Available Data

1.2.1 Study Site

The study site is in the Glen Dye catchment close to the terrestrial-aquatic interface of the River Dee in Aberdeenshire Glen Dye is located in North-East Scotland at $56^o56'27$N and $2^o36'00$W. It is a headwater sub-catchment of the River Dee, a high-order river draining into the North Sea. The sensors were deployed at the Scottish Environment Protection Agency (SEPA) Charr gauging flume on the Water of Dye, a 41.7km^2 catchment. Glen Dye is mainly upland in character, with altitude ranging between 100m and 776m. The climate is cold, with mean annual precipitation of 1130mm, of which $<$10% is snow. There is inter-annual variation in temperature with the winter months being December - February and the summer months being June - August. The underlying geology of the catchment is granite, with a small schist outcrop. The interfluves above 450m are covered by extensive peats ($<$ 5m deep) and peaty podzols ($<$ 1m). In some places peat is eroded to the mineral interface. Incised catchment slopes have the most freely-draining humus iron podzols ($<$ 1m deep); the main river valley bottoms generally have freely draining alluvial deposits. For a detailed description of the study site and its geology and climate characteristics, see (Waldron et al., 2007).

1.2.2 Sampling Strategy and Calculation of EpCO_2

Samples for measurement of Dissolved Inorganic Carbon (DIC) concentration were collected approximately every 5 hours over a 24-hour period and 12 times during June 2003

- August 2004. The sampling spanned a wide range of flow conditions. DIC (mmol L^{-1} C) is quantified by direct measurement using a headspace analysis approach (Waldron et al., 2014), to internal precision better than ± 0.03 mmol L^{-1}. This was regressed on discharge, which is measured semi-continuously, to generate a relationship from which DIC is predicted, thus creating a continuous DIC profile (Waldron et al., 2007). The generated relationship between discharge and DIC was indistinguishable from the same relationship constructed 10 years earlier, allowing confidence that this relationship is temporally stable over the constructed three years profile. Troll 9000EXP data loggers (In-Situ, Inc.) were used to generate 15-min frequency time series of temperature, pH and atmospheric pressure from October 2003 to September 2006, spanning three hydrological years. These measurements allowed the excess partial pressure of carbon dioxide ($EpCO_2$) to be indirectly calculated from the continuous DIC profile (Waldron et al., 2014). Estimates of the capacity for CO_2 efflux, are described as the "Excess partial pressure of CO_2", $EpCO_2$, a ratio of over-saturation (for more details, see Neal (1988) and Dawson et al. (2009)). The river system is over-saturated with CO_2 with respect to the atmosphere when $EpCO_2$ exceeds 1 and free CO_2 is in equilibrium with the atmosphere when $EpCO_2$ equals 1. The Troll loggers also generated 15-min frequency time series of specific conductivity (SC). SC in streams and rivers is influenced by the river geology, in addition to the water flow and temperature (United States Environmental Protection Agency EPA, 2012). It is usually higher in low flow periods when the groundwater contribution is proportionally highest (United States Environmental Protection Agency EPA, 2012).

1.3 Exploratory Data Analysis

Standard exploratory tools are initially employed to visualize the fundamental fluctuations of $EpCO_2$, at the Water of Charr, over time and depict insights about its diurnal and seasonal variations. The Exploratory Data Analysis (EDA) also helps in getting primary knowledge about the changes in $EpCO_2$ in relation to the water hydrology and the physicochemical variables such as water discharge, temperature, pH and specific conductivity. Figure 1.1 displays the calculated $EpCO_2$ high-frequency series and the recorded measurements of discharge, temperature, pH and specific conductivity over the three hydrological years starting from October 2003 to September 2006. Each hydrological year runs from October to September and hereafter is abbreviated as HY. The $EpCO_2$ exhibits temporal variability and varies between 0.26 to 10. The average $EpCO_2$ over the whole study period is 2.57 ± 1.01. Thus, our sample point on the Water of Charr is, on average, over-saturated with CO_2 with respect to the atmosphere. Similarly, water discharge is temporally variable (second top row of Figure 1.1) with an average of 1.1 ± 4.5 km^3 through the whole study period. A comparison of total discharge between HYs shows that HY2003/2004 had the wettest summer, HY2004/2005 had the driest winter and HY2005/2006 was the wettest overall with the wettest winter and driest summer (Table 1.1). The coldest months in the three hydrological years are December - February (Third row of Figure 1.1) with an average water temperature of 2.9 ± 1.7^oC; and the warmest months are June - August with an average temperature of 14 ± 2.8^oC. Figure 1.1 also shows the presence of a period of missing observations in the sensor-generated data and consequently the $EpCO_2$ series in July 2004.

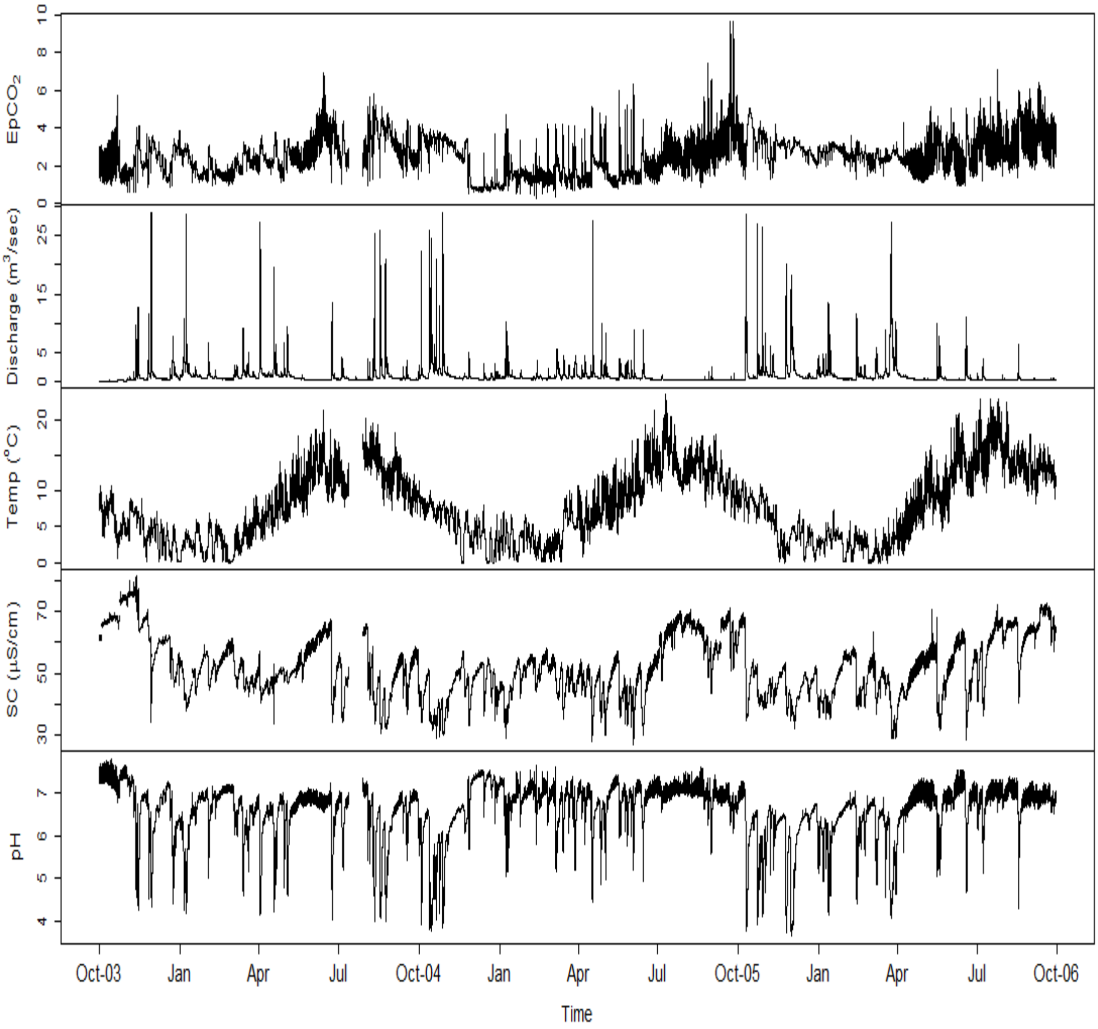


FIGURE 1.1: Time plots of the 15-minute $EpCO_2$ (top), water discharge, temperature, SC and pH (bottom) series from 1^{st} October 2003 to 30^{th} September 2006.

Figure 1.2 illustrates the seasonal and diurnal responses in $EpCO_2$ in each of the HYs. The top box-plots in Figure 1.2 show the average distribution of $EpCO_2$ in both summer and winter, while the middle row plots (d-f) display the monthly distributions of $EpCO_2$. The median $EpCO_2$ (represented by the black bar in the middle of each box)

HY	Discharge m^3/s		
	Winter	Summer	Overall
HY2003/3004	10,083	8882	37,063
HY2004/2005	6496	3880	35,958
HY2005/2006	12,385	3726	40,044

TABLE 1.1: Total water discharge at the water of Charr sampling point across the winter (December – February) and summer (June – August) of each HY and across the whole HY.

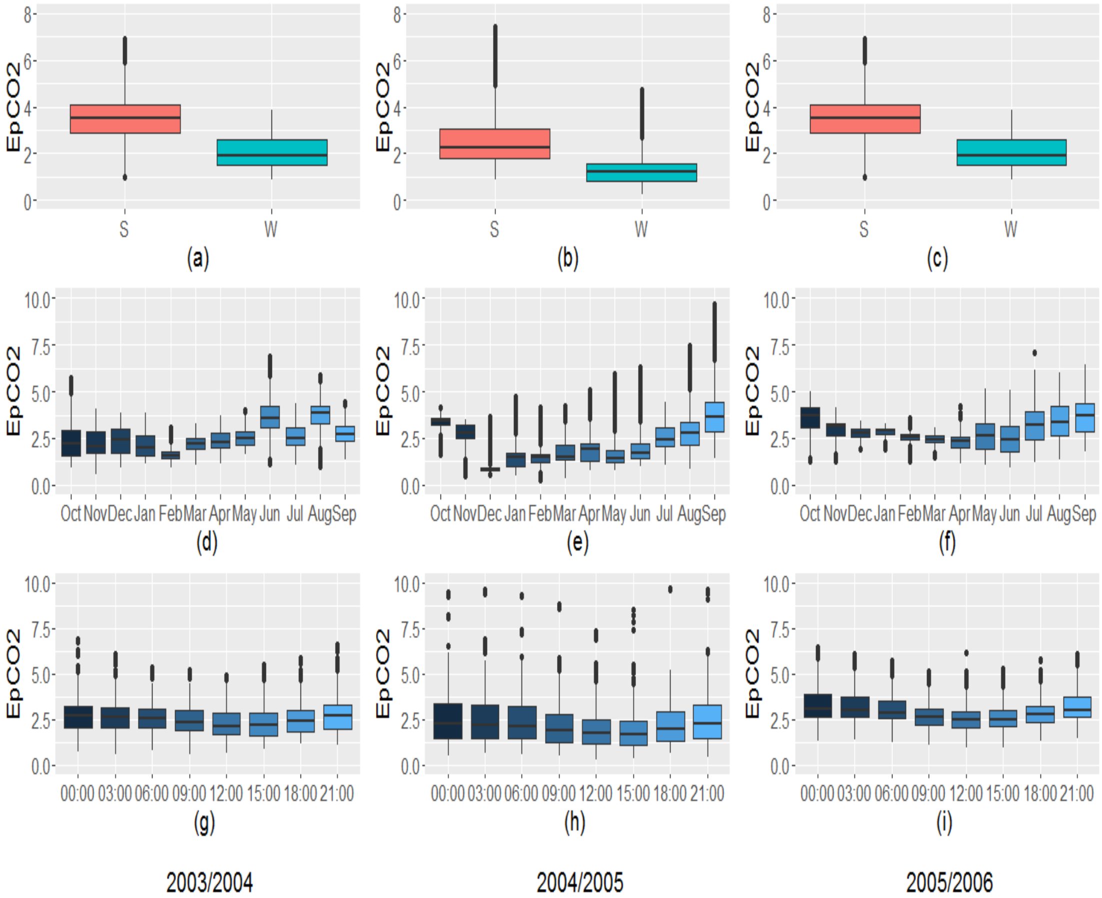


FIGURE 1.2: Box-plots of $EpCO_2$ by season (top), month (middle) and hour (bottom) in the HYs 2003/2004 (right), 2004/2005 (middle) and 2005/2006 (left).

is generally higher in summer (June - August) compared to winter (December - February). $EpCO_2$ is clearly more variable during summer. This is attributed to the greater catchment productivity in summer when more (or sources of) CO_2 are available. The bottom box-plots (g-i) portray the intra-daily variations of $EpCO_2$, where the median $EpCO_2$ is smallest close to midday and largest just after midnight. The drop of $EpCO_2$ during day-time is due to the active biogenic consumption of CO_2 (i.e. photosynthesis) occurring during the sunlight hours and masking the organic carbon respiration activity. It is noticed that $EpCO_2$ exhibits more variability during darkness, attributed to the variability in the aquatic respiration of organic carbon. Figure 1.3 shows that the diel variations are more apparent during the summer months. The mean and variability of $EpCO_2$ are almost constant throughout the 24-hour day from November to January. From February to April, the mean $EpCO_2$ starts to decrease slightly during day time relative to the night hours. This relative drop in $EpCO_2$ during the light hours becomes more pronounced in the subsequent summer months (May - September). The $EpCO_2$ decreases for a shorter time period of the day during late summer compared to earlier summer months, explained by the decrease in daylight duration as autumn approaches.

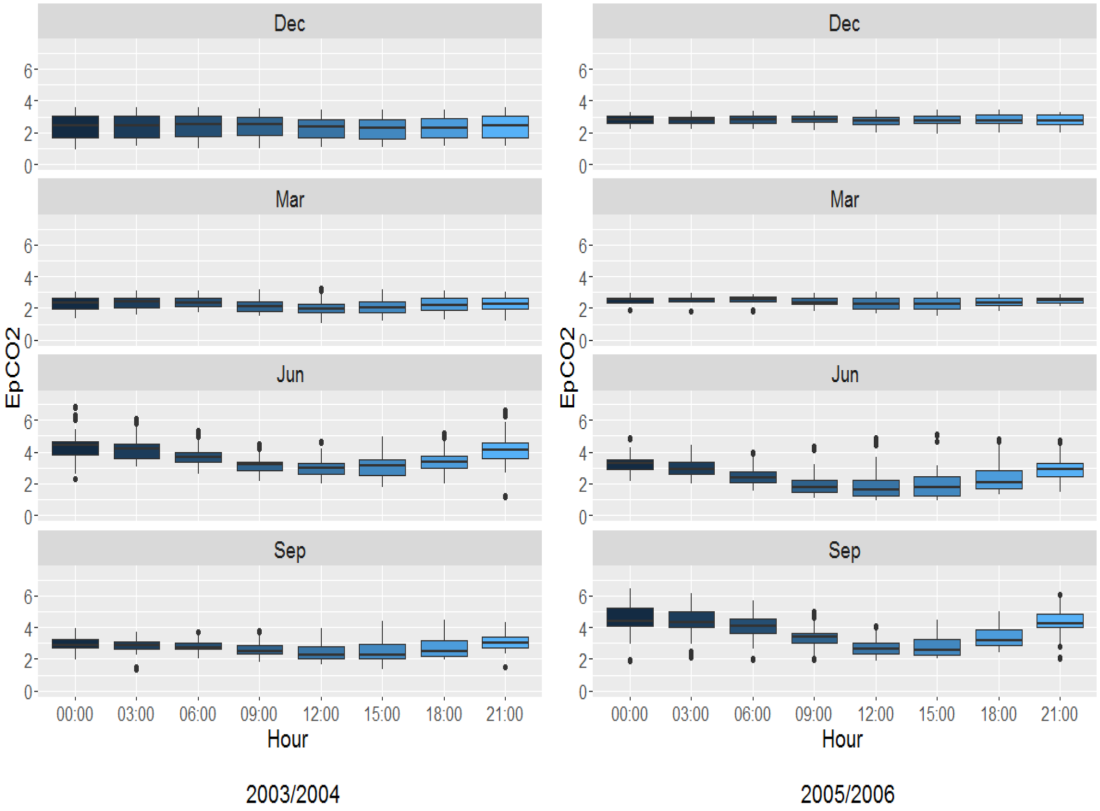


FIGURE 1.3: 3-hourly box-plots of $EpCO_2$ in December, March, June and September of the HYs 2003/2004 (left) and 2005/2006 (right).

Previous studies (Dawson et al., 2009, Li et al., 2012, Yao et al., 2007) indicated the existence of relationships between the partial pressure of carbon dioxide and the climatological and hydrological variables such as water temperature and water discharge. Further standard exploratory plots are therefore used to primarily investigate these relationships in the high-frequency data collected at the water of Charr. Though $EpCO_2$ is indirectly calculated based on the water temperature, Figure 1.4 portrayed a positive relationship between them only during night time that becomes negligible in the daylight hours. The other HYs, not shown here due to space limitations, displayed similar patterns. Temperature itself varies considerably from month to month, which in turn allows

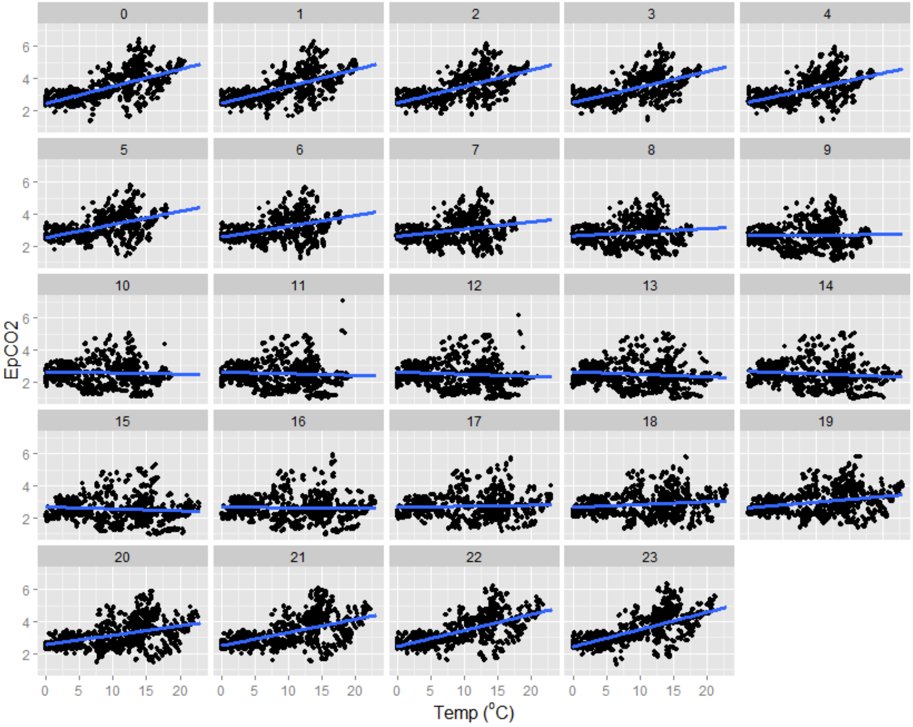


FIGURE 1.4: Scatter plots of $EpCO_2$ versus temperature, with the least-squares regression line, across the 24 hours of the day in the HY2005/2006.

its relationship with $EpCO_2$ to change across the 24-hour day and also across the different months. That is, the relationship between $EpCO_2$ and temperature is dynamic over the different timescales. The relationship between $EpCO_2$ and water discharge is also investigated using scatter plots (see Figure 1.5). It is evident that simple exploratory plots are not effective tools in describing such complex and inter-related hydrological high-frequency data, as the response of $EpCO_2$ to flow events may differ depending on preceding events and differences in summer and winter biological productivity. Animated 3-D plots, available at https://static-content.springer.com/esm/art%3A10.1007%2Fs10651-015-0329-4/MediaObjects/10651_2015_329_MOESM1_ESM.pdf provided a more enhanced representation of the dynamic interaction between $EpCO_2$, water discharge and temperature over time. Advanced statistical time series techniques are evidently needed to explore and analyze the main patterns of $EpCO_2$ and its relationship with river hydrology over time.

The following section gives a brief description of the most common standard time series analysis techniques used to scrutinize and analyze environmental data in general and hydrological data in particular. After that, Section 1.5 introduces a more advanced tool for exploring time series in both frequency and time domains simultaneously.

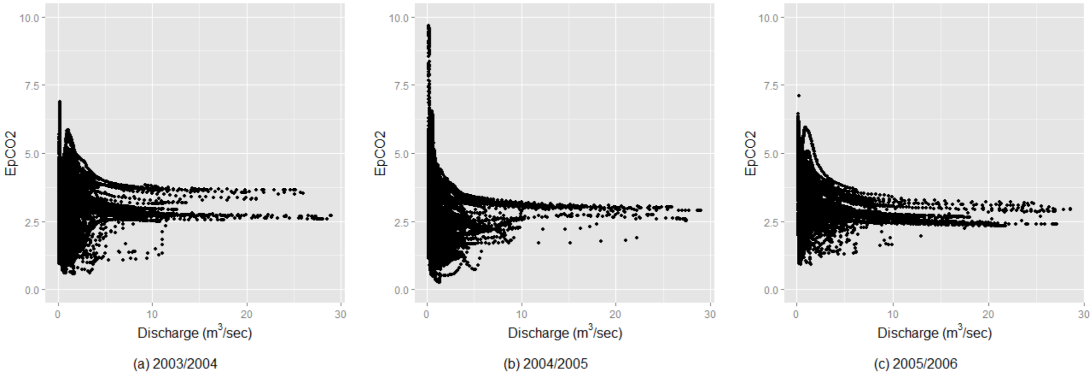


FIGURE 1.5: Scatter plots of $EpCO_2$ versus water Discharge (m^3/s) in the HYs (a) 2003/2004, (b) 2004/2005 and (c) 2005/2006.

1.4 Time Series Modeling of Hydrological Data

Hydrological data are usually collected over time or/and space. A sequence of equally spaced data points recorded over a continuous time period at regular intervals are known as *time series.* The ultimate objective of time series analysis, in environmental studies, is to describe the pattern of a certain phenomenon over time to identify appropriate control measures. The long-term increase or decrease in the average of a time series, which does not have to be linear, is called *trend.* Whereas, the repeated regular short-term patterns are known as *seasonal* or *cyclic* patterns. The magnitude and period of cyclic patterns tend to be more variable relative to those of seasonal patterns.

The correlation induced by the adjacency of observations in time limits the applicability of many standard statistical methods that typically assume independent and identically distributed errors. Analyzing serially correlated data assuming independence results in smaller standard error estimates. This temporal correlation is less significant if the environmental data are collected further apart in time. However, the current advances in sensor technology permit hydrological data to be recorded at shorter time intervals, which in turn increase the dependence between successive observations and its influence on the estimates of standard errors. The analysis of time series data taking into account the serial dependence between observations is referred to as time series analysis. There are mainly two approaches, not necessarily mutually exclusive, to time series analysis: the time domain approach and the frequency domain approach. Shumway and Stoffer (2011) is the main reference textbook used for providing the following short review of these two approaches.

The *time domain approach* involves two methods for analyzing time series data. The first method aims at describing the behaviour of the time series over time and assumes that the observed time series are generated from the sum of trend, a seasonal and/or cyclic effect and error. In the second method, serial correlation between observations is assumed to be better explained by the dependence of the current observation on the past values; and hence this approach focuses on modeling future values as a function of the current and past values.

According to the first time domain method, a time series observation at a time point t, Y_t, is represented as:

$$Y_t = m_t + S_t + C_t + \epsilon_t, \tag{1.1}$$

where m_t denotes the trend, S_t and C_t are the repetitive seasonal and cyclic effects, respectively and ϵ_t is a white noise random error. The terms m_t, S_t and C_t can be estimated using known parametric functions of time. For example, the trend can be described using a linear or any higher order polynomial regression, and the seasonal/-cyclic effects can be estimated using harmonic regression. Non-parametric regression can also be used to evaluate these terms if they appear to exhibit more complex patterns. Model 1.1 assumes that the variability in time series is only explained by trend and some periodic patterns and that the errors are independent. However, as previously mentioned, independence between adjacent observations in time is not a realistic assumption. Hence, standard model-fitting techniques such as ordinary least squares are no longer valid in time series analysis. Box and Jenkins (1970) developed a class of models, known as Autoregressive Integrated Moving Average (ARIMA) models, to appropriately handle auto-correlation.

In Autoregressive (AR) models, the current value of the series y_t is expressed as a linear combination of the past p values. An autoregressive model of order p, denoted by AR(p), is of the form:

$$y_t = \sum_{h=1}^{p} \rho_h y_{t-h} + \varepsilon_t,$$

where $\rho_h, h = 1, \ldots, p$ are auto-regression coefficients to be estimated and ϵ_t is a white noise random error. Using the lag operator $\mathbb{B}^k y_t = y_{t-k}$, the AR($p$) models can be equivalently expressed as $\varepsilon_t = (1 - \rho_1\mathbb{B} - \rho_2\mathbb{B}^2 - \ldots - \rho_p\mathbb{B}^p)y_t$. An AR($p$) model is considered stationary if all the roots of the equation $(1 - \rho_1\mathbb{B} - \rho_2\mathbb{B}^2 - \ldots - \rho_p\mathbb{B}^p) = 0$ lie outside the unit circle, i.e. all $|\mathbb{B}| > 1$. In Moving Average (MA) models, the current value of the series is modeled in terms of the current and past q errors. Thus, a moving average model of order q, MA(q), is expressed as follows:

$$y_t = \sum_{h=1}^{q} \theta_h \varepsilon_{t-h} + \varepsilon_t,$$

where $\theta_h, h = 1, \ldots, q$ are coefficients of lagged error terms to be estimated and ϵ_t is a white noise error term. Following from this, an Autoregressive Moving Average model denoted by ARMA(p,q), where p is the order of the autoregressive part and q is the moving average order, has the following form:

$$y_t = \sum_{h=1}^{p} \rho_h y_{t-h} + \sum_{h=1}^{q} \theta_h \varepsilon_{t-h} + \varepsilon_t.$$

The above class of ARMA models, including AR and MA models, is principally developed for stationary time series. A time series is stationary if its mean and variance are independent of time and if the covariance between observations depend only on the time lag between them and not the actual time points. This implies that under stationarity the time series does not exhibit a trend nor seasonal/cyclic patterns. In real life applications, most of the time series are not stationary. Therefore, Box and Jenkins (1970) extended the class of ARMA models to the very popular ARIMA models designed for non-stationary data. A time series y_t is known to follow an ARIMA(p,d,q) model if its d^{th} difference follows a stationary ARMA(p,q). If $d=1$, then the first order difference of the series y_t is defined as $\Delta y_t = y_t - y_{t-1}$. The first order difference is often applied to remove trend from the data. Period differencing is rather applied to remove seasonal effects, where a d-period differencing is obtained by $\Delta^d y_t = y_t - y_{t-d}$. After differencing the time series to obtain a stationary series, an ARMA(p,q) process is used to model the $\Delta^d y_t$ series. To determine the order (p and q) of the ARMA model, the sample auto-correlation function (ACF) and the sample partial auto-correlation function (PACF) are computed. The sample auto-correlation function (ACF) is a sequence of auto-correlation coefficients computed at the different time lags h between the values of the time series at time t and the values at time $t-h$. Let $y_1, \ldots, y_N$ be a stationary time series of size N and $\bar{y}$ be the sample mean of the series, then the sample auto-correlation coefficient at lag h is obtained by:

$$r_h = \frac{\sum_{t=1}^{N-h}(y_t - \bar{y})(y_{t+h} - \bar{y})}{\sum_{t=1}^{N}(y_t - \bar{y})^2}$$

The partial auto-correlation function (PACF) is the sequence of auto-correlation coefficients calculated at the different lags h between y_t and y_{t-h} after omitting the effect of $y_{t-1}, \ldots, y_{t-h+1}$. Plots of both the sample auto-correlation and partial auto-correlation coefficients versus the sequence of time lags h are called correlograms. These correlograms are practically used to identify the ARMA model order. For instance, an AR model is identified to model a time series whose ACF shows an exponentially decaying pattern and the number of significant PACF coefficients to lag p is the order p. Alternatively, an exponentially decaying PACF suggests an MA(q) model and the order q is determined by the number of significant ACF coefficients to lag q. For more details, see Box and Jenkins (1970) and Shumway and Stoffer (2011).

The above time domain approaches can be combined together by modeling first the trend and seasonal components using either parametric or non-parametric regression methods, then accounting for the remaining temporal auto-correlation in the residuals. The auto-correlation in the errors can be modeled using an ARMA process to adjust the standard errors of the estimated model parameters. Most of the environmental and hydrological variables do not have a predetermined functional form over time. Therefore, a wide range of non-parametric flexible regression models (Hastie and Tibshirani, 1990) have been developed and used to model the pattern of environmental data smoothly over space and time, see Bowman et al. (2009), Ferguson et al. (2008), Giannitrapani et al. (2011). In most of these applications, an AR(1) has been used to account for the remaining temporal correlation in the error term. The main objective in most environmental applications is to describe the mean structure; and hence it is sufficient to employ a simple model for the errors, effectively considered as a nuisance parameter, to account for the main effects of correlation.

According to the above exploratory analysis, the $EpCO_2$ exhibits non-linear patterns over time and complex relationships with water hydrology. Therefore, non-parametric flexible regression, described in more detail in the following chapter, is considered as a useful tool for modeling the $EpCO_2$ high-frequency data.

The *frequency domain approach* assumes that any time series involves periodicity and that the variability in time series data is related to periodic and systematic variations. These periodic variations are usually represented by sine and cosine cycles of different frequencies. In this approach, the total variance in a time series is partitioned by evaluating the variance associated with each time frequency. The distribution of the variance over the different frequencies is called the spectral density/power spectrum. The power spectrum shows how much signal lies within each frequency. Wavelet analysis is a very popular tool for analyzing non-stationary time series in both the frequency and time domains. Examples of using wavelets in analyzing hydrologic time series can be found in Franco-Villoria et al. (2012), Labat (2005), Sen (2009), White et al. (2005). In view of that, wavelets offer an advanced useful exploratory tool to study the different temporal variations of a non-stationary time series.

The exploratory data analysis has clearly indicated that the $EpCO_2$ high-frequency time series exhibits a wide range of temporal variability. To gain a better understanding of how the $EpCO_2$ varies in both the time and frequency domains, wavelet analysis is employed. In this thesis, wavelet analysis is mainly used as an exploratory tool for visualizing the variability in the high-frequency time series of $EpCO_2$ across the different timescales throughout time. This wavelet analysis will help in identifying the dominant seasonal and cyclical variations in the time series. The methodology of wavelets and the analysis of the $EpCO_2$ data using wavelets are presented in detail in the following section.

1.5 Wavelets Analysis

In wavelet analysis, the time series is first decomposed into different frequency components then each component is analyzed with a resolution matched to its scale. Wavelet analysis is quite attractive for its time-scale localization and has the advantage of analyzing a time series by combining both approaches, time and frequency domains. The theory of wavelets has developed over the past decade to cover a wide range of applications such as non-stationary signal processing, noise filtering, data reduction and singularity detection, which make the approach a useful choice for analyzing high-frequency time series.

A wavelet is a small wave that grows and decays within a limited period of time. Mathematically, a wavelet is a real valued function, ψ, of a real variable, t (t here refers to time), that satisfies two conditions (Percival and Walden, 2006):

$$\int_{-\infty}^{\infty} \psi(t)dt = 0, \tag{1.2}$$

and

$$\int_{-\infty}^{\infty} \psi^2(t)dt = 1. \tag{1.3}$$

The first condition 1.2 implies that if $\psi(t)$ has some positive values, it also has some negative values; that is, $\psi(t)$ has the form of a wave. The second condition 1.3 implies that $\psi(t)$ has to be non zero over a finite interval $[-T, T]$ which is a small interval compared to the interval $(-\infty, \infty)$ on which the whole function is defined. Hence, wavelets are oscillating objects that decay quickly. Wavelets may have an additional condition known as the admissibility condition, which allows the reconstruction of the original signal from its wavelet transform. A wavelet $\psi(.)$ is said to be admissible if (Percival and Walden, 2006):

$$0 < C_\psi \equiv \int_0^{\infty} \frac{|\Psi(\theta)|^2}{\theta} d\theta < \infty, \tag{1.4}$$

where $\Psi(\theta)$ is the Fourier transform of the wavelet $\psi(t)$ at frequency θ, given by:

$$\Psi(\theta) \equiv \int_{-\infty}^{\infty} \psi(t)e^{-i2\pi\theta t}dt$$

.

Wavelets are generated by the dilation and translation of a predetermined mother wavelet as follows (Nason, 2008):

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k),$$

where j and k are integers representing the scale and location, respectively. The set of wavelets $\{\psi_{jk}(x)\}_{j,k\in\mathbb{Z}}$ forms an orthogonal basis for the space to which the signal x belongs. Let $x(t)$ be a real valued function of an independent variable t denoting time that can be decomposed using wavelets in the function space L^2 as follows:

$$x(t) = \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} d_{j,k}\psi_{j,k}(t),$$

where

$$d_{j,k} = \int_{-\infty}^{\infty} x(t)\psi_{jk}(t)dt.$$

The $\{d_{jk}\}_{j,k\in\mathbb{Z}}$ are known as the wavelet coefficients of the signal x, obtained by projecting the original signal x onto the mother wavelet $\psi_{j,k}(t)$. A large value of j assesses the variation of the signal on a large scale.

There exists two versions of wavelets; the Discrete Wavelet Transform (DWT) and the Continuous Wavelet Transform (CWT). The DWT decomposes a time series into discrete different scales. While in the CWT, a time series is decomposed continuously using any arbitrary scale. That is, the DWT is a sampled version of the CWT that uses only a limited number of translated and dilated versions of the mother wavelet to decompose the original signal. The aim here is to use wavelets to describe the behaviour of the time series at a set of time scales and therefore, the main focus will be on DWT. A detailed description of the DWT is given below, where the theory results have been mainly extracted from Percival and Walden (2006).

1.5.1 Discrete Wavelet Transform

The DWT is an orthogonal transform of the time series $\{x_t : t = 1, \ldots, N\}$. Let $\{x_t\}$ be equally spaced in time such that each observation x_t is collected at time $t\Delta t$, where Δt is the time interval between adjacent observations. Also, let $\{x_t\}$ be of length $N = 2^J$, where J is a positive integer. Then, the DWT of level J_0 $(J_0 \leq J)$ of the time series $\{x_t\}$ is defined by:

$$\mathbf{W} = \boldsymbol{\mathcal{W}}\mathbf{X}, \tag{1.5}$$

where $\mathbf{X}$ is the time series vector of length N, $\mathbf{W}$ is the DWT coefficients vector of length $N = 2^J$ whose n^{th} element is the n^{th} DWT coefficient denoted by W_n, $n = 1, \ldots, N$

and $\boldsymbol{\mathcal{W}}$ is an $N * N$ real valued matrix defining the DWT such that $\boldsymbol{\mathcal{W}}^T\boldsymbol{\mathcal{W}} = \mathbf{I}_N$. The matrix $\boldsymbol{\mathcal{W}}$ is constructed based on the chosen wavelet filter (see Section 1.5.2 for details). Thanks to the orthonormality of $\boldsymbol{\mathcal{W}}$, the series $\mathbf{X}$ can be reconstructed by $\mathbf{X} = \boldsymbol{\mathcal{W}}^T\mathbf{W}$. The first $N - \frac{N}{2^{J_0}}$ elements of the vector $\mathbf{W}$ of the DWT coefficients are known as wavelet coefficients and the last $\frac{N}{2^{J_0}}$ elements are called scaling coefficients. Each wavelet coefficient is associated with changes on a particular scale $\tau_j \Delta t$, where $\tau_j = 2^{j-1}$, at a localized set of times. There are $\frac{N}{2^j}$ wavelet coefficients associated with each scale $\tau_j \Delta t$, for $j = 1, \ldots, J_0$. These wavelet coefficients will be large or close to zero depending on the amount of variation in the series at the corresponding scale and time. Whereas, the $\frac{N}{2^{J_0}}$ scaling coefficients are associated with variations on scales $\tau_{J_0+1}\Delta t$ and higher. The wavelet and scaling coefficients change with scale and time and denoted by $W_{j,t}$, where j and t refer to the scale τ_j and time t, respectively. Due to the orthonormality of $\boldsymbol{\mathcal{W}}$, the coefficients within each scale are almost independent.

The vector $\mathbf{W}$ of the DWT coefficients can be decomposed into $J_0 + 1$ vectors denoted by $\mathbf{W}_1, \ldots, \mathbf{W}_{J_0}, \mathbf{V}_{J_0}$. Each vector $\mathbf{W}_j$, $j = 1, \ldots, J_0$, contains $\frac{N}{2^j}$ wavelet coefficients associated with variations in the time series over a scale of $\tau_j = 2^{j-1}$. While the vector $\mathbf{V}_{J_0}$ contains the $\frac{N}{2^{J_0}}$ scaling coefficients. Following from this, the rows of the DWT matrix $\boldsymbol{\mathcal{W}}$ can be grouped into $J_0 + 1$ submatrices $\boldsymbol{\mathcal{W}}_1, \ldots, \boldsymbol{\mathcal{W}}_{J_0}, \boldsymbol{\mathcal{V}}_{J_0}$, where $\boldsymbol{\mathcal{W}}_j$ is an $N * \frac{N}{2^j}$ matrix, and $\boldsymbol{\mathcal{V}}_{J_0}$ is an $N * \frac{N}{2^{J_0}}$ matrix. Thus, $\mathbf{W}_j = \boldsymbol{\mathcal{W}}_j\mathbf{X}$ and $\mathbf{V}_{J_0} = \boldsymbol{\mathcal{V}}_{J_0}\mathbf{X}$.

As mentioned above, due to the orthonormality of $\boldsymbol{\mathcal{W}}$, the original signal $\mathbf{X}$ can be recovered from its DWT as follows:

$$\mathbf{X} = \boldsymbol{\mathcal{W}}^T\mathbf{W} = \sum_{j=1}^{J_0} \boldsymbol{\mathcal{W}}_j^T\mathbf{W}_j + \boldsymbol{\mathcal{V}}_{J_0}^T\mathbf{V}_{J_0}, \tag{1.6}$$

and $\|\mathbf{W}\|^2 = \|\mathbf{X}\|^2$. Now, let $\mathbf{D}_j = \boldsymbol{\mathcal{W}}_j^T\mathbf{W}_j$ for $j = 1, \ldots, J_0$ and $\mathbf{S}_{J_0} = \boldsymbol{\mathcal{V}}_{J_0}^T\mathbf{V}_{J_0}$, then the multiresolution analysis (MRA) of $\mathbf{X}$, used to reconstruct the original time series from the wavelet and scaling coefficients, is defined by:

$$\mathbf{X} = \sum_{j=1}^{J_0} \mathbf{D}_j + \mathbf{S}_{J_0}, \tag{1.7}$$

$\mathbf{D}_j$ is known as the wavelet detail at scale τ_j, $j = 1, \ldots, J_0$, and $\mathbf{S}_{J_0}$ is called the wavelet smooth at the maximal scale τ_{J_0}. The wavelet smooth $\mathbf{S}_{J_0}$ represents a smooth version of $\mathbf{X}$ and can be regarded as the long-term trend of the time series. Following this representation, the MRA of a time series helps identifying the long-term trend and cyclic components of the series and how these components vary with time.

The matrix $\boldsymbol{\mathcal{W}}$ defining the DWT is formed based on the choice of the filter. There exist two types of filters in wavelet analysis: wavelet filter and scaling filter. The application of both filters allow separating the low-frequency components of the signal from its high-frequency components in J_0 steps following a pyramid algorithm, as will be explained in the following section.

1.5.2 Filtering

The distribution of the variance of a time series over the frequency components θ is described by the spectral density $f(\theta)$. To extract a particular signal from the time series, a linear filter is used. A filter is a sequence $\{u_l\}$ whose Discrete Fourier Transform (DFT) or transfer function is given by:

$$A(\theta) \equiv \sum_{l=-\infty}^{\infty} u_l e^{-\mathrm{i}2\pi\theta l}, \qquad -\infty < \theta < \infty,$$

with associated squared gain function $|A(\theta)|^2$ and a corresponding polar representation $A(\theta) = |A(\theta)|e^{\mathrm{i}\omega(\theta)}$, where i denotes the imaginary unit. The most common filters in wavelet analysis are the zero and linear phase filters.

To introduce the *zero phase filters*, let $\{u_l^o : l = 0, \ldots, N-1\}$ denotes the filter $\{u_l\}$ periodized to length N and $\{U_k^o\}$ be the corresponding DFT, then:

$$U_k^o \equiv \sum_{l=0}^{N-1} u_l^o e^{-\mathrm{i}2\pi kl/N} = U(\theta_k) \quad \text{with} \quad \theta_k \equiv \frac{k}{N}.$$

Now, consider a time series $\{x_t : t = 0, \ldots, N-1\}$ whose DFT is given by:

$$\mathcal{X}_k \equiv \sum_{t=0}^{N-1} x_t e^{-\mathrm{i}2\pi kt/N}, \tag{1.8}$$

and which can be reconstructed using the inverse DFT as follows:

$$x_t = \frac{1}{N} \sum_{k=0}^{N-1} \mathcal{X}_k e^{\mathrm{i}2\pi kt/N}.$$

This time series $\{x_t\}$ can be circularly filtered with the periodized filter $\{u_l^o\}$ obtaining:

$$Y_t \equiv \sum_{l=0}^{N-1} u_l^o x_{t-l \bmod N}, \qquad t = 0, \ldots, N-1.$$

Let the DFT of the filtered series $\{Y_t\}$ be $\{U_k^o \mathcal{X}_k\}$, then $\{Y_t\}$ can be rewritten equivalently in terms of its inverse DFT as follows:

$$Y_t = \sum_{k=0}^{N-1} U_k^o \mathcal{X}_k e^{-\mathrm{i}2\pi kt/N}.$$

The transfer function or the DFT of the filter, $U(.)$, can be written using the polar representation as $U(\theta) = |U(\theta)|e^{\mathrm{i}\omega(\theta)}$. $\omega(\theta_k) = 0$ for all k, if $\{u_l\}$ is a zero filter. This makes $U(\theta_k) = |U(\theta_k)|$ and $U_k^o = |U_k^o|$, and so the filtered series $\{Y_t\}$ can be equivalently expressed as:

$$Y_t = \sum_{k=0}^{N-1} |U_k^o| \mathcal{X}_k e^{-\mathrm{i}2\pi kt/N}.$$

According to this expression, filtering $\{x_t\}$ with the periodized filter $\{u_l^o\}$ results in a filtered series $\{Y_t\}$ that differs from $\{x_t\}$ in the amplitudes of the sinusoids and not in their phases. This allows the events in $\{Y_t\}$ to be aligned to the events in $\{x_t\}$.

To define *linear phase filters*, set the circularly advanced filtered series $\{Y_t\}$ by ν units as:

$$Y_t^\nu \equiv Y_{t+\nu \bmod N}, \qquad t = 0, \ldots, N-1,$$

where ν is an integer such that $1 \leq |\nu| \leq N-1$. Then,

$$\begin{aligned} Y_t^\nu \equiv Y_{t+\nu \bmod N} = \sum_{l=0}^{N-1} u_l^o x_{t+\nu-l \bmod N} &= \sum_{l=-\nu}^{N-1-\nu} u_{l+\nu}^o x_{t-l \bmod N} \\ &= \sum_{l=0}^{N-1} u_{l+\nu \bmod N}^o x_{t-l \bmod N}. \end{aligned}$$

That is, advancing the filtered series $\{Y_t\}$ by ν units corresponds to using a filter whose coefficients are circularly advanced by ν units $\{u_l^{(\nu)} \equiv u_{l+\nu} : l = \ldots, -1, 0, 1, \ldots\}$. The circular filter $\{u_{l+\nu \bmod N}^o : l = 0, \ldots, N-1\}$ are obtained by periodizing the filter $\{u_l^{(\nu)}\}$ to length N. Thus, the phase properties of $\{u_{l+\nu \bmod N}^o\}$ can be obtained from the transfer function of $\{u_l^{(\nu)}\}$, given by $U^{(\nu)}(\theta) \equiv U(\theta)e^{\mathrm{i}2\pi\theta\nu}$. If $\{u_l\}$ is a zero phase filter, then $U(\theta) = |U(\theta)|$ and $U^\nu(\theta) \equiv |U(\theta)|e^{\mathrm{i}2\pi\theta\nu}$. Hence, $\{u_l^{(\nu)}\}$ is a linear phase filter with phase function $\omega(\theta) = 2\pi f\nu$ for a real-valued constant ν. A linear phase filter can be easily changed to a zero phase filter, if ν is an integer.

As mentioned above, the matrix $\mathcal{W}$, which defines the DWT, depends on both the wavelet filter and scaling filter. A real valued *wavelet filter* $\{h_l : l = 0, \ldots, L-1\}$ of width L, where L must be an even integer and $h_0 \neq 0$, $h_{L-1} \neq 0$ and $h_l = 0 \; \forall \; l < 0$ and $l \geq L$, must satisfy the following three conditions:

1. $\sum_{l=0}^{L-1} h_l = 0$,
2. $\sum_{l=0}^{L-1} h_l^2 = 1$,
3. $\sum_{l=0}^{L-1} h_l h_{l+2n} = \sum_{l=-\infty}^{\infty} h_l h_{l+2n} = 0 \quad \forall n \in \mathbb{Z}_+$ (i.e. the wavelet filters are orthogonal to its even shifts).

The first condition indicates that the wavelet filter is associated with a difference operator. This in turn implies that the wavelet filter is a high-pass filter capturing the high frequency (small scale) components of the signal and measures the deviations from the smooth components. The first level wavelet coefficients $\mathbf{W}_1$ are obtained by circularly filtering the time series $\{x_t : t = 0, \ldots, N-1\}$ with $\{h_l\}$ and retaining every other value of the output, which is known as downsampling by two, as follows:

$$W_{1,t} = \sum_{l=0}^{L-1} h_l x_{2t+1-l \bmod N} = \sum_{l=0}^{N-1} h_l^o x_{2t+1-l \bmod N}, \qquad t = 0, \ldots, \frac{N}{2} - 1,$$

where $\{W_{1,t} : t = 0, \ldots, \frac{N}{2} - 1\}$ are the elements of $\mathbf{W}_1$ and $\{h_l^o : l = 0, \ldots, N-1\}$ is $\{h_l\}$ periodized to length N, i.e.

$$h_l^o = \begin{cases} h_l & 0 \leq l \leq L-1, \\ 0 & L \leq l \leq N-1. \end{cases}$$

While the wavelet filter $\{h_l\}$ is used to form the first $\frac{N}{2}$ rows of the DWT matrix $\boldsymbol{\mathcal{W}}$, the scaling filter is used to construct the $\frac{N}{2} * N$ matrix of $\boldsymbol{\mathcal{V}}_1$ which forms the last $\frac{N}{2}$ rows of $\boldsymbol{\mathcal{W}}$. On the contrary, the scaling filter is a low-pass filter that captures the low frequency (large scale) components of the series.

The *scaling filter* $\{g_l\}$ is also known as the quadrature mirror filter that corresponds to $\{h_l\}$, since: $g_l = (-1)^{l+1} h_{L-1-l}$ and $h_l = (-1)^l g_{L-1-l}$. A scaling filter $\{g_l\}$ must satisfy the following three conditions:

1. $\sum_{l=0}^{L-1} g_l = \sqrt{2}$ (or $-\sqrt{2}$),
2. $\sum_{l=0}^{L-1} g_l^2 = 1$,
3. $\sum_{l=0}^{L-1} g_l g_{l+2n} = \sum_{l=-\infty}^{\infty} g_l g_{l+2n} = 0 \quad \forall n \in \mathbb{Z}_+$.

The rows of $\boldsymbol{\mathcal{V}}_1$ (the scaling coefficients associated with unit scale) are obtained by circularly filtering the time series $\{x_t, t = 0, \ldots, N-1\}$ with $\{g_l\}$ and retaining every other value of the output as follows:

$$V_{1,t} = \sum_{l=0}^{L-1} g_l x_{2t+1-l \bmod N} = \sum_{l=0}^{N-1} g_l^o x_{2t+1-l \bmod N}, \qquad t = 0, \ldots, \frac{N}{2} - 1,$$

where $\{V_{1,t}, t = 0, \ldots, \frac{N}{2} - 1\}$ are the elements of $\mathbf{V}_1$ and $\{g_l^o : l = 0, \ldots, N-1\}$ is $\{g_l\}$ periodized to length N.

After filtering the time series $\mathbf{X}$ of length $N = 2^J$ into the $\frac{N}{2}$ first level wavelet coefficients $\mathbf{W}_1$ and the $\frac{N}{2}$ first level scaling coefficients $\mathbf{V}_1$, the remaining wavelet and scaling coefficients are obtained recursively in $J-1$ steps, forming the pyramid algorithm (see Figure 1.6). At each step j, $j = 2, \ldots, J$, the vector $\mathbf{V}_{j-1}$ of length $\frac{N}{2^{j-1}}$ is filtered into the vectors $\mathbf{W}_j$ and $\mathbf{V}_j$ which are of length $\frac{N}{2^j}$ each. That is, at every j^{th} step, the vector $\mathbf{V}_{j-1}$ is treated as $\mathbf{X}$ and the elements of $\mathbf{V}_{j-1}$ are filtered using $\{h_l\}$ and $\{g_l\}$, then the filter outputs are sub-sampled by 2^j to form $\mathbf{W}_j$ and $\mathbf{V}_j$. The elements of $\mathbf{W}_j$ are the wavelet coefficients at level j and the elements of $\mathbf{V}_j$ are the scaling coefficients at level j. Finally, the DWT coefficients vector $\mathbf{W}$ is formed by the $J+1$ vectors $\mathbf{W}_1, \ldots, \mathbf{W}_J, \mathbf{V}_J$.

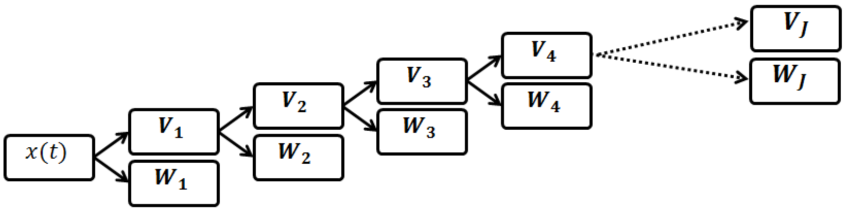


FIGURE 1.6: Pyramid Algorithm.

The filtering operation in the DWT involves circular filtering, where the signal is treated as a portion of a larger periodic sequence with period N. This could create problems at both ends of the series. The coefficients affected by circularity assumptions are known as boundary coefficients. The number of boundary coefficients increases as the filter width increases and also as the scale increases. One way to delineate the boundary regions, where the coefficients are affected by circular shifting due to their proximity to the boundaries of the series, is to put vertical lines on the plots of DWT coefficients and the plots of details and smooth MRA components. The boundary problems arise in two occasions. The first occurs when the DWT is used with a non-dyadic series, i.e. a series of length $N \neq 2^j$. In this case, the boundary problems can be solved by removing observations or completing the series to be dyadic. The second occasion concerns the filtering operation, which has to be applied on all observations and requires $L-1$ observations to be available before time t. In this case, completing the series is the only possible solution. Completing the series can be done by padding the ends of the series with zeros, or fitting a polynomial model to replace the required data at the ends of the series, or mirroring the last observations to complete each end of the series.

1.5.3 Choice of Filter

The choice of the wavelet/scaling filter depends on the features of the signal under study and the aim of the analysis. Based on the properties of the series and the purpose of the analysis, it might be interesting to choose a filter with one or more of the following desired properties.

- *Symmetry*: Symmetric filters ensure no phase shift in the output series. Most of the filters are asymmetric. However, this property is less desired if the maximal overlap discrete wavelet transform (See Section 1.5.5) is used, as it already ensures that the original series and its filter coefficients are aligned.
- *Orthogonality*: Orthogonal filters ensure that the wavelet and scaling coefficients contain different information. The Daubechies and Least Asymmetric wavelets (See Section 1.5.4) are orthogonal.
- *Smoothness*: The degree of smoothness is measured by the number of continuous derivatives of the basis function. A smooth wavelet is required for a smooth time series and the smoothness of the filter increases as the filter length increases.
- *Number of vanishing moments*: The number of vanishing moments is computed as half of the filter length. This number has an impact on the ability of the wavelet to account for the behaviour of the signal.

The Daubechies family of filters, presented in the following section, is a very popular choice in time series analysis. Choosing the filter width L is as crucial as choosing the wavelet filter. The choice of the wavelet filter and its width is a trade-off between introducing undesirable artefacts by choosing wavelet filters of short width and increasing the computations and the number of boundary coefficients by choosing filters of large width. The best strategy is to start with a preliminary filter of small width (2 to 6) and keep increasing the width until the analysis becomes free of artefacts.

1.5.4 Daubechies Filters

Daubechies (1992) proposed a family of filters that provides a DWT, which can be easily interpreted as differences of adjacent averages. The class of Daubechies filters is defined by the squared gain function $\mathcal{G}^{(D)}(.)$ of the associated scaling filters. As the filter width L increases, the squared gain function $\mathcal{G}^{(D)}(.)$ converges to the squared gain function of an ideal low-pass filter and the number of scaling filters sequences $\{g_l : l = 0, \ldots, L-1\}$

with the same $\mathcal{G}^{(D)}(.)$ increases. These filters are distinguished by their phase functions $\omega^{(G)}(.)$ in the polar representation of their transfer functions $G(.)$ given by:

$$G(\theta) \equiv \left[\mathcal{G}^{(D)}(\theta)\right]^{1/2} e^{i\omega^{(G)}(\theta)}.$$

This results in a number of sequences $\{g_l\}$ with different phase functions $\omega^{(G)}(.)$ to choose from. Daubechies (1992) suggested a couple of choices for the best sequence of filters to use.

1. Daublets D(L) ($L = 2, 4, \ldots$), where the external phase filter of width L is the scaling filter denoted by $\{g_l^{(ep)}\}$ satisfying:

   $$\sum_{l=0}^{m} g_l^2 \leq \sum_{l=0}^{m} [g_l^{(ep)}]^2, \qquad \text{for} \qquad m = 0, \ldots, L-1,$$

   for any other filter $\{g_l\}$ having the same squared gain function $\mathcal{G}^{(D)}(.)$.

2. Symmlets LA(L) ($L = 8, 10, \ldots$), where the least asymmetric filter of width L is the scaling filter denoted by $\{g_l^{(la)}\}$ whose phase function $\omega^{(G)}(.)$ is as close as possible to that of a linear phase filter.

The LA filters have approximately linear scaling and wavelet phase functions:

$$\omega^{(G)}(\theta) \approx 2\pi\theta\nu \qquad \text{and} \qquad \omega^{(H)}(\theta) \approx -2\pi\theta(L - 1 + \nu),$$

where ν is an odd negative integer. Hence, approximate zero phase scaling and wavelet filters can be obtained by circularly advancing the filtered series by $|\nu|$ and $|L - 1 + \nu|$ units, respectively. Therefore, the least asymmetric filter, LA(.), is a very common choice in time series analysis to obtain DWT coefficients that align in time with the original signal.

An alternative to the DWT known as the maximal overlap discrete wavelet transform, which does not require padding or truncating the time series when its length is not a power of two, is presented in the following section.

1.5.5 Maximal Overlap Discrete Wavelet Transform

The Maximal Overlap Discrete Wavelet Transform (MODWT) is a modified version of the DWT, which similarly decomposes the time series into a set of wavelet details plus a smooth component. The MODWT has some advantages over the DWT making it more preferable. First, the MODWT is well defined for any sample size N that is not a power of two. Second, the MODWT coefficients and its associated MRA are not affected by the choice of the starting point of the time series or the choice of the wavelet filter. Third, the MODWT details and smooths are associated with zero phase filters, which easily makes the features in the MRA align with the events in the original signal. However, the penalties for these advantages are that it is a highly redundant nonorthogonal transform and computationally intensive.

The MODWT maintains a complete resolution of the signal at each frequency/scale. The MODWT coefficients are obtained by circularly filtering the time series $\mathbf{X}$ with the MODWT wavelet filter $\{\tilde{h}_l\}$ and scaling filter $\{\tilde{g}_l\}$ without downsampling. That is, the MODWT of level J_o yields the wavelet coefficients vectors $\widetilde{\mathbf{W}}_1, \ldots, \widetilde{\mathbf{W}}_{J_o}$ and the scaling coefficients vector $\widetilde{\mathbf{V}}_{J_o}$, each of dimension N. The MODWT wavelet filter $\{\tilde{h}_l\}$ and scaling filter $\{\tilde{g}_l\}$ are defined by $\tilde{h}_l = h_l/\sqrt{2}$ and $\tilde{g}_l = g_l/\sqrt{2}$, respectively. The t^{th} elements of the vectors $\widetilde{\mathbf{W}}_1$ and $\widetilde{\mathbf{V}}_1$ are obtained by circularly filtering $\{x_t\}$ with the MODWT wavelet filters $\{\tilde{h}_l\}$ and scaling filters $\{\tilde{g}_l\}$ as follows:

$$\widetilde{W}_{1,t} = \sum_{l=0}^{L-1} \tilde{h}_l x_{t-l \bmod N} \quad \text{and} \quad \widetilde{V}_{1,t} = \sum_{l=0}^{L-1} \tilde{g}_l x_{t-l \bmod N}, \qquad t = 0, \ldots, N-1.$$

By analogy to the DWT multiresolution analysis, the original signal $\mathbf{X}$ can be recovered as the sum of detail and smooth components as follows:

$$\mathbf{X} = \sum_{j=1}^{J_o} \widetilde{\mathbf{D}}_j + \widetilde{\mathbf{S}}_{J_o},$$

where J_o is the maximum level of decomposition. The t^{th} element of the vector $\widetilde{\mathbf{D}}_j$ is obtained by:

$$\widetilde{D}_{j,t} = \sum_{l=0}^{L-1} \tilde{h}_l \widetilde{W}_{j,t+l \bmod N} = \sum_{l=0}^{N-1} \tilde{h}_l^o \widetilde{W}_{j,t+l \bmod N}, \qquad j = 1, \ldots, J_o \;, \; t = 0, \ldots, N-1.$$

where $\{\tilde{h}_l^o : l = 0, \ldots, N-1\}$ is $\{\tilde{h}_l : l = 0, \ldots, L-1\}$ periodized to length N. An analogous statement holds for the smooth component.

In this thesis, the MODWT is used instead of the DWT and from the Daubechies family of wavelets the least asymmetric filter of width 8, LA(8), is chosen to filter the EpCO$_2$ signal. The least asymmetric filter is chosen since it has a zero phase filter allowing the association of the MODWT coefficients with the actual times. The filter width 8 is selected as it yields similar analysis as larger widths but with smaller number of coefficients affected by boundary conditions.

1.5.6 Analysis of Variance

The variance of the signal can be decomposed into different scale/frequency components based on the wavelet and scaling coefficients of the MODWT. Thus, we have:

$$\| \mathbf{X} \|^2 = \sum_{t=0}^{N-1} x_t^2 = \sum_{j=1}^{J_0} \widetilde{\mathbf{W}}_j^2 + \widetilde{\mathbf{V}}_{J_0}^2,$$

where $\| \mathbf{X} \|^2$ is the energy of the signal $\mathbf{X}$ and $\widetilde{\mathbf{W}}_j$ are the wavelet coefficients measuring the deviations of the signal $\mathbf{X}$ from its long run mean at the different scales, whereas $\widetilde{\mathbf{V}}_{J_0}$ is the vector of the scaling coefficients equal to the sample mean of the signal $\mathbf{X}$.

Consider the time series $\{x_t, t = 0, \ldots, N-1\}$ and let $\{\tilde{h}_{j,l} : l = 0, \ldots, L_j - 1\}$ denotes the j^{th} level MODWT wavelet filter and $L_j = (2^j - 1)(L-1) - 1$ be the filter width, associated with scale $\tau_j = 2^{j-1}$. Then, the circularly filtered series at scale τ_j is given by:

$$\widetilde{W}_{j,t} = \sum_{l=0}^{L_j - 1} \tilde{h}_{j,l} x_{t-l \bmod N}, \qquad t = 0, \ldots, N-1.$$

The time-independent wavelet variance for scale τ_j is defined as the variance of the $\widetilde{W}_{j,t}$, i.e.:

$$v_{\mathbf{X}}^2(\tau_j) = \mathrm{VAR}\{\widetilde{W}_{j,t}\}.$$

The wavelet variance decomposes the variance of the time series on a scale by scale basis, and hence the sample variance of the time series, $\sigma_{\mathbf{X}}^2$, is obtained by:

$$\sigma_{\mathbf{X}}^2 = \sum_{j=1}^{\infty} v_{\mathbf{X}}^2(\tau_j).$$

The time-independent wavelet variance at scale τ_j is estimated by:

$$\hat{v}_{\mathbf{X}}^2(\tau_j) = \frac{1}{M_j} \sum_{t=L_j-1}^{N-1} \widetilde{W}_{j,t}^2, \tag{1.9}$$

where $M_j = N - L_j - 1$ denotes the number of boundary coefficients at scale τ_j. This makes $\hat{v}_{\mathbf{X}}^2(\tau_j)$ an unbiased estimator of the wavelet variance. Accordingly, the sample variance of the time series $\hat{\sigma}_{\mathbf{X}}^2$ can be estimated by:

$$\hat{\sigma}_{\mathbf{X}}^2 = \sum_{j=1}^{\infty} \hat{v}_{\mathbf{X}}^2(\tau_j). \tag{1.10}$$

However, this estimate could provide misleading results if the filter width is too small.

Under the assumption that $\{\widetilde{W}_{j,t}\}$ is a Gaussian stationary process with mean zero, the estimator $\hat{\upsilon}^2_{\mathbf{X}}(\tau_j)$ is asymptotically normally distributed with mean $\upsilon^2_{\mathbf{X}}(\tau_j)$ and variance $2R_j/M_j$, where $R_j = \int_{-1/2}^{1/2} \tilde{f}_j^2(\theta)d\theta$ with $\tilde{f}$ being the spectral density function of $\{\widetilde{W}_{j,t}\}$. Based on this result, the confidence interval for $\upsilon^2_{\mathbf{X}}(\tau_j)$ might have a negative lower confidence limit. Therefore, $\hat{\upsilon}^2_{\mathbf{X}}(\tau_j)$ is re-normalized to obey a chi-squared distribution such that:

$$\frac{\eta\hat{\upsilon}^2_{\mathbf{X}}(\tau_j)}{\upsilon^2_{\mathbf{X}}(\tau_j)} \sim \chi^2_\eta,$$

where $\eta = \frac{2\mathbb{E}^2\{\hat{\upsilon}^2_{\mathbf{X}}(\tau_j)\}}{\text{VAR}\{\hat{\upsilon}^2_{\mathbf{X}}(\tau_j)\}}$ is the *equivalent degrees of freedom* adjusted to account for the correlation between $\hat{\upsilon}^2_{\mathbf{X}}(\tau_j)$. This equivalent degrees of freedom can be estimated by either $\hat{\eta}_1 = \frac{M_j\hat{\upsilon}^4_{\mathbf{X}}(\tau_j)}{R_j}$ or $\hat{\eta}_2 = \max\{M_j/2^j, 1\}$. $\hat{\eta}_1$ provides a good approximation if the number of coefficients at each scale τ_j is large enough (i.e. $M_j > 128$), while $\hat{\eta}_2$ is reasonably accurate for smaller sample sizes. The $100(1-2\alpha)\%$ confidence interval for $\upsilon^2_{\mathbf{X}}(\tau_j)$ is then approximated by:

$$\left[\frac{\eta\hat{\upsilon}^2_{\mathbf{X}}(\tau_j)}{Q_n(1-\alpha)}, \frac{\eta\hat{\upsilon}^2_{\mathbf{X}}(\tau_j)}{Q_n(\alpha)}\right],$$

where $Q_n(\alpha)$ is the α^{th} quantile of a χ^2_η.

The above results rely on the assumption of constant variability over time, which might not hold if the time series is non-stationary. In this case, the time-dependent wavelet variance at scale τ_j can be estimated by:

$$\hat{\upsilon}_{\mathbf{X},t}(\tau_j) = \frac{1}{N_s}\sum_{u=-(N_s-1)/2}^{(N_s-1)/2} \widetilde{W}^2_{j,t+|\nu_j^{(H)}|+u \bmod N}, \tag{1.11}$$

where $\widetilde{W}_{j,t+|\nu_j^{(H)}|+u \bmod N}$ are the $\widetilde{W}_j$ circularly advanced by $|\nu_j^{(H)}|$ units to align the events with the original time series. The choice of N_s relies on the nature of data; for example, a natural choice for monthly data would be $N_s = 12$. Similarly to the time-independent wavelet variance, a $100(1-2\alpha)\%$ confidence interval for the time-dependent variance can be obtained.

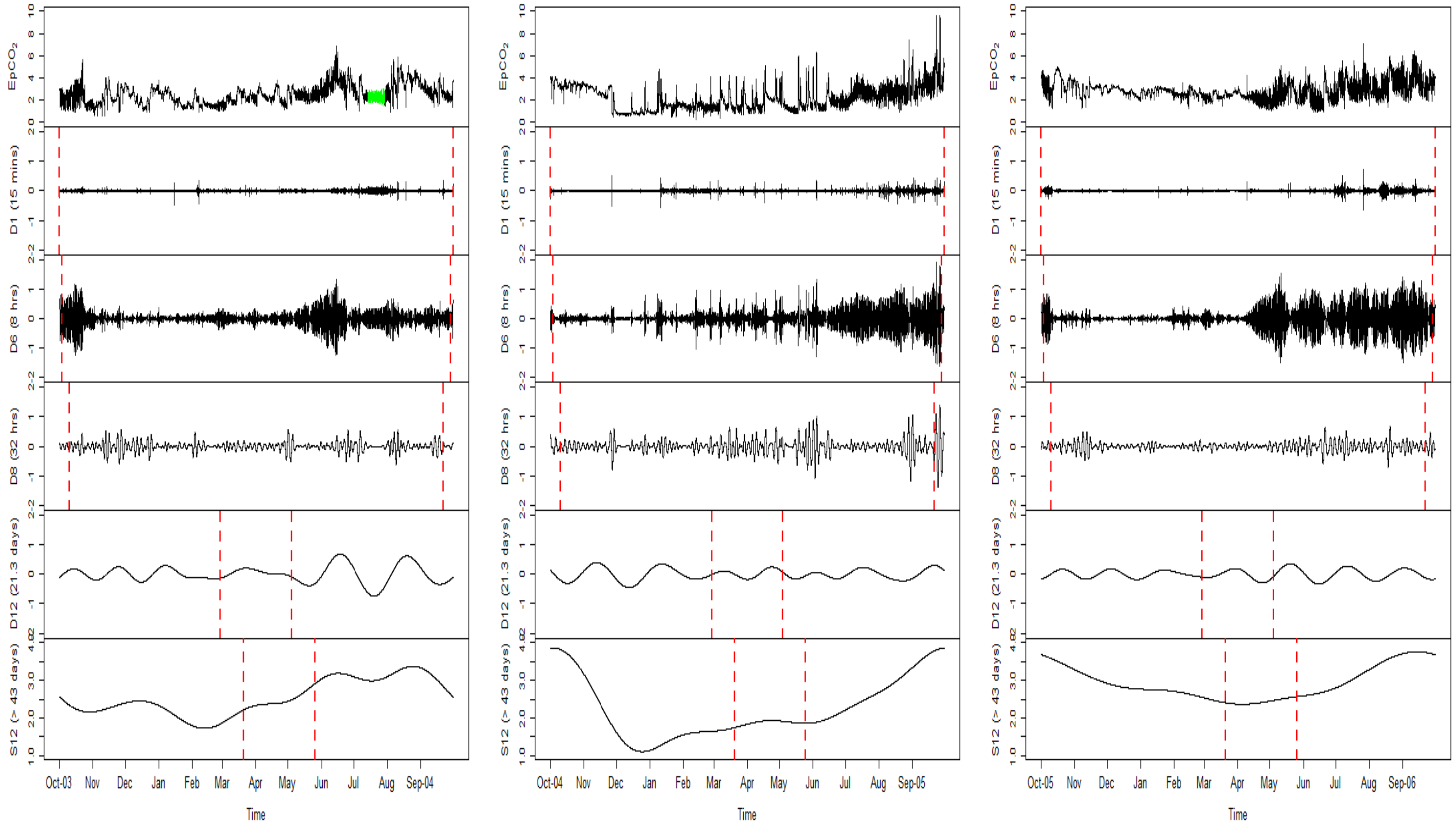
1.5.7 Wavelets Analysis for the EpCO$_2$ Data

Wavelet analysis has been applied to the high-frequency time series of EpCO$_2$ as a means of exploring its temporal variations over the different timescales. As explained above, the result of wavelet analysis is a time-scale decomposition of the original signal, which helps identify the cyclical components over different frequencies, as well as the long-term trend (Nason, 2008, Percival and Walden, 2006). The wavelet transform cannot be applied to time series with missing data. The EpCO$_2$ series in 2003/2004 has 1544 missing values in total, one in February 2004 and the rest in July 2004, which represent 4.4% of the total record. Each of the pH, temperature, and specific conductivity series of 2003/2004 and 2004/2005 has less than 5% missing values of the total record. All the missing values are first imputed using linear interpolation. The interpolation is done separately for each month and for each time within the month to better reproduce the variability of the series; i.e. missing values at 11:00 am in June are imputed using only recorded values at 11:00 am in June from the same HY, and so on. To induce some randomness, normally distributed random errors are then added to the interpolated values. These imputed values are shown in green in Figure 1.7. This imputation has been done using the R function `approxfun`.

According to Section 1.5.5, MODWT is preferred over DWT since it does not have any restrictions on the series length; N does not have to be necessarily equal to 2^J. Therefore, the MODWT is used to decompose the original EpCO$_2$ signal of each hydrological year and the associated time series of temperature, discharge, pH and conductivity into the different timescales. The MRA will then be based on the MODWT rather than the DWT, and hence the choice of the wavelet filter is not vital. But since it is of interest to align events in time, the least asymmetric filter with width equal to 8, LA(8), is chosen to circularly filter the original high-frequency signal. As previously explained, the least asymmetric filters form a special class of the Daubechies filters with phase function very close to that of a linear phase filter, making it easy to line up features in the filtered series with the original series (Percival and Walden, 2006). A filter width equal to 8 provides a good smooth representation of the corresponding time series and is chosen after comparing a series of wavelet transforms obtained for a range of filter width values. The wavelet transform corresponding to smaller filter width values resulted in sharp peaks in the individual elements of the time series decomposition, and greater width values did not make any difference. The `wmtsa` R package developed for wavelet analysis by Constantine and Percival in 2013 is used to obtain the MRA of the studied time series via the MODWT of the corresponding series.

The MODWT with LA(8) filter decomposes the EpCO$_2$ series for each of the hydrological years into 12 wavelet details and one smooth component, $\mathbf{X} = \sum_{j=1}^{12} \mathbf{D}_j + \mathbf{S}_{12}$, where 12 is the maximum number of scales. The wavelet details ($\mathbf{D}_j$, $j = 1, \ldots, 12$) reflect changes in the original series over scales of $15(2^{j-1})$ minutes and the smooth component ($\mathbf{S}_{12}$) is a smooth version of the data representing the overall trend and relates to variations over $\sim$43 days and higher.

Figure 1.8 illustrates the estimated time-independent wavelet variances $\hat{\upsilon}^2_{\mathbf{X}}(\tau_j)$ along with the corresponding confidence intervals against the corresponding scales τ_j, $j =$

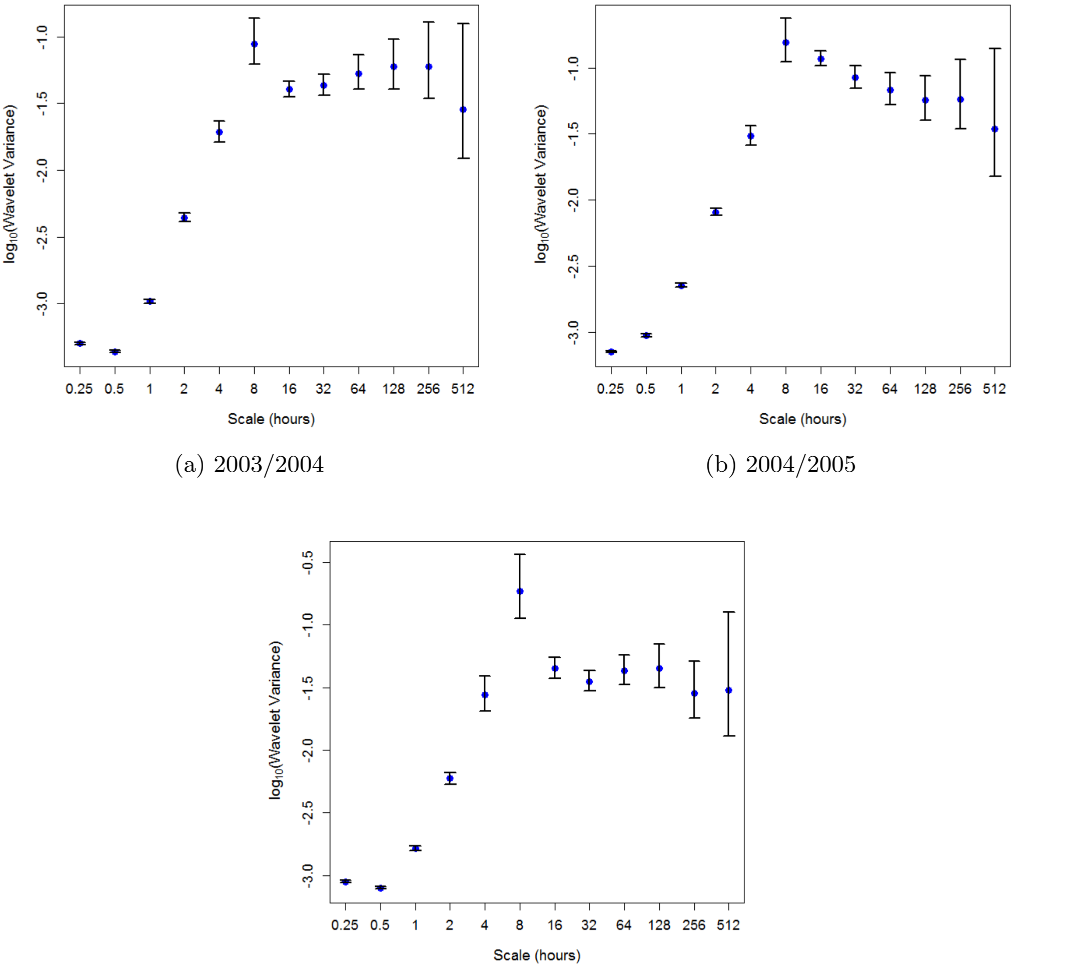


(a) 2003/2004

(b) 2004/2005

(c) 2005/2006

FIGURE 1.7: Multi-resolution analysis of EpCO$_2$ series in the HYs (a) 2003/2004, (b) 2004/2005 and (c) 2005/2006. The wavelet details $\mathbf{D}_1$ (15 mins), $\mathbf{D}_6$ (8 hrs), $\mathbf{D}_8$ (32 hrs) and $\mathbf{D}_{12}$ ($\sim$ 22 days) are on the same scale, different from the original series (top) and the smooth component $\mathbf{S}_{12}$ ($\geq$ 44 days) (bottom). The Red dashed vertical lines indicate the areas affected by boundary conditions. The green lines refer to the imputed values.



(a) 2003/2004

(b) 2004/2005

(c) 2005/2006

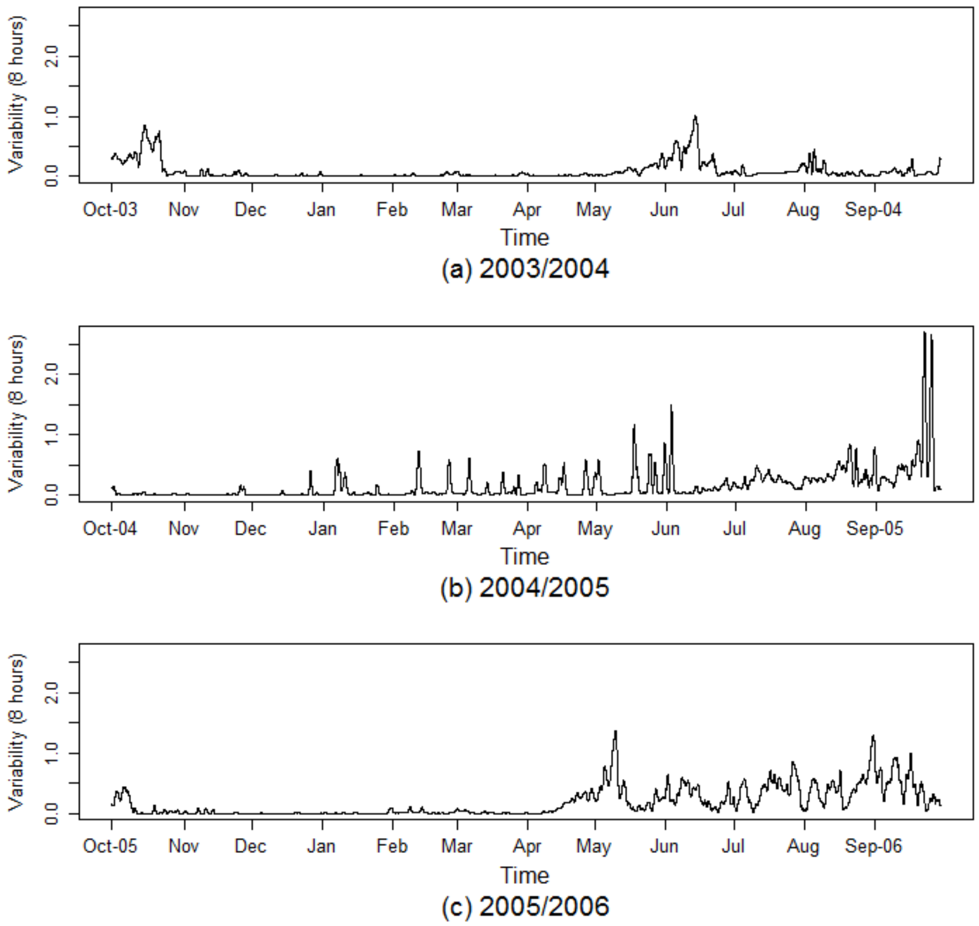
FIGURE 1.8: Time-independent wavelet variance of the EpCO$_2$ series in the HYs (a) 2003/2004, (b) 2004/2005 and (c) 2005/2006 for the scales $\frac{15}{60}\tau_j$, $j = 1, \ldots, 12$. Confidence intervals (black bars) are based on χ^2 distribution with $\hat{\eta}_1$ degrees of freedom.

$1, , \ldots, 12$. At each scale τ_j, the degrees of freedom of the chi-square distribution used to construct the confidence intervals are estimated using $\hat{\eta}_1$ since the number of coefficients M_j is greater than 128 (see Section 1.5.6). It is evident that the 6^{th} wavelet detail is the major contributor to the sample variance of the EpCO$_2$, reflecting changes in the time series over the scale of $15/60(2^{6-1}) = 8$ hours. That is, values in $\mathbf{D}_6$ are expected to be large when an average over 8 hours differs from values surrounding it. This feature is considered representative of the daylight cycle, since it reflects changes over a scale of 8 hours. The sample variance of the EpCO$_2$ series $\hat{\sigma}^2_{\mathbf{X}}$ (given by Equation 1.10) is estimated to be 0.4, 0.62 and 0.45 in the hydrological years 2003/2004, 2004/2005 and 2005/2006, respectively. This shows that the EpCO$_2$ of the hydrological year 2004/2005 has the highest variability.

The MRA of the EpCO$_2$ series for the wavelet details $\mathbf{D}_1$, $\mathbf{D}_6$, $\mathbf{D}_8$ and $\mathbf{D}_{12}$, and the smooth component $\mathbf{S}_{12}$ is displayed in Figure 1.7. Each of the detail series $\mathbf{D}_1$, $\mathbf{D}_6$, $\mathbf{D}_8$, $\mathbf{D}_{12}$ reflects changes in the original series on a scale of 15 minutes, 8 hours, 32 hours and $\sim$22 days, respectively. While, the wavelet smooth $\mathbf{S}_{12}$ represents the trend and relates to changes over 43 days and higher. The MRA indicated that the detail components $\mathbf{D}_j$, $j = 1, \ldots, 4$ (only $\mathbf{D}_1$ is shown here due to space limitations) are the least variable reflecting small scale variability and can be related to weather or hydrological events. Therefore, these high-frequency scales capture the uncommon EpCO_2 levels which might be influenced by short-lived changes in the water hydrology such as intense periods of rainfall. The detail component $\mathbf{D}_6$ is the main contributor to the sample variance of the EpCO_2 series (see Figure 1.8) and also the associated temperature series (not shown here) reflecting the presence of an intra-daily cycle related to the daylight cycle. This diel cycle is not constant throughout each hydrological year and larger fluctuations are observed during summer when a pronounced daily cycle is present. This MRA shows that the EpCO_2 of the dry summer of the hydrological year 2005/2006 exhibits this diel pattern clearly for a longer time period compared to the wetter summers of the HYs 2003/2004 and 2004/2005, reflecting different balances of external (e.g. climatological) and internal (biological processing) drivers of EpCO_2. The wavelet detail $\mathbf{D}_8$ is the second major source of variability after $\mathbf{D}_6$ and can be seen as an approximation for the daily (day to day) variations. Whilst, $\mathbf{D}_{12}$ reflects changes over nearly 22 days and hence can be considered as a proxy for the monthly variability. The wavelet smooth $\mathbf{S}_{12}$ of the EpCO_2 series of the HYs 2003/2004 and 2005/2006 shows a decrease in the overall trend of EpCO_2 reaching its minimum in March-April and followed by a gradual increase until the beginning of September. However, in the hydrological year 2004/2005, the EpCO_2 reaches its minimum earlier in December before it continues to increase until the end of the hydrological year. The wavelet smooth $\mathbf{S}_{12}$ appears to repeat the same pattern in all hydrological years but with different magnitude; and hence can be regarded as the seasonal component of the full series.

Figure 1.9 shows the time-dependent wavelet variance of the wavelet detail $\mathbf{D}_6$, the major contributor to the sample variance of the EpCO_2 series. The figure emphasizes the non-constant variance of the component over time, with more variability generally observed in summer. The intra-daily variability is considerably higher during the summer season of the last two HYs compared to that of 2003/2004. However, in the hydrological year 2004/2005, some intermittent increases in the variability are evident at the end of winter and during spring, possibly reflecting changes in in the external drivers (e.g. heavy rainfall) of resultant EpCO_2. It is also noticed that the highest variability in the intra-daily cycle of EpCO_2 occurs at the end of the second hydrological year.

To gain a better understanding of the intra-daily variability of EpCO_2 in relation to the changes in the other hydrological variables, the 6^{th} wavelet detail series of the MRA of



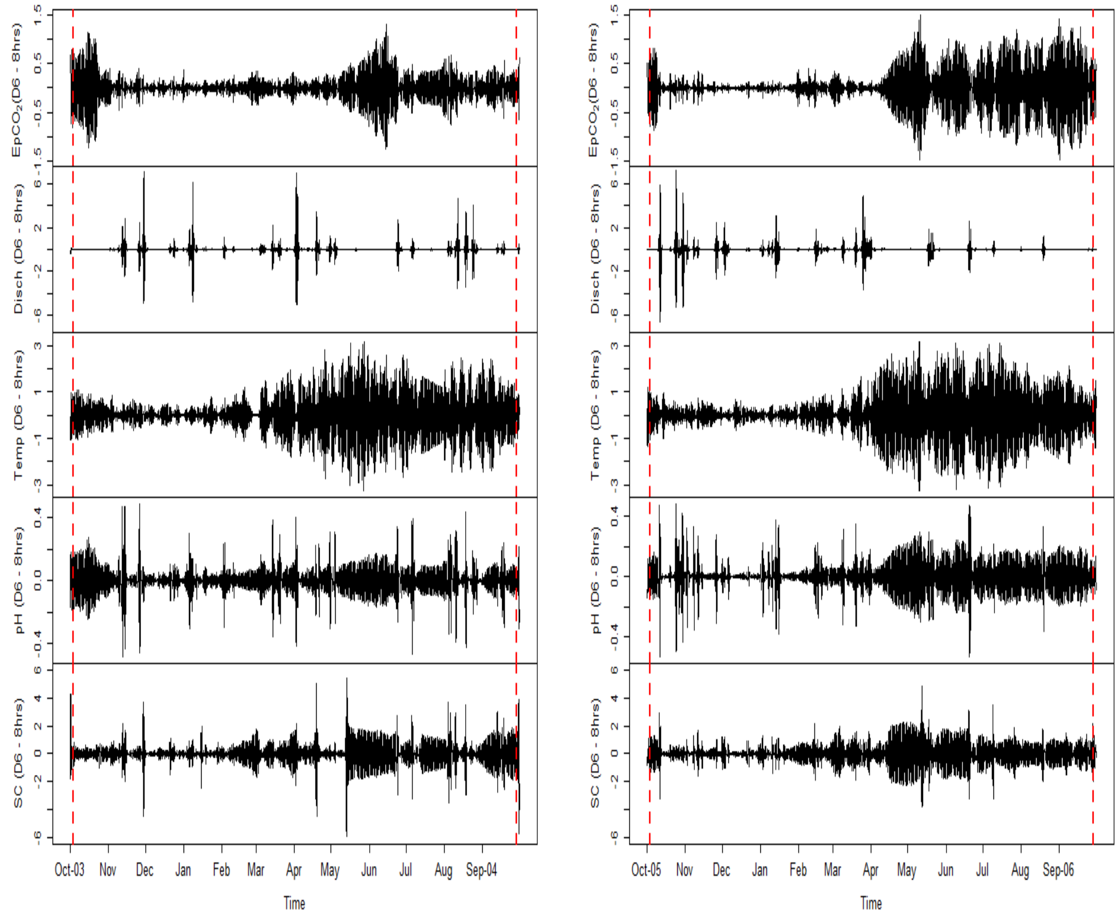
(a) 2003/2004

(b) 2004/2005

(c) 2005/2006

FIGURE 1.9: Time-dependent wavelet variability for the intra-daily cyclic component $\mathbf{D}_6$ in the HYs (a) 2003/2004, (b) 2004/2005 and (c) 2005/2006.

the different hydrological variables was obtained. Figure 1.10 compares the 6^{th} wavelet detail series, representing changes over a scale of 8 hours, of the different hydrological variables and the $EpCO_2$ series. The timing, extent and number of occurrences of hydrological events differ from one HY to another. The highest $EpCO_2$ variability is usually associated with little changes in discharge, consistent with internal fluvial carbon cycling, while hydrological events are associated with compressed $EpCO_2$ variability. The periods when pH and SC show most change occur with flow events. The variability in $EpCO_2$ evolves coherently with the variability in temperature, in itself a proxy for seasonality: $EpCO_2$ appears to be more variable during summer when there are larger fluctuations between day and night temperatures. The changes in temperature and discharge influence the SC and pH, which in turn influence the $EpCO_2$ across the different years.



(a) 2003/2004 (b) 2005/2006

FIGURE 1.10: Plot of the 6th wavelet detail (8 hrs scale) of MRA of $EpCO_2$ (top), flow, temperature, conductivity and pH (bottom) series in the HYs (a) 2003/2004 and (b) 2005/2006. The dashed vertical lines indicate the areas affected by boundary conditions.

In brief, the primary EDA and wavelet analysis highlighted the seasonal and diurnal fluctuations of $EpCO_2$ and the changes in these variations within and between the individual hydrological years. They also revealed that the hydrodynamics might contribute to part of the $EpCO_2$ variability although the nature of these relationships is very complex and difficult to explore and visualize through exploratory tools. It is not yet clear from the above exploratory analysis whether or not the temporal patterns in $EpCO_2$ can be described entirely or partially by hydrology. Therefore, advanced modeling of $EpCO_2$ is required to better describe and analyze the variations in $EpCO_2$ and its relationship with water hydrology. This will be the main focus of the following chapter.

1.6 Aims and Objectives

This thesis aims at using existing statistical methods and developing new statistical tools to effectively and efficiently explore and analyze hydrological high-frequency time series.

This aim will be achieved by first investigating the challenges of modeling and analyzing hydrological high-frequency data using standard statistical methods, then using and developing more advanced statistical techniques to explore and analyze such hydrological high-frequency data taking into account the identified challenges. Throughout the thesis, the high-frequency $EpCO_2$ data described above are used as an illustrative data set. The primary exploratory analysis of $EpCO_2$ data, presented in Sections 1.3 and 1.5, have shown the evidence of inter-annual and intra-annual variations in the $EpCO_2$. In addition, they have highlighted an $EpCO_2$ intra-daily pattern that varies over time, possibly a result of changes over time in the river physical characteristics or climatological conditions. Identifying the different intra-daily patterns of $EpCO_2$ and understanding the drivers of each pattern are of interest to ecologists. This has motivated the main objectives of this thesis:

1. Investigating the different complexities of modeling high-frequency time series data and the effect of persistent correlation on statistical modeling;
2. Using and developing efficient statistical methods to identify the different daily patterns of $EpCO_2$ and the climatological and hydrological drivers of each pattern, taking into account the nature of such high-frequency time series and the complex structure they involve;
3. Developing statistical procedures to investigate whether and how the covariance structure/spectral density of a high-frequency data evolve over time;
4. Developing flexible statistical techniques that account for the dynamic changes in the covariance structure/spectral density to provide a more adequate dimension reduction for non-stationary high-frequency time series.

1.7 Thesis Structure

The thesis consists of 6 chapters including this introductory chapter. The remainder of the thesis is structured as follows:

Chapter 2 continues to explore the challenges of modeling and analyzing high-frequency data which arise from environmental applications using the $EpCO_2$ data, described earlier, as an illustrative data set. The chapter introduces non-parametric regression and generalized additive models as common methods of modeling non-linear and complex patterns and relationships over time. Using the above results of wavelet analysis, a set of additive models is fitted to better describe the temporal variability in $EpCO_2$ and its relationship with river hydrology.

Chapter 3 focuses on efficiently analyzing the identified intra-daily pattern of $EpCO_2$ using a functional data analysis approach (Ramsay and Silverman, 1997). The main objective of this chapter is to explore the main sources of variability across the different daily patterns using functional principal component analysis and try to classify these daily patterns into a set of clusters and determine the internal and external drivers of each class using functional clustering techniques.

Chapter 4 explores the effect of taking into account the temporal correlation using a frequency domain approach on reducing the data dimensionality and on identifying the main sources of variability in $EpCO_2$, by introducing the recently proposed methodology of dynamic functional principal components (Hormann et al., 2014). This methodology has been extended in this chapter to temporally dependent functional data estimated using a wider range of basis functions. The chapter then presents a novel clustering approach based on the dynamic scores to classify temporally correlated data, while accounting for the auto-correlation and disregarding any fine fluctuations between individual functional observations. The results of classifying the intra-daily patterns of $EpCO_2$ using this methodology is presented later in the chapter.

Chapter 5 extends the methodology presented in Chapter 4 to account simultaneously for the auto-correlation and non-stationarity in the time series, by allowing the dynamic functional principal components to vary smoothly over time. The effectiveness of this methodology has been investigated through an extensive simulation study. Based on these time-varying dynamic functional principal components, an inferential procedure is also suggested to test the stationarity of the functional time series. Subsequently, the chapter proposes an extension to the clustering method proposed in Chapter 4 and compares the results of clustering the $EpCO_2$ daily patterns using both approaches.

Chapter 6 summarizes and briefly discusses the main findings of the thesis. The chapter discusses the challenges and drawbacks of analyzing hydrological high-frequency data using standard time series analysis techniques and the benefits of using the proposed methodology to provide a more adequate representation of the data after adjusting for both auto-correlation and non-stationarity. Finally, it highlights the limitations of the used methodology and suggests further directions for future work.

Chapter 2

High-Frequency Time Series Modeling

As noted in Chapter 1, time series can be analyzed using either a time domain approach or a frequency domain approach. The main focus of Chapter 1 was to explore and visualize the dynamics of a high-frequency hydrological data set in both time and frequency domains using the available data of EpCO_2. With the aid of wavelets, which have the advantage of analyzing a time series in both time and frequency domains simultaneously, seasonal and diurnal fluctuations of EpCO_2 were identified. This initial exploratory analysis has also shown that these variations are not constant and vary over time and that the water hydrodynamics co-vary with the EpCO_2 and its temporal patterns. However, the nature of these relationships is highly dynamic and difficult to explore using ordinary exploratory analysis or wavelets. Moreover, the wavelet analysis has failed to account for the persistent temporal correlation between the high-frequency measurements. For these reasons, advanced flexible regression modeling is needed to gain better understanding of the temporal variations in EpCO_2 and its complex relationships with water hydrology in the time domain. This will be achieved using additive models, which provide a very useful framework that allows smoothing to be incorporated in a flexible regression structure (Hastie and Tibshirani, 1990).

Non-parametric flexible regression modeling, including additive models, has proven to be very useful in describing the pattern of many environmental phenomenon smoothly over time and space. Additive models have been widely used in the analysis of air pollution. A non-parametric additive model framework has been used by Bowman et al. (2009) to model the sulphur dioxide pollution smoothly across Europe in space and time; while Giannitrapani et al. (2011) have modeled the sulphur dioxide pollution data at one European site over time using an additive model with smooth terms of the different time components. Pearce et al. (2011) have also assessed the relationship between the daily air pollutant concentrations of PM_{10} and NO_2 and the local-scale meteorology in Melbourne, Australia after controlling for long-term trends, seasonality, weekly emissions, spatial variation, and temporal persistence using additive models. Additive models have also been extensively employed in the analysis of water quality trends and the description of relationships between water quality variables. For instance, McMullan et al. (2007) have fitted an additive model to identify the spatial and temporal trends as well as the factors affecting the water quality, measured by the dissolved oxygen, of the river Clyde. Interaction terms have been added to the model to assess how the effect of the explanatory variables on the dissolved oxygen change smoothly over space. Ferguson et al. (2008) have also employed an additive model with univariate and multivariate smooth terms to explore trends and seasonality for various water quality variables, one at a time, at Loch Leven. A range of additive models have also been developed to describe the kind of lakes suspectable to cyanobacterial blooms (high risk waterborne toxic biological substances) using the available explanatory variables, including water color, alkalinity, retention time and total phosphorus (Carvalho et al., 2011). The aim of this chapter is to model and assess the temporal variations in $EpCO_2$ in relation to the water hydrodynamics, using additive models.

Linear regression attempts to model the relationship between a variable of interest (Y) and its explanatory variable (X) by fitting the following linear equation:

$$y_i = \beta_o + \beta_1 x_i + \epsilon_i, \qquad i = 1, \dots, n,$$

where ϵ_i denotes an independent normally distributed error term with mean zero and variance σ^2; and β's are the model parameters to be estimated. In some applications, the explanatory variable is not linearly related with the response variable. In such situation, it is better to model the relationship without specifying the form of the regression function, this is known as *non-parametric regression.* The model relating one explanatory variable to the response variable using a smooth function is expressed as follows:

$$y_i = g(x_i) + \epsilon_i, \qquad i = 1, \dots, n, \tag{2.1}$$

where g is a smooth function whose shape is unrestricted and needed to be estimated and ϵ_i denotes an independent normally distributed error term with mean 0 and variance σ^2.

In non-parametric regression, in general, there are two main choices to be made. The first choice concerns the type of the smooth function to be used to estimate g, which depends on the nature of data and the aim of the study. Smooth functions can be fitted using kernel smoothing methods, such as locally weighted regression and local polynomial regression, or spline smoothing methods, including smoothing and regression splines. All these methods are presented in full details in Bowman and Azzalini (1997), Hastie and Tibshirani (1990) and Wood (2006). The second choice concerns the degree of smoothness which is a trade-off between getting low bias (by interpolating the data to the noise) or having large variance (by obtaining a smoother function) (Hastie and Tibshirani, 1990).

The remainder of this chapter is organized as follows. Firstly, the most commonly used methods of smoothing are described in the following section. Then, the approaches used to determine the degree of smoothness are presented. Additive models and their fitting procedures are introduced afterwards in Section 2.2. In Section 2.3, the different methods of accounting for the auto-correlation in the residuals of the fitted models and adjusting the standard errors of the estimated smooth functions are explained. Finally, the results of modeling EpCO_2 using non-parametric regression and the statistical issues of modeling such high-frequency data using additive models are presented and discussed in Section 2.4.

2.1 Smooth Functions

There are various methods and types of smooth functions that can be used to estimate the non-parametric relationship between a covariate and a response. Some of these methods are used to provide a visual description of the underlying relationship between the explanatory variable and the response variable, while others are used to model and estimate the dependence of the mean of Y on X. Although the methods of estimating smooth functions differ in philosophy and style, the estimates produced at the end are very similar. Therefore, it is always recommended to use a more simple and computationally efficient smoother depending on the situation. Amongst the most frequently used methods of smoothing are the kernel methods and the splines, described in more detail below.

2.1.1 Kernel Smoothing Methods

Generally speaking, the kernel smoother uses a kernel function to define a sequence of weights for each observation on the covariate axis to be assigned to the neighbouring observations. The kernel function is a unimodal density function with a pre-specified scale parameter, centered and reaching its maximum value at the target observation, and decreases monotonically as the observations get further away from the target point

x (Ruppert et al., 2003). This means that the weights given to the observations closer to the observation of interest, x, are larger than the weights given to the further ones (Bowman and Azzalini, 1997). The scale parameter, known as the smoothing parameter, determines the width of the kernel function at each observation of interest and hence the degree of smoothing applied.

2.1.1.1 Loess

Locally weighted regression smoother, known as Loess, is a non parametric regression method that is often used to obtain a graphical description of the underlying patterns in a data set. Although Loess is computationally intensive; it offers great flexibility for the regression surface to be fitted. In Loess, weighted least squares is employed to fit linear or polynomial regression functions of the covariates at each target point using the other points in its neighbourhood (Cleveland, 1979).

At each target point, x, a neighbourhood that contains a pre-specified proportion of the data points around the target point is determined by identifying the k nearest neighbours to this target point, denoted by $N_k(x)$ (Ruppert et al., 2003). Then, the Euclidean distance between the target point x and the furthest away point within each neighbourhood is computed by:

$$\Delta(x) = \max_{N_k(x)} |x - x_i|.$$

This process is repeated for all the target points in the data set (Cleveland, 1979). Then, weights, calculated based on a pre-specified weight function, are assigned to each observation within the neighbourhood. The weights w_i are usually obtained using the tri-cube weighting function defined by:

$$w_i = W\Big(\frac{|x - x_i|}{\Delta(x)}\Big),$$

where

$$W(u) = \begin{cases} (1-u^3)^3 & \text{for } 0 \le u < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Afterwards, these weights are employed within the weighted least squares to obtain the locally weighted smooth (Cleveland, 1979). The proportion of data in each neighbourhood is the smoothing parameter for Loess and is usually known as the span (Hastie and Tibshirani, 1990). The span is the main choice to be made to control the smoothness of the surface estimated using Loess method.

2.1.1.2 Local Polynomial Regression

Local polynomial regression is another smoothing method that uses kernel functions centered around each observation to assign weights to the surrounding observations. Local polynomial regression, unlike Loess, uses fixed width bins. After choosing the kernel function $W(.)$ defining the weighting structure, a polynomial of degree d is fitted at each target point x using weighted least squares as follows:

$$\min_\beta \sum_{i=1}^{n} \{y_i - (\beta_o + \beta_1(x_i - x) + ... + \beta_d(x_i - x)^d)\}^2 W(x_i - x; \lambda). \tag{2.2}$$

The estimated intercept $\hat{\beta}_o$ is the estimator $\hat{g}(x)$ of the smooth function at location x, and λ is the smoothing parameter determining the width of the kernel density surrounding each target observation. As λ increases, the amount of observations covered by the kernel density becomes larger and the fitted curve becomes smoother and approaches the fitted least squares regression line (Bowman and Azzalini, 1997). The most commonly used kernel density in local polynomial regression is the Gaussian density, where the standard deviation is considered as the kernel smoothing parameter (Bowman and Azzalini, 1997, Hastie and Tibshirani, 1990, Ruppert et al., 2003):

$$W(x_i - x; \lambda) = \exp\bigg(-\frac{1}{2}\Big(\frac{x_i - x}{\lambda}\Big)^2\bigg).$$

According to this Gaussian weighting kernel function, observations within a span of 2λ on each side of the target point contribute to the estimate obtained at this point. In practice, local linear ($d = 1$) regression is the most commonly used among all local polynomial estimators for its sufficient flexibility and its superior behaviour near the boundaries of the region where data are collected.

In all the kernel smoothing methods, the smooth function g is estimated by minimizing the local regression least squares criterion given by Equation 2.2. The parameters $\boldsymbol{\beta} = (\beta_o, \beta_1, \ldots, \beta_d)^\top$ used to defined the smooth function g are then estimated using weighted least squares by:

$$\hat{\boldsymbol{\beta}} = \Big(\mathbf{X}^\top \mathbf{W} \mathbf{X}\Big)^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{y},$$

where $\mathbf{y}$ is the vector of the response variable values, $\mathbf{X}$ is the design matrix containing the covariate data and $\mathbf{W}$ is the matrix of assigned weights.

2.1.2 Spline Smoothing Methods

Spline smoothing methods represent an alternative for kernel smoothing methods that provide an estimate for the smooth function not by locally fitting polynomials but rather by fitting piecewise polynomials. A spline function is a curve constructed from polynomial segments joined together smoothly at predefined subintervals. A k^{th} order spline is then a piecewise polynomial function of order k subject to continuity constraints (Bowman and Azzalini, 1997). The points at which the segments are joined are known as knots (Hastie and Tibshirani, 1990) and the order k of the polynomial is the degree of the polynomial segments plus one. Cubic spline functions, obtained by fitting piecewise polynomials of order 4, are the most commonly used splines. There are two types of splines, smoothing splines and regression splines, which mainly differ in the way the smooth function g is estimated.

2.1.2.1 Smoothing Splines

Using *smoothing splines*, the smooth function g is approximated over the closed interval $[a, b]$ by dividing the whole interval into subintervals such that $a \leq s_1 \leq .... \leq s_m \leq b$, then fitting a k^{th} order polynomial segment within each of the intervals $[a, s_1], [s_1, s_2], \ldots, [s_m, b]$ as follows:

$$g(x) = \begin{cases} g_o(x) & : \quad a \leq x \leq s_1 \\ g_1(x) & : \quad s_1 \leq x \leq s_2 \\ \ldots & : \quad \ldots \\ g_m(x) & : \quad s_m \leq x \leq b. \end{cases}$$

The points s_i, $i = 1, \ldots, m$ are the internal knots. Typically, the knots in smoothing splines are the observed unique x values. The smoothing splines require the function g and its j^{th} derivative to be continuous at $s_1, \ldots, s_m$ for each $j = 1, \ldots, k-1$. Cubic smoothing splines are among the most commonly used, where not only the values of the (cubic) polynomial segments are equal at the knots but also the first and second derivatives at the end of one segment are equal to the first and second derivatives at the start of the next segment ensuring the continuity and smoothness of g at each knot.

The smooth function g in Equation 2.1, using smoothing splines, is estimated as $\hat{g}$ by minimizing $\sum_{i=1}^{n}(y_i - g(x_i))^2$. This leads to a regression function that exactly interpolates the data points. However, the typical goal of fitting smoothing splines is estimating a smooth function that is close to the data ignoring the finer scale random noise. Therefore, a second term called the roughness penalty is added to penalize the curvature in the function and g is estimated such that it minimizes the modified least squares criterion as follows:

$$\min_g \sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int_a^b \left(\mathcal{D}^J g(x)\right)^2 dx, \tag{2.3}$$

where λ is a fixed constant known as the smoothing parameter (Bowman and Azzalini, 1997). As λ increases, the influence of the roughness penalty imposed increases relative to the goodness of fit which makes the estimated curve smoother. Conversely as λ approaches zero, the function $\hat{g}$ becomes increasingly locally variable and exactly predicting the data points when $\lambda = 0$. That is, λ controls the trade-off between the goodness of fit and the smoothness of the curve (Hastie and Tibshirani, 1990). $\mathcal{D}^J$ is the J^{th} order derivative defining the roughness penalty. A very popular roughness penalty is the integrated square of the second order derivative, reflecting the curvature at x, given by:

$$\int_a^b \left(\mathcal{D}^2 g(x)\right)^2 dx = \int_a^b \left(g''(x)\right)^2 dx.$$

This second order roughness penalty term suggests that the function $\hat{g}$ minimizing the modified least squares criterion, is a natural cubic spline (Bowman and Azzalini, 1997, Hastie and Tibshirani, 1990). In a natural cubic spline, a piecewise cubic polynomial is fitted in each subinterval of $[a, b]$, and not only the function $\hat{g}$ and its first and second order derivatives are continuous at the knots but also the value of the second and third derivatives of g at the start and end points a and b are both equal to zero (Hastie and Tibshirani, 1990). These supplementary two conditions ensure that the function is linear beyond the boundary knots. Natural cubic splines are often used to produce interpolating splines.

The selection of an appropriate λ is of crucial importance to smoothing splines. In the case of cubic smoothing splines, the number of parameters is equal to the number of observations since the number of parameters defining a spline is the number of interior knots (equal to $n-2$ if there are n unique values of x in the data) plus the degree of the piecewise polynomials minus one, i.e. $n-2+3-1$. This means that although the fitted smooth function has n degrees of freedom, the influence of the smoothing parameter λ results in a function which is smoother than this large number of parameters implies. This large number of parameters is computationally inefficient (Hastie and Tibshirani, 1990), which is considered a potential drawback of the smoothing splines. To overcome this problem, penalized regression splines are frequently employed.

2.1.2.2 Regression Splines

Regression splines fundamentally rely on a set of functions called basis functions. This procedure involves approximating the smooth function, $g(.)$ using a weighted sum of some individual functions, as follows:

$$g(x) = \sum_{p=1}^{P} \psi_p(x) a_p, \tag{2.4}$$

where ψ_p are referred to as basis functions. The choice of the basis functions is one of the crucial choices to be made in regression splines. The most commonly used basis functions are the polynomial *B-spline* basis functions developed by De Boor (2001), known for their flexibility and computational efficiency as a result of their compact support property. This property means that the basis functions are local, such that each basis function is only non-zero over the interval between a small number of adjacent knots, which results in a sparse design matrix (Hastie and Tibshirani, 1990, Ruppert et al., 2003). To define a B-spline basis, the number of basis functions and the number and position of knots have to be specified. The number of basis functions in a B-spline basis is equal to the number of interior knots plus the order of the polynomial defining the basis functions. To define a B-spline basis with P basis functions of degree d, $P+d+1$ knots are needed to be defined. Let the vector $s = (s_1, \ldots, s_{P+d+1})$ be the vector of knots such that $s_1 \leq s_1 \leq \ldots \leq s_{P+d+1}$ and the interval (s_{d+1}, s_{P+1}) is the interval over which we want to estimate the smooth function, then a $(d+1)^{th}$ order spline can be expressed as:

$$g(x) = \sum_{p=1}^{P} \psi_p^d(x) a_p, \tag{2.5}$$

where the individual B-spline basis functions $\psi_p^d(x)$ are defined recursively for $d > 0$ by the Cox-De Boor formula (De Boor, 2001):

$$\psi_p^d(x) = \frac{x - s_p}{s_{p+d} - s_p} \psi_p^{d-1}(x) + \frac{s_{p+d+1} - x}{s_{p+d+1} - s_{p+1}} \psi_{p+1}^{d-1}(x), \quad p = 1, \ldots, P,$$

and

$$\psi_p^0(x) = \begin{cases} 1 & \text{if } s_p \leq x < s_{p+1} \\ 0 & \text{otherwise.} \end{cases}$$

The most frequently used regression splines in practice are the cubic B-splines, where d is equal to 3, as they guarantee an appropriate degree of smoothness. After choosing the appropriate basis system to be used to approximate the smooth function g, the positions of the knots have to be determined. It is very common in regression B-splines to have fewer knots than observations and to place the knots at equidistant points, especially when the observed data are regularly spaced across the interval of interest. Other rules for the placement of the knots can be applied, for example, more knots can be included in regions that involves larger amount of curvature than others.

For a given set of P known basis functions, Equation 2.5 can be re-written in vector-matrix notation as follows:

$$\mathbf{y} = g(\mathbf{x}) = \boldsymbol{\Psi}(\mathbf{x})\mathbf{a}, \tag{2.6}$$

where $\boldsymbol{\Psi}(\mathbf{x})$ is the $n \times P$ matrix of basis functions values $\{\psi_p(x_i)\colon\ p = 1, \ldots, P$ and $i = 1, \ldots, n\}$; and $\mathbf{a}$ is the vector $(a_1, \ldots, a_P)^\top$ of corresponding basis coefficients to be estimated. Using regression spline smoothing, the least squares estimates of the basis coefficients a_p, $p = 1, \ldots, P$, are obtained by minimizing the following residuals sum of squares:

$$\begin{aligned}
&\sum_{i=1}^{n}\Big(y_i - \sum_{p=1}^{P} a_p \psi_p(x_i)\Big)^2 \\
&= \sum_{i=1}^{n}\Big(y_i - \boldsymbol{\psi}(x_i)^\top \mathbf{a}\Big)^2 \\
&= \|(\mathbf{y} - \boldsymbol{\Psi}\mathbf{a})\|^2,
\end{aligned} \tag{2.7}$$

where $\boldsymbol{\psi}(x_i) = \Big(\psi_1(x_i), \ldots, \psi_p(x_i)\Big)^\top$ and $\boldsymbol{\Psi}$ is the design matrix defining the basis functions. The least square estimate of $\mathbf{a}$ obtained by differentiating Equation 2.7 with respect to $\mathbf{a}$ is given by:

$$\hat{\mathbf{a}} = \Big(\boldsymbol{\Psi}^\top \boldsymbol{\Psi}\Big)^{-1} \boldsymbol{\Psi}^\top \mathbf{y}.$$

By multiplying both sides of the above equation by $\boldsymbol{\Psi}$, we obtain:

$$\boldsymbol{\Psi}\hat{\mathbf{a}} = \hat{\mathbf{y}} = \boldsymbol{\Psi}\Big(\boldsymbol{\Psi}^\top \boldsymbol{\Psi}\Big)^{-1} \boldsymbol{\Psi}^\top \mathbf{y},$$

where $\boldsymbol{\Psi}\Big(\boldsymbol{\Psi}^\top \boldsymbol{\Psi}\Big)^{-1} \boldsymbol{\Psi}^\top$ is a square matrix of order n known as the smoothing matrix denoted by $\mathbf{S}$; and hence:

$$\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}. \tag{2.8}$$

The degree of smoothness in regression splines is controlled by the number of basis functions, referred to as basis dimension. This basis dimension is determined by the degree of spline and the number of knots. Increasing the number of knots and hence the basis dimension allows the smooth function to closely follow the data points. However, the selection of the number and locations of knots is a quite complicated discrete process (Wood, 2006).

2.1.2.3 Penalized Regression Splines

Instead of changing the basis dimension to control the smoothness of the estimated function, *penalized regression splines* fix the basis dimension at a large number then add a continuous roughness penalty term on the coefficients of the basis functions to the least squares criterion in Equation 2.7 as follows:

$$\|(\mathbf{y} - \mathbf{\Psi}\boldsymbol{a})\|^2 + \lambda \int_a^b [g''(x)]^2 dx, \tag{2.9}$$

where λ is the smoothing parameter controlling the trade-off between the model fit and the model smoothness. The penalty term $\int_a^b [g''(x)]^2 dx$ can be re-written as a quadratic form in $\mathbf{a}$ as $\mathbf{a}^\top \mathbf{D}\mathbf{a}$, where $\mathbf{D} = \int \boldsymbol{\psi}''(x)\boldsymbol{\psi}''^\top(x)dx$ is a square penalty matrix of order P specific to the basis chosen. Accordingly, Equation 2.9 can be expressed equivalently as:

$$\|(\mathbf{y} - \mathbf{\Psi}\boldsymbol{a})\|^2 + \lambda \mathbf{a}^\top \mathbf{D}\mathbf{a}.$$

The penalized least squares estimates of $a_p : p = 1, \ldots, P$ obtained by minimizing the above penalized criterion with respect to $\mathbf{a}$ are given by (Ruppert et al., 2003):

$$\hat{\mathbf{a}} = \left(\mathbf{\Psi}^\top \mathbf{\Psi} + \lambda \mathbf{D}\right)^{-1} \mathbf{\Psi}^\top \mathbf{y}.$$

By multiplying both sides of the above equation by $\mathbf{\Psi}$, the smoothing matrix $\mathbf{S}$ of penalized regression splines is obtained by:

$$\mathbf{S} = \mathbf{\Psi}\left(\mathbf{\Psi}^\top \mathbf{\Psi} + \lambda \mathbf{D}\right)^{-1} \mathbf{\Psi}^\top,$$

which clearly depends on the smoothing parameter λ.

A key issue in all non-parametric regression methods is the choice of a smoothing parameter that ensures an appropriate degree of smoothness. Therefore, numerous methods, including subjective and automatic procedures, have been developed for determining the optimal smoothing parameter. Some of these methods are discussed in the following section.

2.1.3 Choice of the Smoothing Parameter

There exist various methods for the selection of the optimal smoothing parameter, including cross validation, generalized cross validation and AIC. In general, the smooth curve becomes flatter and the bias increases as the smoothing parameter increases; whereas the curve becomes more wiggly and the variance increase, as the smoothing parameter decreases (Bowman and Azzalini, 1997). By decreasing the smoothing parameter, the estimated smooth function interpolates the data to the noise and therefore the variance increases.

One way for selecting the smoothing parameter is to pre-specify the required degrees of freedom for the smoother. The degrees of freedom define the degree of model complexity, which is itself a trade-off between having low bias and obtaining a more complex model with large variance (Hastie and Tibshirani, 1990, Ruppert et al., 2003).

In a more complicated situation, one may need an automatic procedure for choosing the optimal smoothing parameter ensuring a reasonable level of smoothing. There are several methods to automatically select the best smoothing parameter. One method is Cross Validation (CV), which leaves out one data point at a time and fits the smooth function using the rest of the data. The fitted smooth function is then used to predict the left-out data point (Hastie and Tibshirani, 1990). That is, CV evaluates the amount of smoothing through determining how well each observation is estimated using the rest of the data. This is, mathematically, done by choosing the smoothing parameter that minimizes the following expression (Ruppert et al., 2003):

$$CV = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{g}_{-i}(x_i))^2, \tag{2.10}$$

where $\hat{g}_{-i}(x_i)$ denotes the smooth function estimated at x_i using the data remaining after excluding the i^{th} observation. This method is clearly computationally inefficient since it has to be calculated N times for different values of the smoothing parameter (Hastie and Tibshirani, 1990).

To overcome the computational inefficiency of cross validation, Generalized Cross Validation (GCV) was developed by Craven and Wahba (1979). GCV is a frequently used procedure for the selection of smoothing parameter, calculated only once for a sequence of values of the smoothing parameter as indicated in the following formulae:

$$GCV = \frac{n\text{RSS}}{n - \text{tr}(\mathbf{S})}, \tag{2.11}$$

where RSS is the residuals sum of squares calculated as $\sum_{i=1}^{n}(y_i - \hat{g}(x_i))^2$ and $\mathbf{S}$ is the smoothing matrix. The GCV values are then plotted versus the sequence of smoothing parameter values and the optimal smoothing parameter is chosen such that it corresponds to the minimum GCV value (Hastie and Tibshirani, 1990, Ruppert et al., 2003).

Another commonly used method for the selection of smoothing parameter is the Akaike's information criteria (AIC) (Akaike, 1973), which tries to balance between minimizing the residual sum of squares (or equivalently maximizing the likelihood) and increasing the degree of model complexity (Ruppert et al., 2003). The AIC is computed by Equation 2.12 and one would pick the smoothing parameter minimizing this criteria:

$$\mathrm{AIC} = -2\log(\mathcal{L}) + 2\mathrm{df}_{\mathrm{par}}, \tag{2.12}$$

where $-2\log(\mathcal{L})$ is the residual deviance of the model with $\mathcal{L}$ being the maximized value of the likelihood function; and $\mathrm{df}_{\mathrm{par}}$ is the degrees of freedom of the model parameters (Hastie and Tibshirani, 1990). A corrected version of AIC, AICc, was also proposed to take small sample sizes into account (Sugiura, 1978). The Bayesian Information Criteria (BIC) (Schwarz, 1978) provides another equivalent criterion to AIC derived in a Bayesian framework and is known to penalize the number of parameters more strongly than AIC. For all these criteria, the minimum value indicates the optimal smoothing parameter.

The methods of choosing the optimal smoothing parameters can be grouped into two main classes. The first group involves the procedures that try to minimize the model prediction error, by optimizing criteria such as CV, GCV or AIC. The second group includes the methods that treat smooth functions as random effects, so that the smoothing parameters are the variance parameters which can be estimated by maximum (marginal) likelihood or restricted maximum likelihood. For more information about the second class of estimating the smoothing parameters, see Section 2.3.2. Likelihood methods tend to be more robust and are more preferable in practice to using GCV for smoothing parameter selection (Wood, 2011), especially in the presence of auto-correlation between observations.

2.2 Additive Models

The models relating more than one explanatory variable to the response variable using smooth functions are called additive models and can be expressed as follows:

$$y_i = \beta_o + g_1(x_{1i}) + g_2(x_{2i}) + ... + g_Q(x_{Qi}) + \epsilon_i, \tag{2.13}$$

where the g's are the smooth functions whose shapes are unrestricted and needed to be estimated and ϵ_i denotes an independent normally distributed error term with mean 0 and variance σ^2. A parametric model component can be incorporated in addition to the non parametric component, this is referred to as semi-parametric regression (Ruppert et al., 2003). The smooth function could also be a joint smooth function of more than one covariate as follows:

$$y_i = \beta_o + g_1(x_{1i}) + g_2(x_{2i}) + g_3(x_{3i}, x_{4i}) + ... + \epsilon_i.$$

Nowadays, additive models become of great use for expressing the relationship between a certain response variable and its explanatory variables. This is due to the fact that most of these relationships are not linear and models should account for this non linearity to enhance the prediction accuracy. Examples can be found in Bowman et al. (2009), Ferguson et al. (2008), Giannitrapani et al. (2011), Lamon et al. (1996), Underwood (2009) and McMullan et al. (2007).

Additive models can also be extended to Generalized additive models (GAMs), where the error term ϵ_i are no longer normally distributed but rather have a distribution that belongs to the exponential family of distributions. A wide range of applications of GAMs, where the response is not normally distributed, are available in Wood (2006). In a GAM, the mean $E(Y|X_1, \ldots, X_Q)$ is linked to the explanatory variables via the following link function:

$$\mathcal{G}(E(Y|X_1, \ldots, X_Q)) = \mathcal{G}(\mu) = \beta_o + g_1(x_{1i}) + g_2(x_{2i}) + ... + g_p(x_{Qi}),$$

where the parameter β_o and the smooth functions $g_1, \ldots, g_Q$ are estimated either using the local scoring procedure or the penalized iteratively re-weighted least squares. A wide area of applications of generalized additive models is epidemiology, where it is of interest to model the effect of air pollution on health and mortality. For example, Schwartz (1994) has used generalized additive models to study the effect of air pollution on respiratory illness in different cities of the world. Whereas, Kelsall et al. (1997) have developed poisson regression models to estimate the increased risk of daily mortality associated with air pollution while controlling for longer-term time trends, seasonality and weather conditions using generalized additive models. GAMs are out of the scope of this thesis, see Hastie and Tibshirani (1990) and Ruppert et al. (2003) for more details and Wood (2006) for more applications.

As previously mentioned in Chapter 1, additive models can be used to model the trend and seasonal components, where the covariates are the different time components. For example, to model the global trend of a time series using a smooth function, the covariate could be a continuous time variable and the model is then of the form:

$$y_i = g(\text{time}_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2).$$

If the phenomenon under study appears to exhibit, in addition to the overall trend, periodic/seasonal patterns over time; a suitable model is of the form:

$$y_i = g_1(\text{time}_i) + g_2(\text{month}_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$

where the covariates are the continuous time (decimal year) variable and the month reflecting the global trend and seasonal components, respectively. In a situation where there is evidence for a varying seasonal pattern over time, a bivariate term can be used as follows:

$$y_i = g(\text{time}_i, \text{month}_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2).$$

To fit any of these additive models, there are different approaches based on the method of smoothing used for each of the covariates. The most commonly used approaches are the backfitting algorithm introduced in Section 2.2.1 and the penalized regression spline fitting approach presented in Section 2.2.2. Recently, Wood et al. (2015) have also developed more practical generalized additive model fitting procedures for large data sets in the case in which the smooth terms in the model are represented by using penalized regression splines.

2.2.1 Backfitting Algorithm

The backfitting algorithm is a simple and general algorithm, which can be used to fit any additive model involving any method of smoothing. Before starting the fitting algorithm, a method of smoothing is specified for each explanatory variable in the model such that $\hat{g}_q(\mathbf{x}_q) = \boldsymbol{\mathcal{S}}_q\mathbf{y}$. Then, the smoothing is performed iteratively for each of the covariates using the j^{th} partial residuals as response (Bowman and Azzalini, 1997), as follows:

$$\hat{g}_j = \boldsymbol{\mathcal{S}}_j(\mathbf{y} - \hat{\beta}_o - \sum_{q \neq j} \hat{g}_q(\mathbf{x}_q)),$$

where $\mathbf{x}_q$ is the vector of the q^{th} covariate values and $\boldsymbol{\mathcal{S}}_j$ is the smoothing matrix of the response $\mathbf{y}$ against the j^{th} partial residuals vector, $\mathbf{y} - \hat{\beta}_o - \sum_{q \neq j} \hat{g}_q(\mathbf{x}_q)$, resulting from subtracting all the current model estimates from the response variable, except the estimate of the j^{th} smooth.

The backfitting algorithm works as follows (Hastie and Tibshirani, 1990):

1. Initialize $\hat{\beta}_o = \bar{y}$ and obtain $\hat{g}_j^{(o)}, j = 1, \ldots, J$ by fitting a smooth function for each of the covariates separately.

2. Iterate for $j = 1, \ldots, Q, 1, \ldots, Q, \ldots$

   $$\hat{g}_j^{(\mathcal{I})} = \boldsymbol{\mathcal{S}}_j(\mathbf{y} - \hat{\beta}_o - \sum_{q=1}^{j-1} \hat{g}_q^{(\mathcal{I})}(\mathbf{x}_q) - \sum_{q=j+1}^{Q} \hat{g}_q^{(\mathcal{I}-1)}(\mathbf{x}_q))$$

   where $\mathcal{I}$ is the iteration number.

3. Repeat the previous step until convergence i.e. until the RSS= $\boldsymbol{\mathcal{S}}_j(\mathbf{y} - \hat{\beta}_o - \sum_{q \neq j} \hat{g}_q^{(l)}(\mathbf{x}_q))^2$ fails to decrease, that is the smooth functions $\hat{g}_j, j = 1, \dots, p$ change less than a pre-specified threshold.

One limitation of the above backfitting procedure is that it does not incorporate the estimation of the smoothing parameter within the fitting algorithm. Therefore, the penalized regression spline fitting procedure which can involve implicit selection of the smoothing parameter is more preferred (Wood, 2006).

2.2.2 Penalized Regression Spline Fitting Approach

This approach of model fitting is mainly used when approximating the smooth functions with regression splines, see Section 2.1.2.3. Suppose that each of the smooth functions of the covariates in an additive model can be expressed in terms of spline basis functions, i.e. $g_q(\mathbf{x}) = \sum_{p=1}^{P} \psi_p(\mathbf{x}) a_p$. By imposing a penalty term on each smooth function of the form $\int_a^b [g_q''(\mathbf{x})]^2 d\mathbf{x}$, the penalized regression spline fit of the whole additive model is obtained by minimizing the following expression with respect to $\mathbf{A}$:

$$\|(\mathbf{y} - \boldsymbol{\Psi}\mathbf{A})\|^2 + \sum_{q=1}^{Q} \lambda_q \mathbf{A}^\top \mathbf{D}_q \mathbf{A}, \tag{2.14}$$

where $\boldsymbol{\Psi}$ is the matrix of basis functions values for all covariates and $\mathbf{A}$ is the combined vector of basis coefficients for all covariates. That is, if each covariate q is approximated with P_q basis functions, then the matrix $\boldsymbol{\Psi}$ is of order $N \times \sum_{q=1}^{Q} P_q$ and the column vector $\mathbf{A}$ is of length $\sum_{q=1}^{Q} P_q$. To overcome identifiability problems, a constraint is put on each of the individual smooth functions throughout the model fitting ensuring that there are no free constants in the smooth terms. Each term is constrained to have mean zero after fitting the model. These constraints force the Effective Degrees of Freedom (EDF) of each smooth function to be $\text{tr}(\boldsymbol{\mathcal{S}}_q) - 1$, instead of $\text{tr}(\boldsymbol{\mathcal{S}}_q)$, reflecting the existence of a redundant constant in $Q - 1$ of the Q terms before fitting the model (Bowman and Azzalini, 1997).

In Equation 2.14, λ_q and $\mathbf{D}_q$ are the smoothing parameter and the penalty matrix assigned to the q^{th} smooth function, respectively. The penalty matrix $\mathbf{D}_q$ is determined based on the basis system chosen to fit the q^{th} smooth function. Each penalty matrix $\mathbf{D}_q$ is of order $\sum_{q=1}^{Q} P_q \times \sum_{q=1}^{Q} P_q$ and is zero everywhere except the cells associated with the penalized coefficients of the q^{th} smooth term. The estimate $\hat{\mathbf{A}}$ which minimizes the modified least squares criterion in Equation 2.14 is given by (Hastie and Tibshirani, 1990, Ruppert et al., 2003):

$$\hat{\mathbf{A}} = (\mathbf{\Psi}^\top \mathbf{\Psi} + \sum_{q=1}^{Q} \lambda_q \mathbf{D}_q)^{-1} \mathbf{\Psi}^\top \mathbf{y}.$$

The overall smoothing matrix $\mathbf{S}$ is then given by: $\mathbf{\Psi}(\mathbf{\Psi}^\top \mathbf{\Psi} + \sum_{q=1}^{Q} \lambda_q \mathbf{D}_q)^{-1} \mathbf{\Psi}^\top$. The smoothing parameters λ_q can be either automatically selected using one of the procedures listed in Section 2.1.3 or implicitly estimated using the restricted maximum likelihood method, see Section 2.3.2 and Wood (2006) for more details.

2.2.3 Model Comparisons

One way to compare pairs of (generalized) additive models, to asses which model fits the data better, is through the use of approximate F-tests (Hastie and Tibshirani, 1990). Because two additive models are not necessarily nested, the test can only be used approximately. Suppose there are two models, a full model M_1 and a reduced model M_0 with degrees of freedom for error df_1 and df_0, respectively. To compare these two models, an approximate F-test is used and the following F-statistic is calculated:

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(df_0 - df_1)}{\mathrm{RSS}/df_1},$$

where RSS_0 and RSS_1 are the residuals sum of squares of the models M_0 and M_1, respectively. The F-statistic is then compared to an F-distribution with $(df_0 - df_1, df_1)$ degrees of freedom.

The residuals sum of squares of each model can be expressed as follows:

$$\mathrm{RSS} = \mathbf{y}^\top (\mathbf{I} - \mathbf{S})^\top (\mathbf{I} - \mathbf{S}) \mathbf{y},$$

where $\mathbf{S}$ is the corresponding smoothing matrix. The corresponding degrees of freedom for independent errors can be expressed in terms of the smoothing matrix $\mathbf{S}$ (Hastie and Tibshirani, 1990) as the trace of $(\mathbf{I} - \mathbf{S})^T (\mathbf{I} - \mathbf{S})$, which is equivalent to:

$$df = n - \mathrm{tr}(2\mathbf{S} - \mathbf{S}\mathbf{S}^\top).$$

2.3 Additive Models with Correlated Errors

Additive models typically assume that the errors ϵ_i are mutually independent, whilst time series observations, by nature, can be correlated in adjacent time points. In a time series analysis context, additive models will only describe how the response variable is statistically related to the explanatory variables without accounting for the dependence of the response on its past values. However, in many environmental studies, the response variable of interest may involve autocorrelation, and hence it is necessary to account for this autocorrelation appropriately when modeling. Autocorrelation causes problems to the estimation of additive models, since they essentially assume that each observation is independently distributed. Failure to account for this autocorrelation may underestimate the standard errors of the estimated smooth curves, which in turn makes the estimates inefficient and the inference about those estimates unreliable. Nevertheless, autocorrelation may influence the smoothing parameter selection from automatic procedures.

Some researchers have suggested the use of a two-stage additive model fitting procedure to account for the autocorrelation in the residual analysis, especially when the fundamental purpose of the model is to *describ*e the behaviour of the variable of interest (see Bowman et al. (2009), Ferguson et al. (2008), Giannitrapani et al. (2011) and Chen et al. (2014)). Alternatively, Lin and Zhang (1999) have proposed the use of Generalized Additive Mixed Models (GAMMs), an extension to Generalized Linear Mixed Models (GLMMs). GAMMs account for the autocorrelation between observations in the random effects part of the model. Other researchers have proposed a Bayesian approach (Fahrmeir et al., 2004, Fahrmeir and Lang, 2001) to account for the auto-correlation in model fitting. However, the computational cost of this approach is quite high, especially when the sample size is large. The following two subsections explain in detail the two-stage GAM fitting procedure and the GAMMs.

2.3.1 Two-stage Additive Model Fitting Procedure

One way to account for serial time dependence in the response variable in additive models is a two-stage additive model fitting procedure (TSP). The first stage involves fitting an additive model assuming independent distributed errors as in Section 2.2. The second stage entails fitting an appropriate correlation structure to the residuals of the fitted additive model. Consider the case where the residuals ϵ_i in Equation 2.13 are significantly autocorrelated, and let ϵ_i follow a pre-specified correlation structure. One of the most commonly temporal correlation structures are the ARMA processes, presented in Chapter 1.

In practice, simple autoregressive processes are able to explain most of the autocorrelation structures between observations and are widely used to approximate ARMA processes. For example, Ferguson et al. (2009) accounted for the temporal correlation in the residuals using a multivariate AR(1) after modeling the changing relationship between the water quality variables at Loch Leven located in Scotland using a multivariate-varying coefficient model. An AR(1) process was also incorporated to capture the remaining correlation structure in the errors after using a non-parametric additive model framework to model the sulphur dioxide pollution smoothly across Europe in space and time (Bowman et al., 2009). Giannitrapani et al. (2011) have showed that an AR(1) process has provided a good description for the error term remaining after modeling the sulphur dioxide pollution data at one European site over time using an additive model with smooth terms of the different time components. The main objective in most environmental applications is to describe the mean structure and hence it is sufficient to employ a simple model for the errors to account for the main effects of correlation. More complex time series models can also be employed if there is a strong evidence for a more complex correlation structure in the errors (Giannitrapani et al., 2011).

Let $\mathbf{V}$ be the correlation matrix of the errors, ϵ_i, remaining after fitting an additive model to describe the mean pattern of the response variable in relation to the explanatory variables. An estimate for this correlation matrix, $\hat{\mathbf{V}}$, can be obtained from the data based on a specified ARMA correlation structure. Each smooth function of the additive model has an estimate of the form $\hat{g}_j = \boldsymbol{\mathcal{S}}_j \mathbf{y}$, where $\boldsymbol{\mathcal{S}}_j$ is the smoothing matrix of component j, and the standard errors are readily available as the square root of the diagonal entries $\boldsymbol{\mathcal{S}}_j \hat{\mathbf{V}} \boldsymbol{\mathcal{S}}_j^\top \sigma^2$. The error variance σ^2 can be estimated from the RSS $= \mathbf{y}^\top (\mathbf{I} - \mathbf{S})^\top (\mathbf{I} - \mathbf{S})\mathbf{y}$ and the approximate degrees of freedom associated with error is given by $\mathrm{tr}\{(\mathbf{I} - \mathbf{S})(\mathbf{I} - \mathbf{S})^\top \mathbf{V}\}$. The variability bands (± 2 s.e.) of the estimated smooth curves can then be adjusted based on the new standard errors (Bowman et al., 2009).

2.3.2 Generalized Additive Mixed Models

An alternative approach to account for the serial correlation between the errors is through Generalized Additive Mixed Models (GAMMs). A GAMM can simultaneously fit a GAM, or an additive model as a special case if the errors are assumed to be normally distributed, and a mixed effect model that accounts for the auto-correlation between the model residuals. A GAMM can be expressed as follows:

$$\mathcal{G}(\mu_i) = \beta_o + g_1(x_{1i}) + g_2(x_{2i}) + g_3(x_{3i}, x_{4i}) + ... + \mathbf{Z_i}\alpha,$$

where the observations y_i are assumed to be conditionally independent with means $E(Y_i|\boldsymbol{\alpha}) = \mu_i$; g_j are the smooth functions of the covariates X_j; $\mathbf{Z}_i$ is the i^{th} row of the random effects model matrix $\mathbf{Z}$; and $\boldsymbol{\alpha} \sim N(\mathbf{0}, \boldsymbol{\Gamma})$ is the vector of random effects coefficients, having an unknown positive definite covariance matrix $\boldsymbol{\Gamma}$ to be estimated from the data (Wood, 2006). GAMMs are fitted using penalized regression, where the smooth functions are approximated by basis functions such as B-splines basis (Ruppert et al., 2003, Wood, 2006).

An additive model can be expressed and fitted as a mixed effects model, by treating each smooth term as the sum of a fixed effects component (unpenalized coefficients), which can be absorbed in $\boldsymbol{\Psi}_i\mathbf{a}$ and a random effects component (penalized coefficients), which can be absorbed in $\mathbf{Z}_i\boldsymbol{\alpha}$. $\boldsymbol{\Psi}_i$ and $\mathbf{Z}_i$ are the i^{th} rows of the fixed and random effects model matrices, respectively. In this mixed effects model, the smoothing parameters are treated as the variance component parameters needed to be estimated by means of maximum likelihood or restricted maximum likelihood (Wood, 2006). For simplicity, consider the case of one univariate smooth function approximated by a B-spline basis. That is, $g(x) = \sum_{p=1}^{P} \psi_p(x)a_p$ with associated roughness measure $\mathbf{a}^T\mathbf{D}\mathbf{a}$, where $\mathbf{D}$ is a semi-positive definite penalty matrix. Let the matrix $\boldsymbol{\Psi}$ be the design matrix, where $\boldsymbol{\Psi}_{\boldsymbol{ip}} = \psi_p(x_i)$ and $\boldsymbol{\Psi}\mathbf{a} = g(\mathbf{x})$. The function g is assumed to be smooth; then by using a Bayesian approach, a prior distribution that is proportional to $\exp(-\lambda\mathbf{a}^\top\mathbf{D}\mathbf{a}/2)$ is specified for the wiggliness of g. However, this prior is an improper prior for $\mathbf{a}$, which does not fit into standard linear mixed modeling approaches. Therefore, a re-parametrization is required so that the new parameters are divided into a set having a proper distribution, treated as random effects, and a set having an improper distribution, treated as fixed effects. This re-parametrization is obtained by using the eigen-decomposition, $\mathbf{D} = \mathbf{U}\mathbf{K}\mathbf{U}^\top$, where $\mathbf{U}$ is an orthogonal matrix whose columns are the eigenvectors of $\mathbf{D}$ and $\mathbf{K}$ is a diagonal matrix whose main diagonal elements are the corresponding eigenvalues of $\mathbf{D}$ ordered in a descending order. Let $\mathbf{K}_+$ denotes the smallest sub-matrix of $\mathbf{K}$ containing all the strictly positive eigenvalues. The new parameters vector can now be written as $(\boldsymbol{\alpha}_R^\top, \mathbf{a}_F^\top)^\top \equiv \mathbf{U}^\top\mathbf{a}$, where $\boldsymbol{\alpha}_R$ are the random effects coefficients and $\mathbf{a}_F$ are the fixed effects coefficients. It is then obvious that $\mathbf{a}^\top\mathbf{D}\mathbf{a} = \boldsymbol{\alpha}_R^\top\mathbf{K}_+\boldsymbol{\alpha}_R$, since $\mathbf{a}_F$ are unpenalized with corresponding zero eigenvalues. The eigenvectors matrix can also be partitioned so that $\mathbf{U} = [\mathbf{U}_R : \mathbf{U}_F]$, and hence $\boldsymbol{\Psi}_F$ and $\boldsymbol{\Psi}_R$ can be defined as $\boldsymbol{\Psi}_F \equiv \boldsymbol{\Psi}\mathbf{U}_F$ and $\boldsymbol{\Psi}_R \equiv \boldsymbol{\Psi}\mathbf{U}_R$. Following from this, the mixed model representation of the smooth function g in terms of a linear predictor and random effects is as follows (Wood, 2006):

$$\boldsymbol{\Psi}_F\mathbf{a}_F + \boldsymbol{\Psi}_R\boldsymbol{\alpha}_R, \quad \boldsymbol{\alpha}_R \sim N(\mathbf{0}, \mathbf{K}_+^{-1}/\lambda), \tag{2.15}$$

where $\mathbf{a}_F$ and λ are the fixed model parameters to be estimated. Let $\boldsymbol{\alpha} = \mathbf{K}_+^{1/2}\boldsymbol{\alpha}_R$ and $\mathbf{Z} = \boldsymbol{\Psi}_R\mathbf{K}_+^{-1/2}$, so that the mixed model representation given by Equation 2.15 can be equivalently expressed as (Wood, 2006):

$$\boldsymbol{\Psi}_F \mathbf{a}_F + \mathbf{Z}\boldsymbol{\alpha}, \quad \boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{I}/\lambda). \tag{2.16}$$

Inserting the terms in Equation 2.16 to a standard generalized linear mixed effects model is done by appending the columns of $\boldsymbol{\Psi}_F$ to the fixed effects design matrix and the columns of $\mathbf{Z}$ to the random effects design matrix, and specifying a covariance matrix for the random effects. This model can then be fitted using penalized maximum likelihood method.

One simple example to illustrate and explain the GAMMs fitting procedure is to assume normality and suppose that the model residuals ϵ_i are not independent in a way such that every two model residuals adjacent in time are correlated. Thus, an autoregressive model of order one, AR(1), can be used to fit the model residuals such that $\epsilon_i = \rho\epsilon_{i-1} + \varepsilon_i$, where ε_i are i.i.d $N(0, \sigma^2)$ and $|\rho| < 1$ and the following additive mixed model is fitted to the data (Ruppert et al., 2003):

$$\mathbf{y} = \boldsymbol{\Psi}_F \mathbf{a}_F + \tilde{\mathbf{Z}}\tilde{\boldsymbol{\alpha}} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}), \tag{2.17}$$

where $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ and $\tilde{\mathbf{Z}}\tilde{\boldsymbol{\alpha}} = [\mathbf{Z}\boldsymbol{\alpha}, \boldsymbol{\alpha}^\rho]$ such that $\boldsymbol{\alpha}^\rho$ is the column vector whose i^{th} element is given by $\rho\epsilon_{i-1}$. Following the model representation in Equation 2.16 and the AR(1) process fitted to the errors ϵ_i, $\boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{I}/\lambda)$ and $\boldsymbol{\alpha}^\rho \sim N(0, \boldsymbol{\Sigma}^\star)$, and COV $\begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\alpha}^\rho \end{bmatrix} = \begin{bmatrix} \mathbf{I}/\lambda & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}^\star \end{bmatrix}$, where $\boldsymbol{\Sigma}^\star$ is the variance-covariance matrix of an AR(1) given by:

$$\boldsymbol{\Sigma}^\star = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \dots & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \rho^2 & \dots & \dots & \rho^{n-2} \\ \vdots & \vdots & \ddots & & \vdots & \vdots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \dots & \dots & \rho^2 & \rho & 1 \end{pmatrix}$$

.

Thus, $\mathbf{y}\backslash\tilde{\boldsymbol{\alpha}} \sim N(\boldsymbol{\Psi}_F\mathbf{a}_F + \tilde{\mathbf{Z}}\tilde{\boldsymbol{\alpha}}, \boldsymbol{\Sigma})$ and $\tilde{\boldsymbol{\alpha}} \sim N(\mathbf{0}, \mathbf{G} = \begin{bmatrix} \mathbf{I}/\lambda & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}^\star \end{bmatrix})$. The coefficients $\mathbf{a}_F$ and $\tilde{\boldsymbol{\alpha}}$ are then estimated by maximizing the penalized log-likelihood of $(\mathbf{y}, \tilde{\boldsymbol{\alpha}})$, which is equivalent to maximizing the following criterion (Ruppert et al., 2003):

$$(\mathbf{y} - \boldsymbol{\Psi}_F\mathbf{a}_F - \tilde{\mathbf{Z}}\tilde{\boldsymbol{\alpha}})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\Psi}_F\mathbf{a}_F - \tilde{\mathbf{Z}}\tilde{\boldsymbol{\alpha}}) + \tilde{\boldsymbol{\alpha}}^\top \mathbf{G}^{-1} \tilde{\boldsymbol{\alpha}}.$$

It is clear that the maximum likelihood estimates of $\mathbf{a}_F$ depends on the estimate of $\tilde{\boldsymbol{\alpha}}$, which in turn depends on $\mathbf{a}_F$. Therefore, the maximum likelihood estimates of both $\mathbf{a}_F$

and $\tilde{\boldsymbol{\alpha}}$ are determined jointly by maximizing the following criterion:

$$(\mathbf{y} - \mathbf{C}\boldsymbol{\theta})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{C}\boldsymbol{\theta}) + \boldsymbol{\theta}^T \mathbf{M}\boldsymbol{\theta},$$

where $\boldsymbol{\theta} = (\mathbf{a}_F^\top, \tilde{\boldsymbol{\alpha}}^\top)^\top$, $\mathbf{C} = (\boldsymbol{X}_F^\top, \tilde{\mathbf{Z}}^\top)^\top$ and $\mathbf{M} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} \end{bmatrix}$; and hence the estimates $\hat{\mathbf{a}}_F$ and $\hat{\tilde{\boldsymbol{\alpha}}}$ are obtained by (Ruppert et al., 2003):

$$\begin{bmatrix} \hat{\mathbf{a}}_F \\ \hat{\tilde{\boldsymbol{\alpha}}} \end{bmatrix} = [\mathbf{C}^\top \boldsymbol{\Sigma}^{-1}\mathbf{C} + \mathbf{M}]^{-1}\mathbf{C}^\top \boldsymbol{\Sigma}^{-1}\mathbf{y}.$$

Next, the estimated values of the random and fixed effects are used to evaluate the profile log-likelihood of the variance components to estimate the smoothing parameter λ and the auto-regressive parameter ρ (Phineiro and Bates, 2000).

It is also possible to rewrite the model given by Equation 2.17 as $\mathbf{y} = \boldsymbol{\Psi}_F\mathbf{a}_F + \boldsymbol{\epsilon}^\star$, where $\boldsymbol{\epsilon}^\star = \tilde{\mathbf{Z}}\tilde{\boldsymbol{\alpha}} + \boldsymbol{\epsilon}$ (Phineiro and Bates, 2000, Ruppert et al., 2003). Subsequently, the variance-covariance matrix of $\mathbf{y}$ is $\mathbf{V} = \tilde{\mathbf{Z}}\mathbf{G}\tilde{\mathbf{Z}}^T + \boldsymbol{\Sigma}$, such that the variance components in $\mathbf{G}$ and $\boldsymbol{\Sigma}$ are estimated using maximum likelihood or restricted maximum likelihood methods. The maximum likelihood estimate of $\mathbf{V}$ is based on the model $\mathbf{y} \sim N(\boldsymbol{\Psi}_F\mathbf{a}_F, \mathbf{V})$:

$$l(\mathbf{a}_F, \mathbf{V}) = -\frac{1}{2}\{n\log(2\pi) + \log|\mathbf{V}| + (\mathbf{y} - \boldsymbol{\Psi}_F\mathbf{a}_F)^\top \mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\Psi}_F\mathbf{a}_F)\}. \tag{2.18}$$

For a fixed $\mathbf{V}$, the maximum likelihood estimate of $\mathbf{a}_F$ is:

$$\hat{\mathbf{a}}_F = (\boldsymbol{\Psi}_F^\top \mathbf{V}^{-1}\boldsymbol{\Psi}_F)^{-1}\boldsymbol{\Psi}_F^\top \mathbf{V}^{-1}\mathbf{y}. \tag{2.19}$$

The estimator $\hat{\mathbf{a}}_F$ is the best linear unbiased estimator for $\mathbf{a}_F$. Whereas, the random effects $\tilde{\boldsymbol{\alpha}}$ can be predicted from $\mathbf{y} - \boldsymbol{\Psi}_F\hat{\mathbf{a}}_F = \tilde{\mathbf{Z}}\tilde{\boldsymbol{\alpha}} + \boldsymbol{\epsilon}$ resulting in the best linear predictor (Ruppert et al., 2003):

$$\hat{\tilde{\boldsymbol{\alpha}}} = \mathbf{G}\tilde{\mathbf{Z}}^\top \mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\Psi}_F\hat{\mathbf{a}}_\mathbf{F}). \tag{2.20}$$

For known $\mathbf{G}$ and $\mathbf{V}$, $\hat{\tilde{\boldsymbol{\alpha}}}$ is the best linear unbiased predictor for $\tilde{\boldsymbol{\alpha}}$.

The variance-covariance matrix $\mathbf{V}$ is obtained by plugging the estimates obtained in Equations 2.19 and 2.20 into the log-likelihood of $\mathbf{V}$ (Ruppert et al., 2003), thus we have:

$$l_p(\mathbf{V}) = -\frac{1}{2}\{n\log(2\pi) + \log|\mathbf{V}| + (\mathbf{y} - \boldsymbol{\Psi}_F\hat{\mathbf{a}}_F)^\top \mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\Psi}_F\hat{\mathbf{a}}_F)\}. \tag{2.21}$$

The maximum likelihood estimates of the parameters defining $\mathbf{V}$ can then be obtained by maximizing numerically this profile log-likelihood with respect to these parameters.

Plugging the estimated variance-covariance matrix $\mathbf{V}$ in Equations 2.19 and 2.20, the estimates of both the best linear estimator of $\mathbf{a}_F$ and the best linear predictor of $\tilde{\boldsymbol{\alpha}}$ are computed, respectively, by:

$$\hat{\mathbf{a}}_F = (\boldsymbol{\Psi}_F^\top \hat{\mathbf{V}}^{-1} \boldsymbol{\Psi}_F)^{-1} \boldsymbol{\Psi}_F^\top \hat{\mathbf{V}}^{-1} \mathbf{y} \tag{2.22}$$

$$\hat{\tilde{\boldsymbol{\alpha}}} = \hat{\mathbf{G}} \tilde{\mathbf{Z}}^\top \hat{\mathbf{V}}^{-1} (\mathbf{y} - \boldsymbol{\Psi} \hat{\mathbf{a}}) \tag{2.23}$$

Notice that the variance parameters can also be estimated using the restricted maximum likelihood (REML) method, which involves maximizing the restricted log-likelihood given by:

$$l_{\text{RE}}(\mathbf{V}) = l_p(\mathbf{V}) - \frac{1}{2} \log |\boldsymbol{\Psi}_F^\top \mathbf{V}^{-1} \boldsymbol{\Psi}_F|$$

The restricted maximum likelihood estimate is more accurate than the maximum likelihood estimate since it takes into account the degrees of freedom for the fixed effects in the model (Ruppert et al., 2003).

A major drawback of GAMMs is that they are computationally inefficient with large time series. They can also be numerically unstable as a result of the confounding between correlation and non-linearity and the increased complexity required for modeling strong and persistent auto-correlation (Wood, 2006). These drawbacks will be discussed in the following Section.

2.4 Modeling the EpCO$_2$ Data

The exploratory data analysis and wavelet analysis results, presented in Chapter 1, highlighted the seasonal and diurnal variations of EpCO$_2$ and the differences in these variations between the individual hydrological years. They also indicated that the variability in water hydrodynamics co-vary with the variability in EpCO$_2$. However, it is not yet clear how the EpCO$_2$ is related to the river hydrology, after accounting for the temporal dynamics and the persistent temporal correlation between the high-frequency measurements. Therefore, advanced flexible regression modeling in general, and additive models in particular, are employed to explain the temporal dynamics of EpCO$_2$ and examine the relationship between EpCO$_2$ and the available physicochemical catchment variables, which are not used in deriving the EpCO$_2$ (i.e. specific conductivity only). The EpCO$_2$ measurements were indirectly calculated from the river discharge, temperature and pH measurements. For this reason, these variables were excluded in all further statistical modelling from the set of explanatory variables used to describe the EpCO$_2$ to avoid multicolinearity and overfitting.

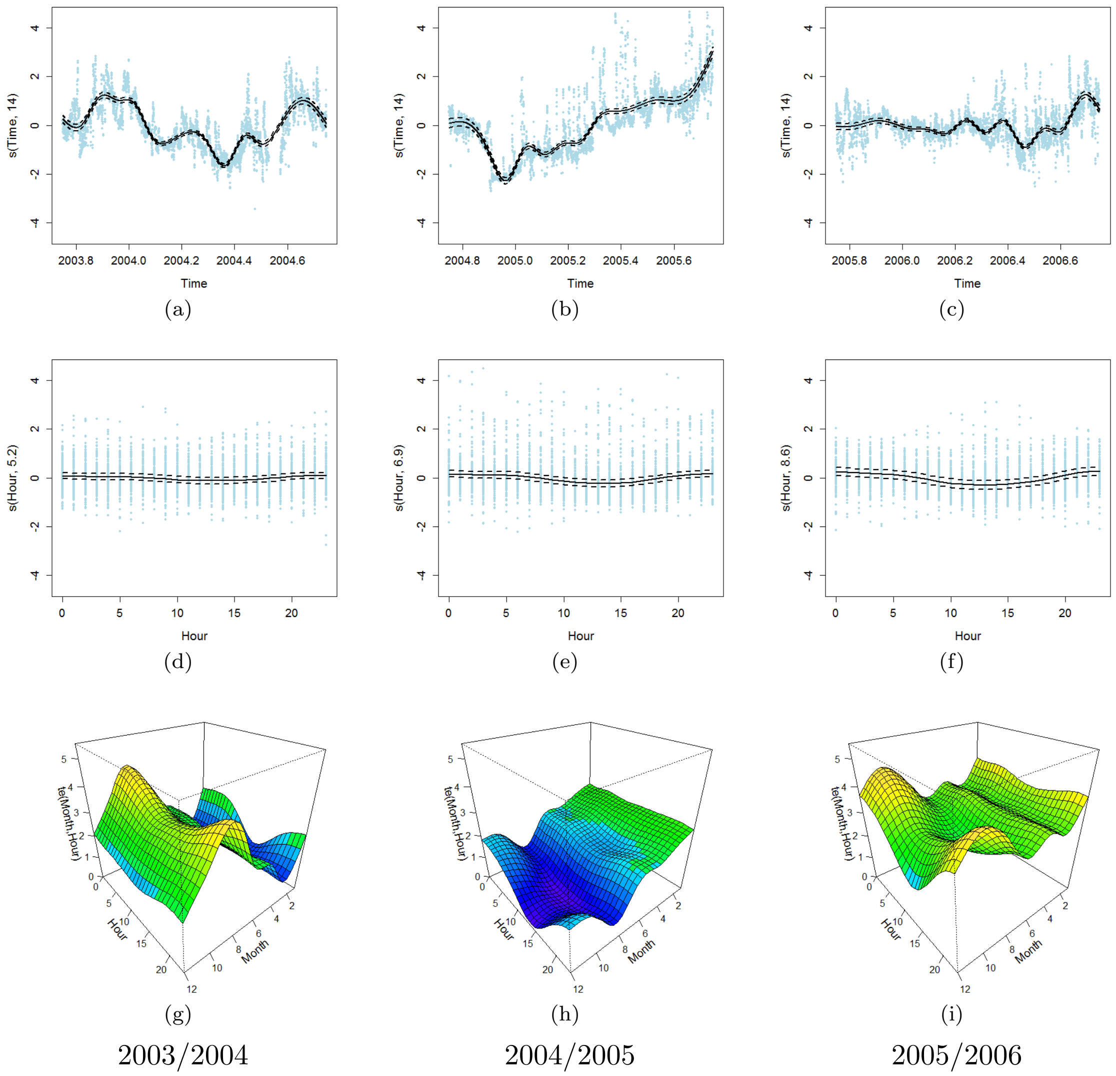
For the fact that each hydrological year is characterized by different climatological and hydrological conditions, additive models are fitted for each HY separately. In these additive models, the univariate smooth terms are fitted using penalized cubic regression splines, except for the periodic effects which are estimated using penalized cyclic cubic regression splines, and the bivariate smooth functions are represented by tensor product splines. Tensor product splines are invariant to linear scaling of covariates and are good to smooth interactions of quantities measured in different units (Wood, 2006). For a detailed discussion on the different types of splines, see (Wood, 2006). The flexibility of each smooth function g is determined by first setting its corresponding basis dimension based on the AIC and sensitivity analysis to identify a smooth interpretable relationship, then automatically selecting the appropriate smoothing parameter using the restricted maximum likelihood (REML) method. Model variable selection is performed using AIC and approximate F-tests. The `mgcv` package (Wood, 2006) supplied with R for fitting generalized additive models and generalized additive mixed models, is used. Additive models are estimated using the fitting routine `bam` which is a computationally efficient alternative for the main routine `gam` for very large data sets (Wood et al., 2015), whereas additive mixed models are fitted using the `gamm` routine (Wood, 2006).

2.4.1 Modeling EpCO$_2$ within each Hydrological Year Separately

Let y_t be a random realization from a normally distributed random variable denoting the EpCO$_2$ measured at time t, $x_t^{\text{Time within year}}$ is the value of the time variable on a decimal year scale representing the time within hydrological year at which the measurement is recorded; and x_t^{Hour} and x_t^{Month} are the corresponding indices for the hour within day and month within year. Then, the temporal dynamics of EpCO$_2$ within each hydrological year are initially investigated through the following additive model:

$$y_t = g_1(x_t^{\text{Time within year}}) + g_2(x_t^{\text{Hour}}) + g_3(x_t^{\text{Hour}}, x_t^{\text{Month}}) + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2), \qquad (2.24)$$

where the smooth functions g_1 and g_2 capture the yearly trend and the diurnal cycle of EpCO$_2$, respectively; while g_3 reflects the changing diurnal effect from one month to another and ϵ_t accounts for the random effects not explained by the additive model. The univariate smooth terms g_1 and g_2 are approximated using penalized cubic regression splines and penalized cyclic cubic regression splines, respectively; with a maximum of 14 degrees of freedom each. The smooth function g_1 is added to the model to estimate the overall trend of EpCO$_2$; and hence using a larger number of knots to estimate it depicts finer variations that are not of interest and using a smaller number of knots misses the fluctuations in EpCO$_2$ from one month to another. The bivariate smooth function g_3 is fitted using a tensor product of 7 cubic regression splines for the month and 6 cyclic cubic regression splines for the hour within day. These bases dimensions allow for a reasonable number of knots over the hour and month axes, to describe and capture the varying diurnal cycle form one season to another. Employing a sensitivity analysis, smaller basis dimensions are believed not to be enough to capture the present variability and larger bases dimensions do not seem to significantly alter the results.



(a) (b) (c)

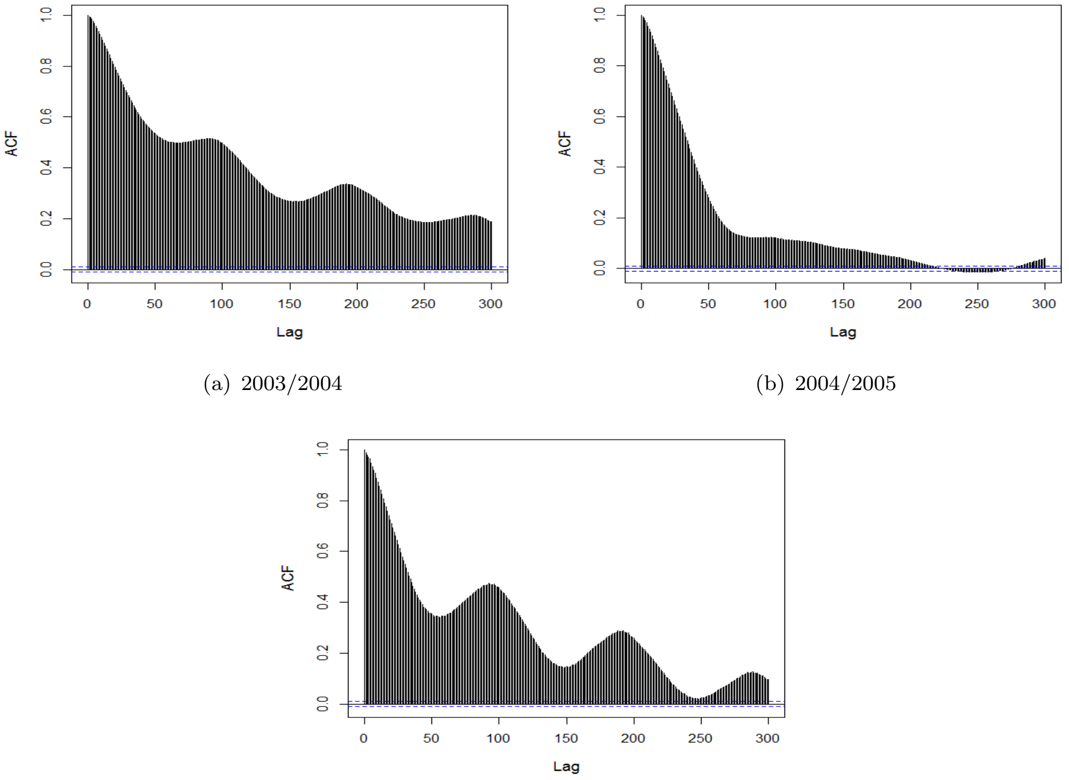
(d) (e) (f)

(g) (h) (i)

2003/2004 2004/2005 2005/2006

FIGURE 2.1: The estimated smooth functions $\hat{g}_1$, $\hat{g}_2$ and $\hat{g}_3$ of Time within year (a-c), Hour (d-f) and the interaction term between Hour and Month (g-i), respectively, of Model 2.24 for the HYs 2003/2004 (left), 2004/2005 (middle) and 2005/2006 (right). The dashed lines are the corresponding ± 2 s.e. bands and the blue dots represent a random sample of the partial residuals. Month 1 is October and Month 12 is September.

Figure 2.1 shows the fitted smooth curve of each covariate in Model 2.24. The top panels indicate the clear dissimilarities in the yearly trend and the within-year variability of $EpCO_2$ between the three HYs. The middle panels present the intra-daily cycle of $EpCO_2$ in each HY. It is evident that the efflux of $EpCO_2$ is generally lower during day time, explained by the increased consumption of CO_2 via photosynthesis during daylight hours. The bottom 2-D plots illustrate the interaction term between the diurnal and seasonal cycles of $EpCO_2$, from which the magnitude of the intra-daily cycle of $EpCO_2$ appears to change over time and is mainly stronger from March to July when a pronounced diel cycle is present. This model explains, on average, only around 55% of the total variability in $EpCO_2$ each year. In addition, the ACF of the models' residuals, in Figure 2.2, show not only significant correlations at high lags but also a remaining periodic pattern every 24 hours that is not captured by the model. This remaining dependence structure could affect the efficiency of the estimates, by underestimating their standard errors as mentioned in Section 2.3, and therefore has to be accounted for.



(a) 2003/2004

(b) 2004/2005

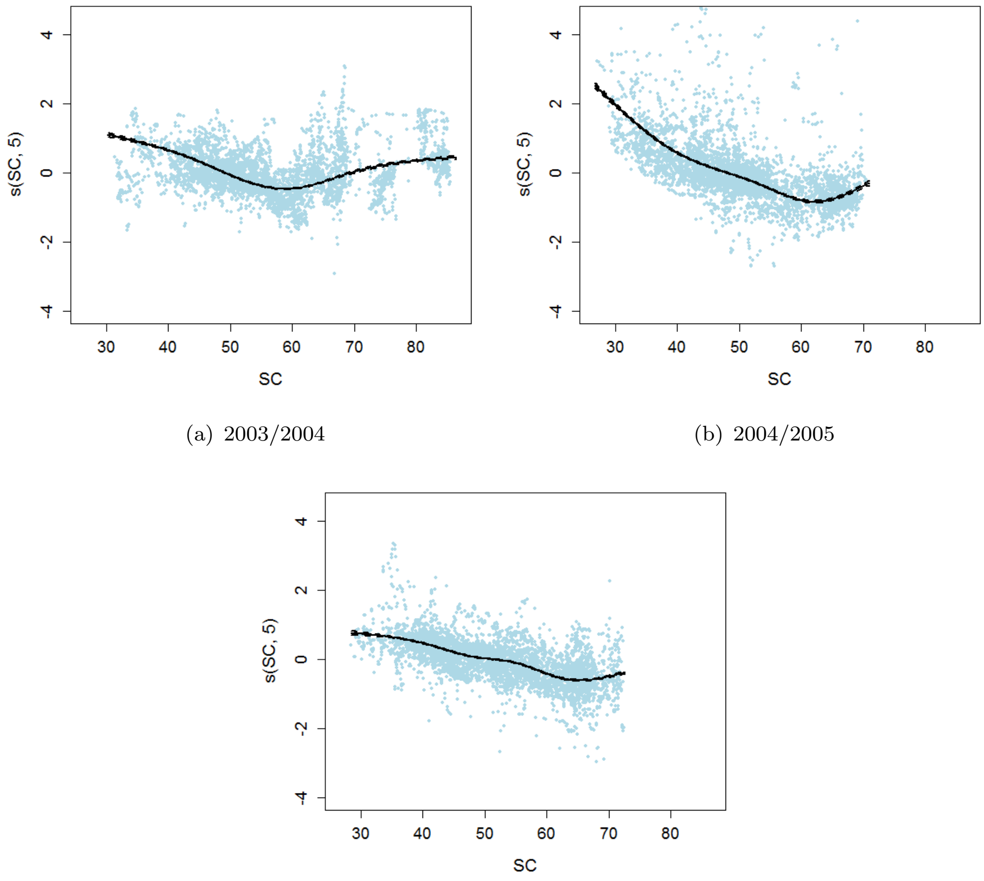
(c) 2005/2006

FIGURE 2.2: ACF of the residuals of Model 2.24 in the HYs (a) 2003/2004, (b) 2004/2005 and (c) 2005/2006.

Part of the evident strong and persistent auto-correlation structure in the residuals could be attributed to other explanatory variables not yet included in the model. Prior ecological knowledge and primary exploratory analysis have suggested that the variations in $EpCO_2$ are not solely related to seasonal and diurnal cycles but also to water hydrodynamics. Therefore, specific conductivity (SC), one of the continuously measured hydrological variables not used in the calculation of $EpCO_2$, is added to the set of explanatory variables describing the EpCO$_2$ in the following model:

$$\begin{aligned} y_t \quad = \quad & g_1(x_t^{\text{Time within year}}) + g_2(x_t^{\text{Hour}}) + g_3(x_t^{\text{Hour}}, x_t^{\text{Month}}) + \\ & g_4(x_t^{\text{Conductivity}}) + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2), \end{aligned} \tag{2.25}$$

where g_j, $j = 1, 2, 3$ are as defined above; and g_4 is approximated using 6 cubic regression splines to capture the main effect of specific conductivity. In addition to the previously observed temporal patterns, $\hat{g}_4$ depicts a (non-linear) negative relationship between EpCO$_2$ and SC up to a SC of ~60 $\mu S/cm$ (Figure 2.3) in all HYs. After adding SC to the set of explanatory variables, the new estimated model explains more than 65%, on average, of the EpCO$_2$ variability in each HY. However, the model residuals of each hydrological year still incorporate long range dependence and a periodic pattern that is not yet captured by the model.



(a) 2003/2004

(b) 2004/2005

(c) 2005/2006

FIGURE 2.3: The estimated smooth function $\hat{g}_4$ of SC in Model 2.25 for the HYs (a) 2003/2004, (b) 2004/2005 and (c) 2005/2006. The dashed lines are the corresponding ± 2 s.e bands and the blue dots are a random sample of the partial residuals of SC.

The ACFs of the residuals from Models 2.24 (Figure 2.2) and 2.25 emphasize the high dependence of $\mathrm{EpCO_2}$ on its previous values. This is mainly attributed to the nature of hydrological time series, in general, and the high-frequency nature of these data, in particular. This dependence structure should be appropriately accounted for when modeling to allow for reliable model inference. Another reason for this remaining complex correlation structure is the non-stationarity of the process and the inter-relationships between the explanatory variables not yet accommodated in the model. The former EDA has shown that the processes controlling the $\mathrm{EpCO_2}$ are time-scale dependent. That is, the $\mathrm{EpCO_2}$ does not only exhibit intra-daily, seasonal and annual fluctuations, but also its pattern and fluctuations change from day to day, from month to month and from year to year according to the underlying time and water hydrology characteristics. Therefore, in the following section, a set of additive models are fitted sequentially to model the variations in $\mathrm{EpCO_2}$ over a day, then over a month, and finally over a full hydrological year. This temporal hierarchy makes use of the models fitted at the lower scale to inform about the higher scale ones.

2.4.2 Temporal Hierarchical Modeling of $\mathrm{EpCO_2}$

Additive models are initially developed for individual days followed by individual months and finally for each hydrological year separately. This sequence of additive models is useful in explaining the variations in $\mathrm{EpCO_2}$ and studying the relationship between $\mathrm{EpCO_2}$ and SC reflecting river hydrology, within a day, a month and a hydrological year. They also describe the differences in variations between the different days, months and hydrological years. Another advantage of this temporal hierarchy is that it better shows the changes and the increased complexity of: (i) the processes driving $\mathrm{EpCO_2}$, (ii) the multi-variate interactions between $\mathrm{EpCO_2}$, water hydrology and time components, and (iii) the temporal correlation structures from the daily to yearly timescales.

2.4.2.1 Daily Additive Models

It is evident from the previous EDA and MRA, in Sections 1.3 and 1.5.7, that the $\mathrm{EpCO_2}$ exhibits a diel cycle with altering magnitude and pattern from one day to another. These alterations could be attributed to seasonal changes or other hydrological conditions. Let y_t denote the $\mathrm{EpCO_2}$ at time point t, and $\mathbf{x}_t = (x_t^{\text{Time within day}}, x_t^{\text{Hour}}, x_t^{\text{SC}})$ be the vector of explanatory variables at time t, where $x_t^{\text{Time within day}}$ is the value of a variable representing the time within the day (using hours and minutes on a decimal hour scale) at which the measurement is recorded, x_t^{Hour} is an index of the hour within day and x_t^{SC} is the measured SC at time t. Then, the daily variations of $\mathrm{EpCO_2}$ can be described

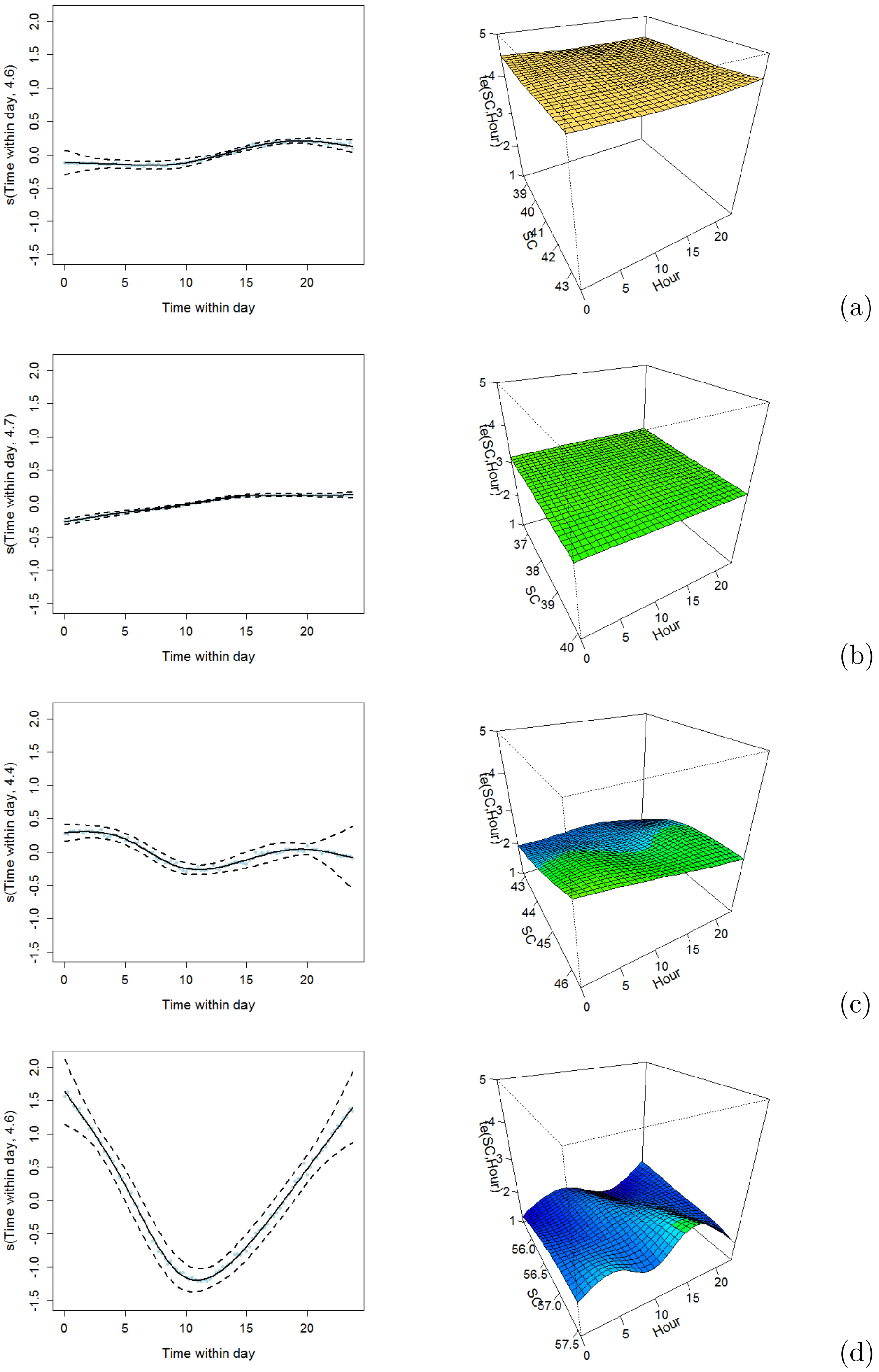


FIGURE 2.4: The estimated smooth functions $\hat{g}_1$ of Time within day (left) and $\hat{g}_3$ of the interaction term between SC and Hour of day (right) of the daily Model 2.26 fitted for the data of (a) 14/10/2005, (b) 14/1/2006, (c) 14/4/2006 and (d) 14/7/2006. The dashed lines are the corresponding ± 2 s.e bands and the blue dots are the partial residuals.

through the following additive model:

$$y_t = g_1(x_t^{\text{Time within day}}) + g_2(x_t^{\text{SC}}) + g_3(x_t^{\text{Hour}}, x_t^{\text{SC}}) + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2) \tag{2.26}$$

where the smooth functions g_j, $j = 1, 2, 3$, capture the daily cycle, the main effect of SC and the bivariate effect of hour within day and SC on EpCO$_2$, respectively; and ϵ_t accounts for the random effects not explained by the additive model. The functions g_1 and g_2 are represented using cubic regression spline bases with dimensions 6 and 5, respectively. Whereas, g_3 is approximated using a tensor product spline of two cubic regression splines with 5 basis functions for SC and 6 basis functions for the hour within day. As SC is not expected to widely vary within the day, a smaller basis dimension is specified for the smooth functions of SC. Whilst, the number of basis functions specified for the time/hour within day in g_1 and g_3 allows the smooth terms to capture the difference between day and night. The model is fitted to the data of some selected days. The model assumptions, including independence of the errors (see Figure 2.5), are shown to be all valid; and the estimated additive model explains about 99% deviance of the data of each selected day. This implies that the above model, to a great extent, has captured all the within-day variability.

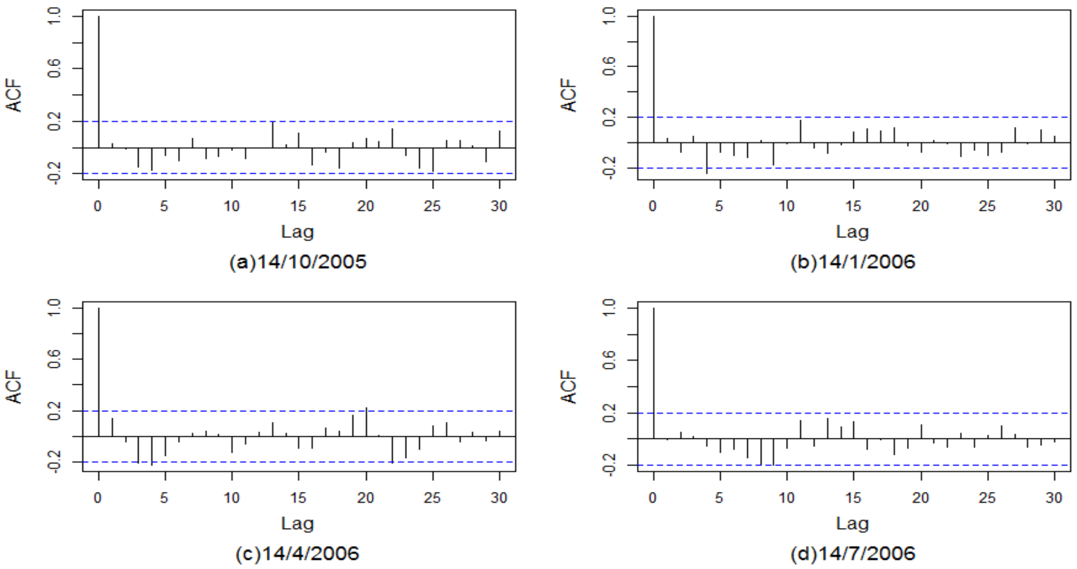


FIGURE 2.5: Residuals' ACFs of Model 2.26 fitted for the days (a) 14/10/2005, (b) 14/1/2006, (c) 14/4/2006 and (d) 14/7/2006.

Figure 2.4 and Table 2.1 show only the results of 14/10/2005, 14/1/2006, 14/4/2006 and 14/7/2006. As can be seen, the fitted smooth functions capture the response of EpCO$_2$ to time of day reflecting changes in the biological activity according to the day-light cycle. This intra-daily cycle of EpCO$_2$ changes from one day to another, according to the seasonal and hydrological conditions and is significantly stronger in the summer days. Despite the small number of observations within each day, the relationship between EpCO$_2$ and SC is significantly changing with time, justifying the multi-variate interactions between EpCO$_2$, hydrodynamics and time at the higher timescales.

	(a) 14/10/2005			**(b) 14/1/2006**		
Smooth terms	edf	F	P-value	edf	F	P-value
$g_1(x_t^{\text{Time within day}})$	4.6	11.4	<0.0001	4.7	85.3	<0.0001
$g_2(x_t^{\text{SC}})$	1	0.3	0.6	1	16.26	0.0001
$g_3(x_t^{\text{Hour of day}}, x_t^{\text{SC}})$	7.2	3.9	<0.0001	10.9	18.8	<0.0001
	(c) 14/4/2006			**(d) 14/7/2006**		
Smooth terms	edf	F	P-value	edf	F	P-value
$g_1(x_t^{\text{Time within day}})$	4.4	19.2	<0.0001	4.6	61.6	<0.0001
$g_2(x_t^{\text{SC}})$	1	13.8	0.0004	1	0.6	0.4
$g_3(x_t^{\text{Hour of day}}, x_t^{\text{SC}})$	8.7	14.4	<0.0001	12	13.8	< 0.0001

TABLE 2.1: Approximate significance of the smooth functions of the daily Model 2.26 fitted for the data of (a) 14/10/2005, (b) 14/1/2006, (c) 14/4/2006 and (d) 14/7/2006.

2.4.2.2 Monthly Additive Models

The EDA has identified seasonal differences in the behaviour and fluctuations of EpCO$_2$. These monthly/seasonal variations are explained via fitting the following additive model for each month of the hydrological year separately:

$$\begin{aligned} y_t \; = \; & g_1(x_t^{\text{Time within month}}) + g_2(x_t^{\text{Hour}}, x_t^{\text{SC}}) + \\ & g_3(x_t^{\text{Hour}}, x_t^{\text{Day of month}}) + g_4(x_t^{\text{Day of month}}, x_t^{\text{SC}}) + \epsilon_t, \end{aligned} \tag{2.27}$$

where $x_t^{\text{Time within month}}$ is the value of a continuous scale variable denoting the time of measurement within month and the other explanatory variables are as defined above. The function g_1 determines the main pattern of EpCO$_2$ within each studied month; and g_2 describes the bivariate effect of hour within day and SC. Based on the results of the daily additive models, the smooth functions g_3 and g_4 are added to the model to capture the changing effects of hour within day and SC from day to day, respectively. The smooth function g_1 is approximated using a cubic regression spline basis with a dimension of 9; and the bivariate smooth functions g_j, $j = 2, 3, 4$, are represented by tensor product splines. The bases' dimensions of the smooth functions of hour within day g_2 and g_3, day of month in g_3 and g_4 and SC in the bivariate terms g_2 and g_4 are set equal to 6,7 and 6, respectively, based on AIC and a sensitivity analysis.

Note that in the interaction terms of Model 2.27, the time variables 'Hour' and 'Day of month' are represented in integer form unlike the 'Time within month' which is defined on a continuous time scale. This is mainly to constraint the dimensionality of the interaction terms involving time components. These restricted integer time scales appear enough for modeling the mean pattern of EpCO$_2$ over the different time components. Because of the strong temporal correlation between the measurements, there was very little loss of information in using these restricted integer scales.

Only the model results of January and June 2006 are presented here due to space limitations. The estimated additive models explain 74% and 93% deviance of the data in January and June, respectively. However, the ACF of the model residuals shows a slowly decaying correlation structure in January (Figure 2.6(a)), and not only significant correlations at high lags but a remaining periodic pattern every 24 hours that is not captured by the model in June (Figure 2.6(b)). These dependence structures affect the efficiency of the estimates and make the inference procedure unreliable. To adjust the standard errors of the estimates by appropriately accounting for this dependence structure, a two-stage additive model fitting procedure is employed. In the winter months, turbulent waters and cold temperature reduce the biological activity and the external drivers such as climatological conditions become the only drivers of changes in EpCO$_2$. The effect of these external drivers usually fades off quite quickly and hence the auto-correlation in the residuals ϵ_t can be accounted for using simple time dependence structures. For example, in January, an autoregressive process of order 1, AR(1), is fitted to the estimated additive model residuals, which has entirely accounted for the remaining time dependence (Figure 2.6(c)). Conversely, in June, a greater degree of structure was displayed in the residuals after fitting the additive model, mainly attributed to the increased biological activity in summer.

AIC indicates that an AR of order 32 (8-hours lag), on average, is sufficient to account for this periodic dependence structure left in the residuals of the summer months. The AR order selected by the AIC coincides with the length of the dominant intra-daily cycle identified by the wavelet analysis. Nevertheless, EpCO$_2$ seems to heavily depend, in particular, on its preceding 2-hours measurements. Therefore, cubic regression splines of the 2- and 8-hours lagged dependent variables are first added to the model and the following model is fitted:

$$\begin{aligned} y_t =& g_1(x_t^{\text{Time within month}}) + g_2(x_t^{\text{Hour}}, x_t^{\text{SC}}) + g_3(x_t^{\text{Hour}}, x_t^{\text{Day of month}}) + \\ & g_4(x_t^{\text{Day of month}}, x_t^{\text{SC}}) + g_5(y_{t-8}) + g_6(y_{t-32}) + \epsilon_t, \qquad (2.28a) \\ \epsilon_t =& \rho\epsilon_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma_\varepsilon^2), \qquad (2.28b) \end{aligned}$$

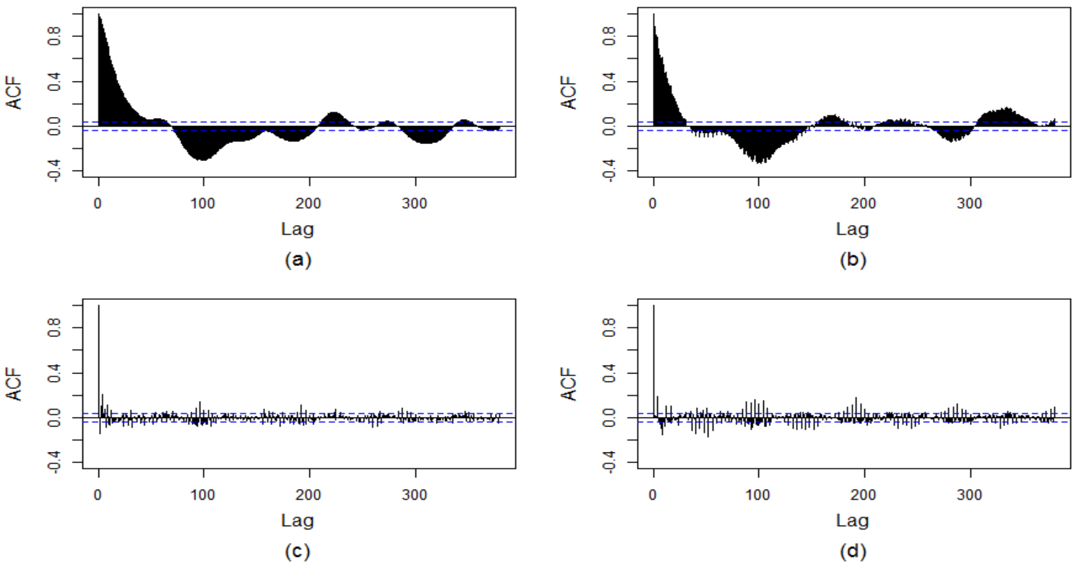
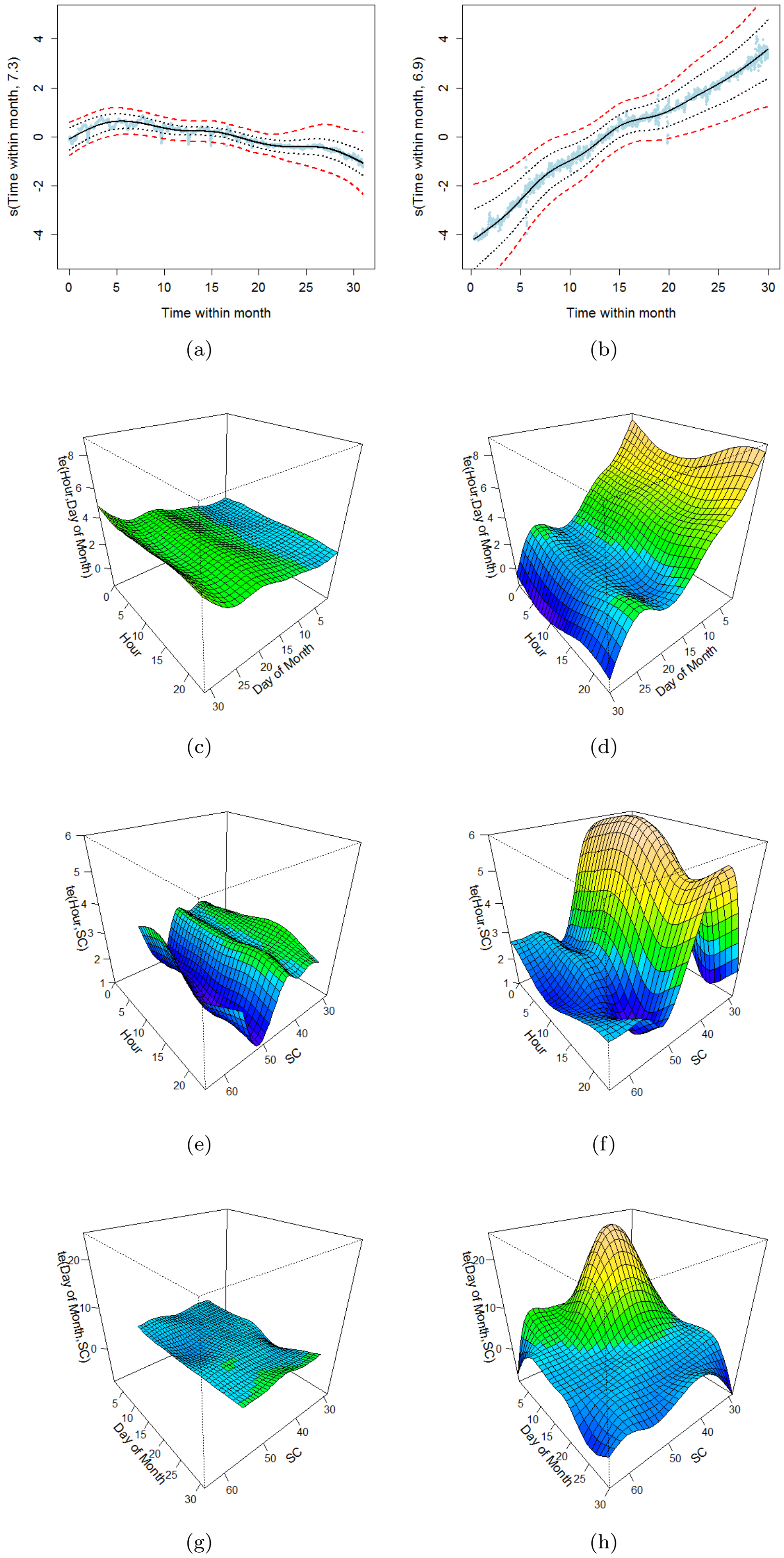


FIGURE 2.6: Residuals' ACFs of Model 2.27 fitted (a) January 2006 and (b) June 2006; and the corresponding ACFs after accounting for the temporal correlation in (c) January via an AR(1) process and in (d) June using the TSP via Model 2.28a - 2.28b.

where y_{t-8} and y_{t-32} are the 2- and 8- hours lagged EpCO$_2$, respectively. The 2-hours lag denotes the average extent of short-term dependence, while the 8-hours lag represents the extent of long-term dependence. Adding the 2- and 8-hours lagged EpCO$_2$ smooth terms to Model 2.27 account for the long range dependence and the periodic dependence structure present in the summer months, then any remaining autocorrelation is accounted for via an AR(1) process (Equation 2.28b, where ρ is the auto-regressive parameter) fitted to the residuals of the adjusted model. The final residuals $\hat{\varepsilon}_t$ appear independent (see Figure 2.6(d)) and the variability bands (± 2 s.e.) of the estimated smooth curves are accordingly adjusted, as explained in Section 2.3.1.

The monthly additive models indicate that the EpCO$_2$ dynamics vary across the different months of the hydrological year. Figure 2.7 illustrates the clear dissimilarities in the trend, variability of EpCO$_2$ and interactions with time and water hydrology between January and June. It is evident that the EpCO$_2$ is more variable in June. Figure 2.7 also shows the evidence of a strong intra-daily cycle in June that is absent in January. The EpCO$_2$ and the magnitude of its diel cycle changes significantly from one day to another within June. The figure indicates that the daylight cycle in June has a greater significant influence on the fluctuations of EpCO$_2$ than water hydrology during the absence of hydrological events (at high SC). Conversely, water hydrology dampens the diel cycle and dominates these fluctuations at periods of hydrological events. In January, EpCO$_2$ does not seem to exhibit an intra-daily cycle and variations are mostly attributed to hydrodynamics.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

FIGURE 2.7: The estimated smooth functions $\hat{g}_1$ of Time within month (a-b) and $\hat{g}_j$, $j = 1, 2, 3$ of the interaction terms between: Hour of day and Day of month (c-d); Hour of day and SC (e-f); and Day of month and SC (g-h) of the monthly additive models after accounting for the temporal auto-correlation in January 2006 (left) and June 2006 (right). The blue dots represent the partial residuals. The black dotted lines in panels (a-b) represent ±2 s.e. bands calculated before accounting for the temporal correlation in the residuals and the corresponding red dashed lines are ±2 s.e. bands calculated based on the adjusted standard errors after accounting for the temporal auto-correlation.

In conclusion, both time and water hydrology contribute to the variations in EpCO$_2$, such that the contribution of the temporal patterns and hydrology is time-dependent and changes from one season to another. Moreover, the temporal correlation remaining between the residuals after accounting for these variations varies seasonally and shows more complex structures in summer.

An alternative method to the two-stage fitting procedure of the additive model is to incorporate the correlation structure through an additive mixed model, as previously explained in Section 2.3.2. The additive mixed model, or more generally the GAMM, simultaneously fits an additive model and a mixed effects model to account for both the mean effect and the auto-correlation effect between the model residuals, respectively. This results in smaller degrees of freedom (i.e. more smooth curves) being selected (see Table 2.2) due to the confounding between non-linearity and correlation. Hence, the remaining correlation structure to be accounted for in the residuals increases, which increases the complexity of modeling required for the residuals. This, in turn, can increase the computational burden and lead to numerical instability in model fitting.

	January		**June***	
Smooth terms	GAM	GAMM	GAM	GAMM
$g_1(x_t^{\text{Time within month}})$	7.3	1.9	6.9	1
$g_2(x_t^{\text{Hour}}, x_t^{\text{SC}})$	25.4	14.8	27.9	7.7
$g_3(x_t^{\text{Hour}}, x_t^{\text{Day of month}})$	21.8	17.2	19.5	13.3
$g_4(x_t^{\text{Day of month}}, x_t^{\text{SC}})$	18.8	8.6	27	10.6
$g_5(y_{t-8})$			4.7	4.6
$g_6(y_{t-32})$			4.9	3.5

Table 2.2: Estimated effective degrees of freedom of the smooth terms in the monthly additive model and additive mixed model fitted for January and June, 2006. (*) In June, both the additive model and additive mixed model involve the 2-hour and 8-hour lagged EpCO$_2$.

2.4.2.3 Yearly Additive Models

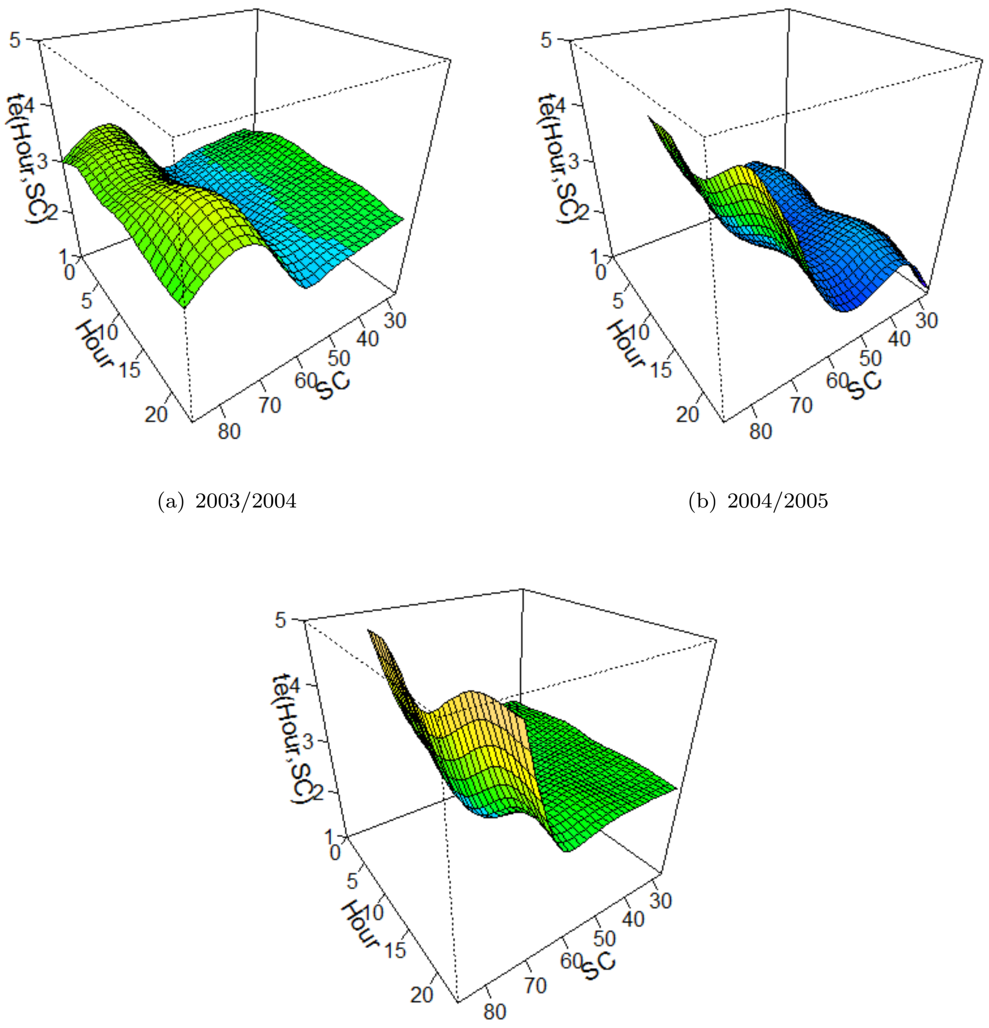
The monthly additive models illustrated intra-annual variations in EpCO$_2$. Moreover, the EDA indicated the non-stationarity of the full time series and the presence of inter-annual variations, as a result of the different climatological characteristics characterizing each HY. Therefore, the model is extended to describe the variations in EpCO$_2$ within each HY separately and highlight the differences between the three hydrological years. It is also noticed that as the model covers a longer time period, the auto-correlation structure becomes more difficult to model. Therefore, the yearly variations of EpCO$_2$

are described through the following TSP:

$$\begin{aligned} y_t =& g_1(x_t^{\text{Time within year}}) + g_2(x_t^{\text{Hour}}, x_t^{\text{Day of year}}) + g_3(x_t^{\text{Hour}}, x_t^{\text{SC}}) \\ &+ g_4(x_t^{\text{Day of year}}, x_t^{\text{SC}}) + g_5(y_{t-8}, \text{by} = \text{Month}) + g_6(y_{t-32}, \text{by} = \text{Month}) + \epsilon_t \end{aligned} \tag{2.29a}$$

$$\epsilon_t = \rho\epsilon_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma_\varepsilon^2), \tag{2.29b}$$

where y_{t-8} and y_{t-32} denote the 2- and 8- hours lagged EpCO$_2$, respectively, which were successful in accounting for the periodic dependence structure in the monthly models. The smooth function g_1 captures the global trend of EpCO$_2$ along each hydrological year



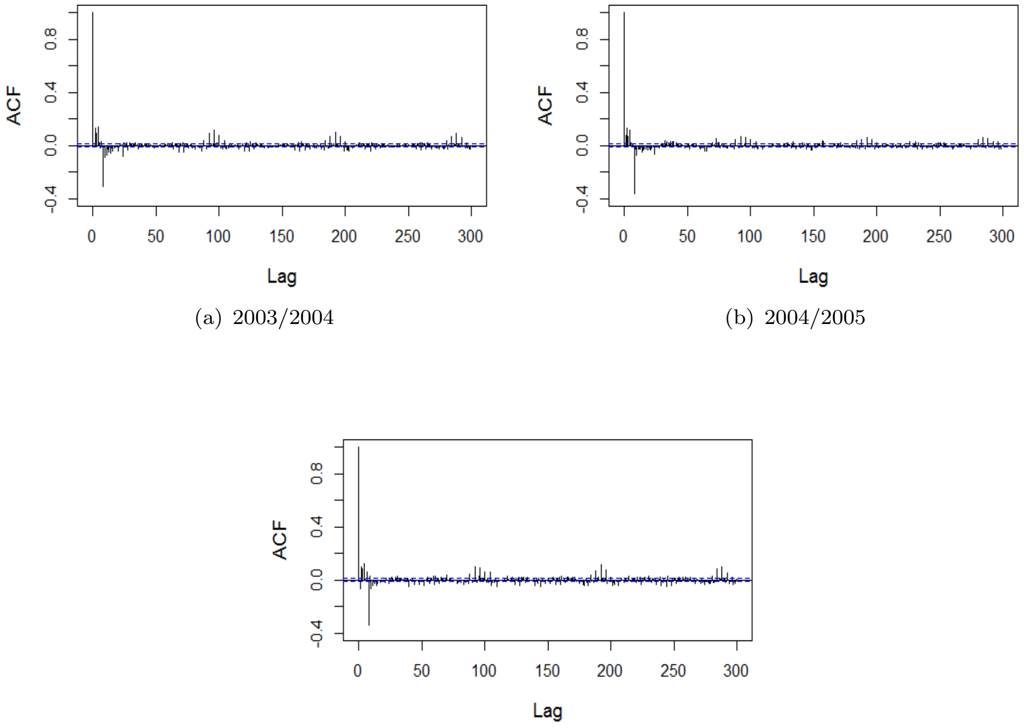
(a) 2003/2004

(b) 2004/2005

(c) 2005/2006

FIGURE 2.8: The estimated smooth functions $\hat{g}_3$ of the interaction term between Hour and SC of the yearly additive Model 2.29a - 2.29b fitted for the HYs (a) 2003/2004, (b) 2004/2005 and (c) 2005/2006.

using a cubic regression spline basis with a dimension of 15; g_2 describes the changing effect of the daily cycle from day to day; g_3 and g_4 explain the bivariate effect of SC with hour and day of year, respectively; and g_5 and g_6 capture the changing effect of the 2- and 8- hours lagged EpCO$_2$ on the current EpCO$_2$ from month to month, respectively. As previous, g_j, $j = 2, 3, 4$ are approximated using tensor product splines. The bases' dimensions specified for the hour within day in g_2 and g_3, the day of year in g_2 and g_4 and the SC in g_3 and g_4 are 6, 7 and 6, respectively; which appear to be reasonable choices based on both AIC and a sensitivity analysis. Since the dependence of EpCO$_2$ on its previous values vary from one month to another, the smooth functions g_5 and g_6 are represented by cubic regression splines for each month, generating a different smooth for each month. The residuals ϵ_t in Equation 2.29a follow an AR(1) (see Equation 2.29b, where ρ is known as the autoregressive parameter and ε_t is a white noise process with mean 0 and variance σ^2_ε).



(a) 2003/2004

(b) 2004/2005

(c) 2005/2006

FIGURE 2.9: ACFs of the residuals ε_t in Equation 2.29b for the HYs (a) 2003/2004, (b) 2004/2005 and (c) 2005/2006.

The fitted additive models explain about 95%, on average, of the variability in EpCO$_2$ in each hydrological year. It is evident that the diel cycle is dominating the changes in EpCO$_2$ at high SC levels (Figure 2.8) i.e. at low flow when in-stream biological processes are most dominant. By incorporating lagged dependent variables in the model, the $EpCO_2$ for the three years seems to exhibit the same patterns but with different magnitude. It is also clear that the autocorrelation in the residuals has been substantially reduced after adding the lagged $EpCO_2$ to the additive model and modeling the residuals via an AR process. However, a little periodic structure is still evident, see Figure 2.9.

In brief, it is evident that the processes controlling the $EpCO_2$ are time and scale dependent. The multi-variate relationships between the $EpCO_2$, water hydrology and time components change from one scale to another and become more complex when the model is extended to describe a longer time period within the hydrological year. In addition, the autocorrelation structure between the residuals remaining after accounting for the temporal and water hydrological changes with time and becomes more persistent and composite at the yearly scale. Therefore, lagged variables and more multi-variate interactions are added to explain the increased variability and account for the persistence of temporal correlations at the larger timescales.

2.5 Summary and Discussion

It is evident that although hydrological high-frequency data involve previously inaccessible information, they pose various challenges to statistical modeling and analysis. This has been demonstrated here using the high-resolution time series of $EpCO_2$ as an illustrative data set. Exploring and modeling these high-resolution sensor data was very complex and challenging because of the differences in the behaviour of the variable of interest over the different timescales, the complex multi-variate relationships which are time and scale dependent, the signal to noise ratio, and the persistent temporal correlation characterizing such hydrological high-frequency data.

The primary EDA showed that $EpCO_2$ is possibly non-stationary and exhibits variations over a wide range of timescales. These variations arise due to changes in the relative strength of external (e.g. climatological) and internal (biological processing) drivers of resultant $EpCO_2$. $EpCO_2$ is generally higher in summer than winter and more variable during summer due to the greater catchment productivity in summer when more CO_2, or sources of, is available, and greater in-stream processing of C results in CO_2 consumption (during day-time) and production (during night-time). This processing can be seen in the intra-daily cycle of $EpCO_2$, which is lowest close to midday (maximum solar radiation to support photosynthesis) and highest just after midnight, when respiration has occurred for longest and so CO_2 concentration is highest. According to the MRA presented in Chapter 1, this intra-daily cycle appeared to be the major time-scale contributor to the variability of the $EpCO_2$ series, which is also not constant throughout the year but larger variability occurs during summer when a pronounced diurnal cycle is present.

The hierarchal additive models fitted over a day, a month and a year showed that the variability of $EpCO_2$ and its relationship with water hydrology are time and scale dependent. The additive models allow temporal variations and the mechanisms controlling $EpCO_2$ across the different timescales to be accommodated. These temporal variations and multi-variate relationships change across the different timescales and become more complex as the model is extended to cover a longer time period within the hydrological year. These additive models indicated that the $EpCO_2$ exhibits a 24-hour dark-light-dark cycle reaching the minimum value at noon. The magnitude of this day/night cycle changes along the year and is more apparent during the summer (May - September) where the $EpCO_2$ reaches its maximum levels. It was also obvious that the water hydrology has an influence on the level of $EpCO_2$. At low flow, DIC concentration is highest (Waldron et al., 2007) and biological activity is greatest (as temperature tends to be higher); event flow, whilst flushing out soil CO_2 so increasing the pool size, ultimately dilutes the DIC pool and so lowers saturation of dissolved carbon dioxide. Turbulent waters and colder temperature reduce biological activity. As such CO_2 over-saturation is reduced and $EpCO_2$ decreases. The diel cycle can still exist, but variability is reduced in winter. Seasonality in flow thus has a significant effect on the $EpCO_2$. The diel cycle appears to dominate the $EpCO_2$ variations in summer during the absence of hydrological events and during low flows (evidenced by higher SC), while high flow events dampen the diel cycle in winter. Hence, the contribution of the temporal and hydrological variations changes with season and timescale. Consequently, the fitted additive models encountered some problems in uniquely identifying the sources of variability and the contribution of each variable to the variability in $EpCO_2$.

Serial autocorrelation is one of the characteristics of high-resolution time series. The residuals of the fitted additive models displayed a periodic autocorrelation structure that persists over a large number of lags. The complexity of the dependence structure increases from daily to yearly timescales. Therefore, modeling these hydrological high-frequency time series by assuming independence is no longer valid. A two-stage fitting procedure has been used, where some lagged dependent variables are added to the model and then an AR process is fitted to the adjusted model residuals. The alternative method of incorporating the correlation structure through fitting a GAMM has been considered as well, and the results of both methods have been compared. The GAMM simultaneously fits a GAM, or an additive model, and a mixed effect model that accounts for the autocorrelation between the model residuals. Incorporating autocorrelation through GAMM results in larger smoothing parameters and smaller degrees of freedom (i.e. smoother functions) being selected and hence the remaining structure to be accounted for in the residuals increases. This increases the complexity of modeling required for the residuals. Whereas in the TSP, the first stage results in optimally selecting smaller smoothing parameters assuming independent errors, which reduces the complexity of modeling required for the residuals in the second stage. After incorporating lagged terms of the dependent variable in the model and a simple correlation structure for the errors, only a small amount of structure is still remaining between the residuals of the yearly models.

In general, the additive models fitted to the high-frequency time series of $\mathrm{EpCO_2}$ highlighted: (i) the complex multi-variate interactions between the covariates and the response variable, (ii) the complex correlation structures persisting over a large number of lags between observations due to the high-frequency nature of the data, and (iii) the identifiability problems in allocating the existing large variability to the signal or noise as a result of the confounding between correlation and non-linearity. The following chapter tries to account for those challenges and presents a different approach for analyzing hydrological high-frequency data. This approach is known as functional data analysis.

Chapter 3

Functional Data Analysis

Typically, real life processes in medicine, meteorology, economics, environment and many others domains are continuous in time. With the enormous technical advances in data collection and storage, it becomes easier to observe and process large amounts of data at arbitrarily high-frequency. This enables us to form a more complete picture of such processes, which would not have been possible in the past with lower frequency data. The previous chapter has shown that although high-frequency data could promote better and deep understanding of environmental process, they pose various challenges in terms of statistical modeling and analysis. These challenges involve the ability to analyze and extract useful information from such big volumes of data, the computational inefficiency, the complex inter-relationships between the determinants of the environmental system, the long-range correlation structure between observations and the high dynamics and non-stationarity of the process. These challenges have been identified and illustrated using the high-resolution time series of $\mathrm{EpCO_2}$.

The primary analysis, including wavelets and GAMs presented in Chapters 1 and 2, have indicated that $\mathrm{EpCO_2}$ exhibits an intra-daily cycle, which is responsible for the major variability in the $\mathrm{EpCO_2}$ series. This intra-daily cycle also varies over time according to the underlying climatological and hydrological conditions. In ecology, it is of interest to study the different intra-daily patterns of $\mathrm{EpCO_2}$ and the main internal and external drivers of each pattern. This fact has motivated our decision to study the intra-daily cycle of $\mathrm{EpCO_2}$ using a functional data analysis approach.

Functional Data Analysis (FDA) (Ramsay and Silverman, 1997) has recently proven to be an appropriate statistical tool for analyzing large volumes of high-frequency data and has subsequently become an attractive and promising field of statistical research. Functional equivalents for many of the standard statistical methods have been introduced. In this chapter, we consider in particular functional principal component analysis as a tool to identify the primary modes of variations in EpCO$_2$ and functional clustering approaches to classify the different intra-daily patterns of EpCO$_2$.

This chapter starts with an introduction to functional data analysis and how to transform discrete data to functional data and perform functional exploratory data analysis. Afterwards, the functional principal component analysis is introduced, then the different approaches of functional clustering are presented. At the end, the results of analyzing EpCO$_2$ data using a functional data analysis approach are presented and the statistical issues of using the traditional functional principal components and functional clustering techniques to analyze such high-frequency time series are discussed.

3.1 Functional Data Analysis (FDA)

FDA is a popular technique used for analyzing data collected as multiple time series, where each series is observed frequently/continuously in time. That is, each time series is viewed as observations of a continuous function collected at a finite series of time points. In this setting, observations are functions and the fundamental unit of interest is the entire function or curve constructed from the observations collected over time. In some applications, each individual function is considered a segment of a longer time series. For example, each function here represents a day and hence the traditional FDA approaches will treat each function/curve as a single entity without being concerned about the correlations between the measurements within the day. Analyzing the EpCO$_2$ data in this format also permits the observation and analysis of the daily patterns of EpCO$_2$, disregarding any problems related to irregularly-spaced or missing data.

Statistically, functional data are considered as realizations of smooth random curves such that each observation x_i is a curve $x_i(t) : i \in \mathbb{Z}, t \in \mathcal{T}$, where $\mathcal{T}$ is the continuous time domain. Each curve $x(t)$ belongs to the Hilbert space of square integrable functions defined on $\mathcal{T}$, $L^2(\mathcal{T})$, with the inner product $\langle f, g\rangle = \int_{\mathcal{T}} f(t)g(t)dt, \forall f, g \in L^2(\mathcal{T})$. In practice, functional data are observed at discrete time points, possibly at high-frequency. That is, a functional observation $x_i(t)$ consists of m_i pairs (t_{ij}, x_{ij}), $j = 1, \ldots, m_i$, where x_{ij} is the observed value of the i^{th} sample path, $x_i(t)$, at the time t_{ij}. Or alternatively, the observations x_{ij} are viewed as discrete numerical representations of the infinite-dimensional objects $x_i(t)$. That is, FDA is concerned about analyzing the curves $x_i(t)$ and not the discrete high-frequency data x_{ij}. For that reason, FDA is considered a computationally efficient data dimensionality reduction technique.

3.1.1 Discrete Data to Functional Data

The first and most crucial step in FDA is to construct the smooth functional curves from their corresponding discrete observations. There are numerous techniques of converting the noisy and raw data into smooth functional data including parametric and non-parametric techniques, which have been presented earlier in Chapter 2. A detailed discussion of the different techniques used for the construction of functional data is available by Ramsay and Silverman (1997) and a description of how these techniques are implemented in the R software package is presented by Ramsay and Graves (2009).

Let the functional data be observed with error such that:

$$x_{ij} = x_i(t_{ij}) + \epsilon_{ij}, \quad i = 1, \dots, N; \qquad j = 1, \dots, m_i, \tag{3.1}$$

where the noise or the measurement error ϵ_{ij} is filtered out by representing the raw data as smooth functions. Then, let the sample paths $x_i(t)$, $i = 1, \dots, N$ belong to a finite-dimensional space generated by a set of basis functions $\{\psi_1(t), \dots, \psi_p(t)\}$ as follows:

$$x_i(t) = \sum_{k=1}^{p} a_{ik}\psi_k(t) = \boldsymbol{\psi}(t)^\top \mathbf{a}_i, \tag{3.2}$$

where $\mathbf{a}_i$ is the vector of basis coefficients $(a_{i1}, \dots, a_{ip})^\top$ to be estimated for the i^{th} sample path and $\boldsymbol{\psi}(t)$ is the vector of the basis functions $(\psi_1(t), \dots, \psi_p(t))^\top$. The goal is to fit a smooth function $x_i(t)$ from each vector of discrete observations $\mathbf{x}_i = (x_{i1}, \dots, x_{im_i})$ by assuming model 3.1 and using the basis expansion 3.2 for each of the N observed sample curves.

To smooth the sample curves, there are multiple choices of basis functions including splines, Fourier series and wavelets. The choice of the basis and its dimension are initially based on the characteristics of the data and the objective of the study (Ramsay and Silverman, 1997). For instance, a Fourier basis is designed for periodic data, while a B-splines basis is a very popular choice for smoothing non-periodic data with continuous derivatives up to certain order. The most common smoothing methods are the spline-based techniques, as they offer flexibility and computationally efficient means of storing information on functions (Ramsay and Silverman, 1997). In addition, spline basis functions enable the required calculations concerning functions to be expressed in a way such that the usual matrix algebra applies. As previously highlighted in Chapter 2, the most commonly used splines are cubic B-splines as they offer an appropriate amount of flexibility.

The coefficients $\mathbf{a}_i$ of the sample path $x_i(t)$ can be estimated by minimizing the following sum of squares:

$$(\mathbf{x}_i - \Psi_i \mathbf{a}_i)^\top (\mathbf{x}_i - \Psi_i \mathbf{a}_i),$$

where $\Psi_i = (\psi_k(t_{ij}))_{m_i \times p}$. Thus, the least squares estimate of $\mathbf{a}_i$ is $(\Psi_i^\top \Psi_i)^{-1} \Psi_i^\top \mathbf{x}_i$ and the fitted curves are known as regression splines. As explained in Chapter 2, the degree of smoothness of these regression splines is determined by the basis dimension which is a complicated discrete process. For this reason, penalized regression splines are more preferred. In penalized regression splines, the roughness penalty term $\lambda \mathbf{a}_i^\top \mathbf{D} \mathbf{a}_i$ is added to the above sum of squares and the coefficients $\mathbf{a}_i$ are then estimated by $(\Psi_i^\top \Psi_i + \lambda \mathbf{D})^{-1} \Psi_i^\top \mathbf{x}_i$. This approach provides a good fit to the data and a more flexible smooth curve. A second order roughness penalty $\mathbf{D}$ suggests that the smooth function minimizing the penalized sum of squares is a cubic B-spline with knots at the sampling points (Wood, 2006). As in GAMs, the smoothing parameter λ can be either selected using one of the automatic procedures including GCV, AIC, BIC, etc. or implicitly estimated using maximum likelihood methods.

3.1.2 Exploratory Functional Data Analysis

Fundamentally, the initial step in any statistical analysis is exploratory data analysis. The summary statistics for univariate data such as the mean and variance have equivalents for functional data. Let $x_i(t), i = 1, \ldots, N$ be a set of N curves, then the functional mean curve is defined by the point-wise average across the N functions:

$$\bar{x}(t) = \frac{1}{N} \sum_{i=1}^{N} x_i(t). \tag{3.3}$$

Similarly, the variance function is defined by the point-wise sample variance of the curves as follows:

$$\mathrm{VAR}_X(t) = \frac{1}{N-1} \sum_{i=1}^{N} \Big[x_i(t) - \bar{x}(t) \Big]^2 .$$

In addition, the dependence between values across different time points can be summarized using the covariance function. The covariance function between any pair of time points t_1 and t_2 is computed by:

$$\mathrm{COV}_X(t_1, t_2) = \frac{1}{N-1} \sum_{i=1}^{N} \Big[x_i(t_1) - \bar{x}(t_1) \Big] \Big[x_i(t_2) - \bar{x}(t_2) \Big],$$

and the associated correlation function is given by:

$$\rho_X(t_1, t_2) = \frac{\text{COV}_X(t_1, t_2)}{\sqrt{\text{VAR}_X(t_1)\text{VAR}_X(t_2))}}.$$

In addition to the above functional summary statistics, visualization methods are often quite useful in displaying the data, highlighting their characteristics and special features. Therefore, with analogy to classical analysis, many exploratory plots have been developed as tools for initially exploring and analyzing functional data. Such exploratory plots are fundamentally used to describe the common patterns or detect peculiar functional observations.

3.1.3 Functional Outliers

As in classical statistical analysis, identifying outliers is crucial before doing any further analysis. To detect functional outliers, Hyndman and Shang (2010) proposed two visualization tools, the functional bagplot and the functional highest density region box-plot. These two plots are based on the first principal components scores. Alternatively, Sun and Genton (2011b) developed a graphical method called the functional box-plot that is based directly on the functional space. By analogy to the univariate box-plot, the first step in constructing a box-plot is ordering the observations. For functional data, Lopez-Pintado and Romo (2009) introduced a graphical depth measure called the "band depth" to order a sample of curves by measuring how deep (central) a curve is. Let $x_{[i]}(t)$ denote the sample curve corresponding to the i^{th} largest band depth value, such that $x_{[1]}(t)$ is the most central curve with the largest band depth value and $x_{[N]}(t)$ is the most outlying curve with the smallest band depth value. If the curve $x(t)$ is regarded as a subset of the plane $G(x) = \{(t, x(t)) : t \in \mathcal{T}\}$, the band $\mathbf{B}$ in $\mathbb{R}^2$ bounded by the curves $x_{i_1}(t), \ldots, x_{i_k}(t)$ is given by:

$$\mathbf{B}(x_{i_1}, \ldots, x_{i_k}) = \{(t, x(t)) : t \in \mathcal{T}, \text{min}_{r=1,\ldots,k} x_{i_r}(t) \leq x(t) \leq \text{max}_{r=1,\ldots,k} x_{i_r}(t)\},$$

and the band depth (BD) of the curve $x(t)$ is defined as:

$$\text{BD}_J(x) = \sum_{j=2}^{J} \text{BD}^{(j)}(x),$$

where $2 \leq J \leq N$ is the number of curves defining a band and $\text{BD}^{(j)}(x)$ is the fraction of bands defined by j different curves containing the whole graph of the curve $x(t)$. Lopez-Pintado and Romo (2009) also proposed a more flexible depth measure called the “modified band depth”, obtained by measuring the proportion of time a curve $x(t)$ falls in the band.

Sun and Genton (2011b) extended the classical definition of the interquartile range (IQR), median and box-plot whiskers to functional data to construct the functional box-plot. Based on either the band depth or its modified version, the α $(0 < \alpha < 1)$ central region is estimated as the band delimited by the α proportion of the deepest sample curves. Thus, the IQR is defined as the envelope delimiting the 50% central region and the median curve is the most central curve $x_{[1]}(t)$. Similar to standard box-plots, the fences or whiskers of a functional box-plot are obtained as 1.5×IQR, and any curve outside these fences is marked as an outlier. These functional box-plots have also been adjusted for correlation, by allowing the factor 1.5 to change based on the covariance structure in the data, see Sun and Genton (2011a) for details.

There exists two types of functional outliers: *magnitude* outliers and *shape* outliers. Generally, magnitude outliers are distant from the mean curve while shape outliers have a different pattern from the other surrounding curves. Sun and Genton (2011b) claimed that the band depth basically depends on the shape of the curves while its modified version determines how central a curve is based on its magnitude. This is because, unlike the band depth, the modified band depth is computed as the proportion of times a curve lies within the band which results in less ties (similar BD values). Accordingly, the band depth and its modified version are used to detect shape and magnitude outliers, respectively.

3.2 Functional Principal Component Analysis (FPCA)

FPCA is considered a key technique in FDA for its leading role in exploring, extracting and summarizing the features characterizing a set of curves. Typically, Principal Component Analysis (PCA) is a very popular dimensionality reduction technique that captures the maximum information present in the original data and simultaneously minimizes the difference between the original data and the new reduced dimensional representation, see Jolliffe (2005) for more details. PCA was one of the first classical multivariate methods to be adapted to functional data, see Ramsay and Silverman (1997). With analogy to multivariate PCA, FPCA has proven to be a very useful tool in identifying the primary modes of variation in a set of functional curves $x_i(t), i = 1, \ldots, N$ after adjusting for the average smooth curve $\bar{x}(t)$ and optimally representing the functional data into a function space of reduced dimension (Ramsay and Silverman, 1997). For example, Henderson (2006) has used an FPCA approach to identify the primary sources of variations in some sediment and nutrient trend data for a number of monitoring sites in three dams in South East Queensland, Australia. FPCA was also employed to analyze the directions of variability in the Canadian temperature data after adjusting for the average yearly temperature across all the weather monitoring stations (Ramsay and Silverman, 1997).

Ramsay and Silverman (1997) and Ramsay and Dalzell (1991) are the main references for the theory and computation of FPCA in this section. Let $X(t)$ be an $L^2(\mathcal{T})$ continuous stochastic process and let $\mu = \{\mu(t) = \mathbb{E}[X(t)]\}_{t\in\mathcal{T}}$ denotes the mean function and $\mathcal{V}$ denotes the covariance operator of $X(t)$ such that:

$$\begin{aligned}\mathcal{V}: \quad L^2(\mathcal{T}) &\rightarrow L^2(\mathcal{T})\\ x \rightarrow \mathcal{V}x &= \int_0^T V(.,t)x(t)dt.\end{aligned}$$

The operator $\mathcal{V}$ is an integral transform with kernel V defined by:

$$V(s,t) = \mathbb{E}[(X(s)-\mu(s))(X(t)-\mu(t))], \quad s,t \in \mathcal{T}$$

Under the L^2 continuity hypothesis, both the mean and covariance functions are continuous.

The spectral decomposition of the covariance operator $\mathcal{V}$ yields a countable set of positive eigenvalues $\{\gamma_m\}_{m\geq 1}$ associated to an orthonormal basis of eigenfunctions $\{\xi_m\}_{m\geq 1}$:

$$\mathcal{V}\xi_m = \gamma_m \xi_m \tag{3.4}$$

such that $\gamma_1 \geq \gamma_2 \geq \ldots$ and $\int \xi_m(t)\xi_{m'}(t)dt = 1$ for $m = m'$ (normalization constraint) and 0 otherwise (orthogonality constraints).

The principal component scores $\{S_m\}_{m\geq 1}$ of $X(t)$ are random variables defined as the orthogonal projection of $X(t)$ on the eigenfunctions of $\mathcal{V}$:

$$S_m = \int_{\mathcal{T}} (X(t)-\mu(t))\xi_m(t)dt, \tag{3.5}$$

where S_m are zero-mean independent random variables with variance γ_m for $m \geq 1$. Accordingly, the Karhunen-Loève expansion holds also for functional data and $\lim_{M\rightarrow\infty} \mathbb{E}\Big[\big(X(t)-\mu(t)-\sum_{m=1}^{M} S_m\xi_m(t)\big)^2\Big] \rightarrow 0$. Hence, $X(t)$ admits the following representation in L^2:

$$X(t) = \mu(t) + \sum_{m\geq 1} S_m\xi_m(t), \quad t \in \mathcal{T}. \tag{3.6}$$

That is, the original process $X(t)$ can be reconstructed by multiplying the principal component scores with their corresponding eigenfunctions and adding up the mean curve.

Each function of this process $X(t)$ can be approximated with a minimum loss of information, with respect to the usual L^2-norm, by truncating Equation 3.6 at the first q terms as follows:

$$X(t) \approx \mu(t) + \sum_{m=1}^{q} S_m \xi_m(t), \quad t \in \mathcal{T}. \tag{3.7}$$

To solve the continuous eigenfunction analysis in Equation 3.4 and compute the functional principal components given by Equation 3.5, the continuous functional eigenanalysis has to be converted into an approximately equivalent matrix eigen-analysis problem. This conversion is addressed in details in the rest of this section.

Let $\{x_1(t), \ldots, x_N(t)\}$ be the observed functional data of the process $X(t)$ and $\bar{x}(t)$ is as defined by Equation 3.3. For each sample path $x_i(t)$, define $z_i(t) = x_i(t) - \bar{x}(t)$ such that the covariance kernel V of the operator $\mathcal{V}$ is estimated by:

$$\hat{V}(s,t) = (N-1)^{-1} \sum_{i=1}^{N} z_i(s) z_i(t).$$

The functional principal components $\{\hat{\xi}_m(t)\}_m \geq 1$ are then the solution of the continuous eigen-equation (Ramsay and Silverman, 1997):

$$\int_0^T \hat{V}(s,t)\hat{\xi}_m(t)dt = \gamma_m \hat{\xi}_m(s). \tag{3.8}$$

The integral in this continuous functional eigen-equation is difficult to solve in closed form. A general strategy to solve it is to convert the functional eigen-equation to the usual discrete or matrix eigen-analysis. One way to do this conversion is to represent each centered function $z_i(t)$ as linear combination of its p basis functions $z_i(t) = \sum_{k=1}^{p} a_{ik}\psi_k(t)$. Define the vector-valued functions $\mathbf{z}(t) = (z_1(t), \ldots, z_N(t))^T$ and $\boldsymbol{\psi}(t) = (\psi_1(t), \ldots, \psi_p(t))^\top$, then the joint expansion for all N curves in vector-matrix notation is given by:

$$\mathbf{z}(t) = A\boldsymbol{\psi}(t), \tag{3.9}$$

where A is an $N \times p$ matrix of basis coefficients, whose rows are the vectors of coefficients $\mathbf{a}_i^\top = (a_{i1}, \ldots, a_{ip})$. Using the basis expansion in Equation 3.9, the estimator of the variance-covariance function, $\hat{V}$ for all $s, t \in \mathcal{T}$, can be expressed in matrix form as:

$$\hat{V}(s,t) = \frac{1}{N-1}\boldsymbol{\psi}(s)^\top A^\top A \boldsymbol{\psi}(t). \tag{3.10}$$

Similarly, each eigenfunction $\hat{\xi}_m(s)$ can be expanded as:

$$\hat{\xi}_m(s) = \sum_{k=1}^{p} b_{mk}\psi_k(s) = \boldsymbol{\psi}(s)^\top \mathbf{b}_m, \tag{3.11}$$

where $\mathbf{b}_m = (b_{m1}, \ldots, b_{mp})$ is the corresponding vector of basis coefficients.

Then, the continuous eigen-analysis problem in Equation 3.8, by replacing $\hat{V}(s,t)$ and $\xi_m(s)$ with their expressions given in 3.11 and 3.10, becomes equivalent to:

$$\frac{1}{N-1}\boldsymbol{\psi}(s)^\top A^\top A\mathbf{W}\mathbf{b}_m = \gamma_m\boldsymbol{\psi}(s)^\top\mathbf{b}_m,$$

where $\mathbf{W} = \int_0^T \boldsymbol{\psi}(t)\boldsymbol{\psi}(t)^\top dt$ is a $p \times p$ symmetric matrix of the inner products between the basis functions obtained by numerical integration. Since this equation holds for all s, it can be rewritten as:

$$\frac{1}{N-1}A^\top A\mathbf{W}\mathbf{b}_m = \gamma_m\mathbf{b}_m,$$

subject to the normalization constraint $\|\hat{\xi}_m\| = \mathbf{b}_m^\top\mathbf{W}\mathbf{b}_m = 1$ and the orthogonality constraints $\mathbf{b}_m^\top\mathbf{W}\mathbf{b}_{m'} = 0$ for all $m \neq m'$. To get the required principal components, define $\mathbf{u}_m = \mathbf{W}^{1/2}\mathbf{b}_m$ where $\mathbf{W}^{1/2}$ is the Cholesky decomposition of the matrix $\mathbf{W}$, and solve the equivalent symmetric eigen-equation:

$$\frac{1}{N-1}\mathbf{W}^{1/2}A^\top A\mathbf{W}^{1/2\top}\mathbf{u}_m = \gamma_m\mathbf{u}_m,$$

The coefficients vector $\mathbf{b}_m$, $m \geq 1$, of the eigenfunction $\hat{\xi}_m(t)$ is estimated by $\hat{\mathbf{b}}_m = (\mathbf{W}^{1/2})^{-1}\mathbf{u}_m$ (such that $\mathbf{u}_m^\top\mathbf{u}_{m'} = 1$ if $m = m'$ and 0 otherwise) and the estimated principal component scores $\hat{S}_{mi}$, which are the principal component values for a specific observation i, are given by:

$$\hat{S}_{mi} = \int_0^T z_i(t)\hat{\xi}_m(t)dt = \mathbf{a}_i^\top\mathbf{W}\hat{\mathbf{b}}_m, \qquad m \geq 1. \tag{3.12}$$

For functional data, the maximum number of non-zero eigenvalues M is defined as the minimum of the observed function values and the number of basis functions. Thus, for each choice of q $(1 \leq q \leq M)$, the q leading principal components define an orthonormal basis system that can be used to approximate the sample curves z_i in L^2 as follows:

$$z_i(t) \approx \sum_{m=1}^{q}\hat{S}_{mi}\hat{\xi}_m(t).$$

The eigenfunctions' coefficients $\hat{S}_{mi}$ define the optimal fit to each function z_i, where each coefficient is referred to as the principal component score of the i^{th} observation on the m^{th} principal component. These orthogonal basis functions provide the best basis of size q minimizing the following sum of squares:

$$\sum_{i=1}^{N}\int[z_i(t) - \sum_{m=1}^{q}\hat{S}_{mi}\hat{\xi}_m(t)]^2dt,$$

which is equivalent to the sum of the discarded eigenvalues. Accordingly, the appropriate number q of principal components to be used can be determined visually by plotting the eigenvalues γ_m against their indices m $(1 \leq m \leq M)$, which is known as a scree plot, then selecting the value of m at which the curve starts to flatten.

This orthonormal basis is an optimal choice for approximating the curves z_i, $i = 1, \ldots, N$, but it is not unique. There exists other equally good orthogonal set defined by:

$$\boldsymbol{\tau} = \mathbf{T}\boldsymbol{\xi},$$

where $\mathbf{T}$ is any orthogonal matrix of order q $(\mathbf{T}^\top = \mathbf{T}^{-1})$ that is referred to as rotation matrix. The new basis functions $\tau_1, \ldots, \tau_q$ are as effective as $\xi_1, \ldots, \xi_q$ at approximating the original curves in q dimensions. Although, there is an infinite number of rotation matrices, it is always of interest to find a rotation that yields an easier to interpret set of components. The VARIMAX rotation is one of the most common strategies that often provides a more interpretable basis system. In this context, let $\boldsymbol{\mathcal{B}}$ be a $q \times N$ matrix representation of the first q principal components, $\xi_1, \ldots, \xi_q$, such that the m^{th} row $(m = 1, \ldots, q)$ of $\boldsymbol{\mathcal{B}}$ contains the evaluated values of $\xi_m(t_1), \ldots, \xi_m(t_N)$ for N equally spaced time points in the interval $\mathcal{T}$. Thus, the matrix $\boldsymbol{\mathcal{A}}$ associated with the rotated basis functions is given by:

$$\boldsymbol{\mathcal{A}} = \mathbf{T}\boldsymbol{\mathcal{B}},$$

where $\mathbf{T}$ is chosen so that it maximizes the variance of α_{mj}^2 given by:

$$\sum_{m=1}^{q}\sum_{j=1}^{q} \alpha_{mj}^2 = \operatorname{tr}(\boldsymbol{\mathcal{A}}^\top \boldsymbol{\mathcal{A}}) = \operatorname{tr}(\mathbf{T}\boldsymbol{\mathcal{B}}^\top \mathbf{T}\boldsymbol{\mathcal{B}}) = \operatorname{tr}(\boldsymbol{\mathcal{B}}^\top \boldsymbol{\mathcal{B}}).$$

This implies that although the rotation preserves the collective amount of the variance, the individual rotated functions account for different percentages of variances.

3.3 Functional Clustering Data Analysis

The previous sections have justified the leading role of functional data analysis in analyzing high-dimensional data and the extension of basic exploratory data analysis methods and standard principal component analysis to functional data. Similarly, standard multivariate cluster analysis techniques are extended to functional data. Cluster analysis is a statistical tool used to identify mutually exclusive homogenous groups (clusters) of observations such that the within-group-object similarity is maximum and the between-group-object similarity is minimum. A major challenge in cluster analysis is to find the group structure and the optimal number of groups which are not known a priori.

A variety of clustering techniques are available in the literature. The most common methods of clustering include hierarchical clustering, K-means algorithm and probabilistic model-based approaches. Hierarchical clustering (Ward and Joe, 1963) involves building a hierarchy of clusters based on a dissimilarity measure between groups using either agglomerative or divisive procedures. Whilst, the K-means algorithm (Hartigan and Wong, 1978, MacQueen, 1967) is the most popular method among the geometric iterative procedures. It assigns objects to clusters, once the number of clusters is specified, such that each object is assigned to the cluster with the nearest mean and the within-cluster sum of squared is minimized. Alternatively, probabilistic model-based approaches (Banfield and Raftery, 1993) have been introduced to account for the disadvantages of both hierarchical and geometric clustering procedures, which are largely heuristic and not based on formal models; and hence formal inference is not possible. Model-based clustering assumes that the observations arise from a finite mixture of distributions, such that each cluster is characterized by a single density function in the mixture. A key advantage of this approach is that it allows the use of model comparison criteria such as AIC, AICc, BIC, etc to select the optimal numbers of clusters by comparing models with different number of clusters. Nevertheless, each observation is accompanied with a probability of cluster membership and hence formal inference can be carried on (Haggarty et al., 2012a). More details about cluster analysis can be found in Everitt et al. (2001).

With analogy to the classical multivariate clustering analysis, functions or curves can be classified into a set of mutually exclusive groups such that dissimilarity between curves within the same group is minimized. Theoretically, functional data are smooth curves belonging to an infinite-dimensional space and hence the main source of difficulty in modeling such data is the curse of dimensionality. But practically, these infinite-dimensional data are represented in a finite-dimensional space of functions; this is known as data reduction. Clustering finite-dimensional data using standard statistical methods can then be performed easily. This is done either through filtering methods or raw-data (regularization) methods. A detailed review of the functional data clustering methods and the filtering techniques is provided by Jacques and Preda (2014). In the filtering methods, each functional datum is firstly approximated using a finite-dimensional set of basis functions, then clustering is performed based on the basis coefficients of the generated smooth functions. While, in the raw-data methods, clustering is performed using the evaluation points of the curves. Raw-data clustering methods have several disadvantages, of which (i) the need of high-dimensional clustering techniques for the large number of evaluation points; (ii) ignoring the functional feature of the data such as continuity and smoothness; (iii) discounting of high correlation between the observations of the same curve; and (iv) the invariance of these methods to any permutation of the curves observation order.

Clustering techniques have been developed for functional data following the classical multivariate clustering approaches. Standard clustering methods including hierarchical, K-means and model-based clustering have been also considered from a functional point of view. Most of the functional clustering algorithms are based on the filtering methods which start with representing the curves into a finite basis of functions. In the majority of applications, the most common choice of basis functions is the spline basis because of their optimal properties mentioned in Section 3.1.1. Trapey and Kinateder (2003) discussed the clustering of functional data using a K-means algorithm, exploiting the connections between the coefficients of the splines coefficients and the values that their continuous functions attain. Whereas, Abraham et al. (2003) proposed a K-means algorithm to cluster functional data based on their estimated B-splines coefficients. A similar functional clustering based on the Partitioning Around Medoids (PAM) algorithm is proposed by Ignaccolo et al. (2008) to classify the air monitoring stations in Piemonte in Italy. Alternatively, Henderson (2006) applied an agglomerative hierarchical clustering to the B-splines coefficients of water quality trend curves for a number of monitoring sites in Australia. Giraldo et al. (2012) also used a hierarchical clustering approach to classify spatially correlated temperature curves based on the coefficients of their Fourier basis functions. To account for the spatial correlation between the curves, they weighted the dissimilarity matrix by the trace-variogram and the multivariate variogram calculated with the basis functions coefficients. A similar hierarchical functional clustering technique is used to identify groups of monitoring stations on the River Tweed that display similar spatio-temporal characteristics in terms of the levels of nitrate pollution (Haggarty et al., 2015). In this paper, the authors accounted for the spatial covariance between functions from sites along a river network using a stream distance metric. Model-based clustering techniques have also been extended to classify functional data. James and Sugar (2003) proposed the first model based clustering method for sparse and irregular functional data, where each curve is represented in terms of its spline basis functions and the corresponding coefficients are distributed according to a mixture Gaussian distributions with parameters specific to each cluster.

Another technique for filtering and reducing the data dimensionality is based on the use of functional principal component analysis, introduced in Section 3.2. Accordingly, many attempts were devoted to cluster functional curves based on their FPC scores. For example, Peng and Muller (2008) used a distance based clustering method to cluster functional data based on the principal components scores. Adelfio et al. (2010) used the trimmed k-means algorithm proposed by Garci-Escudero and Gordaliza (2005) to find clusters of earthquakes according to the FPCA directions of their waveforms. Recently, Lin et al. (2015) developed a 2-dimensional functional principal component analysis to fully capture the space variation in the ovarian and kidney cancer histology image signals; then for accurate feature selection they used a randomized k-means cluster analysis of the principal components. Model-based clustering based on the functional principal component scores have also been considered in Bouveyron and Jacques (2011) and Jacques and Preda (2014).

Among the functional clustering techniques mentioned above, only the hierarchical clustering and the K-means algorithm are described in more details below, in Sections 3.3.1 and 3.3.2, and used to cluster the daily curves of EpCO$_2$. The probabilistic model-based approaches for clustering functional data are out of the scope of this thesis, as they are computationally inefficient especially with large volumes of data and are only considered of added value if further statistical inference is to be performed. All the multivariate clustering methods, as mentioned above, are used to classify the functional data based on their corresponding spline basis coefficients or their functional principal component scores. Recent developments involve the extension of the K-means algorithm to cluster functional data in the function domain. This method is quite useful when it it is not possible to control the discretization of the functions and when the discretization leads to very high-dimensional problems due to the functional nature of the series. For more details, see Garcia et al. (2015).

3.3.1 Hierarchical Clustering

Hierarchical clustering (Ward and Joe, 1963) starts with defining a distance (dissimilarity) matrix between individual observations, then finding clusters of individuals using either an agglomerative or divisive criterion. The agglomerative techniques are the most commonly used, they start with each individual observation belonging to a cluster, then clusters are merged based on the predefined distance matrix to form larger clusters. Whereas, the divisive techniques start with all observations in one cluster, then observations are separated to form finer clusters according to the dissimilarity measure. The results can be represented by a 2-dimensional diagram called a “dendrogram”, which illustrates the arrangement of clusters.

The dissimilarity between a pair of observations depends on the distance between them, quantified by a distance metric. The most commonly used distance metrics are the Euclidean distance, the squared Euclidean distance and the standardized Euclidean distance. Based on the chosen distance metric, a dissimilarity matrix D displaying the distance between each pair of observations can be obtained, where each cell d_{ij} is the distance between between the i^{th} and the j^{th} observation. Afterwards, a linkage criterium such as single, complete or average linkage is employed to determine which clusters should be joined at each stage. In single linkage, the distance between two clusters is defined as the distance between the two closest members. In complete linkage, the distance between two clusters is calculated as the distance between the two furthest members. In average linkage, the distance between two clusters is the distance between the clusters means.

Hierarchical clustering has been extended to functional data by defining a dissimilarity metric that measures the distance between curves (Hitchcock et al., 2006). The distance d_{ij} between two curves, $x_i(t)$ and $x_j(t)$, defined through the L^2 norm is given by (Hitchcock et al., 2006):

$$d_{ij} = \sqrt{\int_{\mathcal{T}} (x_i(t) - x_j(t))^\top (x_i(t) - x_j(t)) dt}.$$

Using the splines expansion in Equation 3.2, the L^2-norm distance d_{ij} can be expressed as follows:

$$d_{ij} = \sqrt{(\mathbf{a}_i - \mathbf{a}_j)^\top \mathbf{W} (\mathbf{a}_i - \mathbf{a}_j)},$$

where $\mathbf{W} = \int_{\mathcal{T}} \boldsymbol{\psi}(t)\boldsymbol{\psi}(t)^\top dt$ such that $\mathbf{a}_i$ and $\mathbf{a}_j$ are the basis coefficients vectors of the smooth curves $x_i(t)$ and $x_j(t)$. For basis functions such as B-splines, the matrix $\mathbf{W}$ is a symmetric matrix of order p obtained by numerical integration (Henderson, 2006). Then, based on the defined dissimilarity matrix, standard hierarchical clustering procedures can be applied (Giraldo et al., 2012).

Although hierarchical clustering has proven to be a good tool for classifying functional data if an appropriate distance metric is used, it fails to identify a reasonable clustering structure for a large data set. Therefore, we will mainly focus in this thesis on the use of the K-means algorithm to classify the daily curves of EpCO$_2$.

3.3.2 K-means Algorithm

The K-means algorithm (Hartigan and Wong, 1978, MacQueen, 1967) aims to cluster a set of individual observations into C clusters, such that each observation belongs to the cluster with nearest mean. This can be achieved by minimizing the within-cluster sum of squares. In FDA, let $m_c(t)$ be the functional mean/center of class c summarized by its splines coefficients $\mathbf{a}^c = (a_1^c, \dots, a_p^c)^\top$. The ultimate goal of the K-means algorithm is then to find the class centers $A = \{\mu_1(t), \dots, \mu_C(t)\}$ minimizing the sum of squared differences between each functional observation $x_i(t)$ in a certain cluster c, $c = 1, \dots, C$, and the functional center $\mu_c(t)$ of this cluster c, represented by their splines coefficients vectors $\mathbf{a}_i$ and $\mathbf{a}^c$, respectively given by (Abraham et al., 2003):

$$\sum_{\mathbf{a}_i \in c} \| \mathbf{a}_i - \mathbf{a}^c \|^2 .$$

This within-cluster sum of squares is minimized over all clusters C; and hence it decreases as the number of clusters increases. Therefore, the number of clusters should be defined a priori. The first step often involves arbitrarily choosing the cluster centers for a predefined number of clusters. The second step of the algorithm is to classify $\mathbf{a}_i$ and hence $x_i(t)$ to the cluster with the closest center. In the third step, the results of the second step are used to recompute the cluster centers $\mu_c(t)$ as the mean of the splines coefficients of the curves assigned to cluster c. The K-means clustering technique is an iterative algorithm where the last two steps are repeated until the cluster centers remain stable. Unfortunately, there is no guarantee that the within-cluster sum of squares will reach a global minimum (Abraham et al., 2003). However, the final results can be checked by repeating the algorithm many times with different initial clusters centers (Krzanowski and Lai, 1988).

3.3.3 Selecting the Optimal Number of Clusters

The main challenge in clustering analysis is the estimation of the optimal number of clusters, which is usually not predefined, for a given data set. Various techniques have been suggested to determine the optimal number of clusters. A review of the different methods used to estimate the optimal number of clusters is provided by Milligan and Cooper (1986) and Gordon (1996). The best performer among these methods is the index proposed by Calinski and Harabasz (1974) calculated based on the ratio of the between and within cluster sum of squares. Alternatively, Kaufman and Rousseeuw (1990) suggested to estimate the optimal number of clusters using the silhouette statistic, which measures how close each object in one cluster is to all other points in the neighbouring clusters. This latter method calculates for each observation the difference between the average distance to the other observations in its cluster and the average distance to the observations in the nearest cluster and estimates the optimal number of clusters as the number maximizing the average difference between both distances over the whole data set.

Another simple technique for selecting the optimal number of clusters is to plot a sequence of the number of clusters C versus their corresponding within-cluster sum of squares. Typically, the within-cluster sum of squares decreases monotonically as the number of clusters increases; but this decrease becomes minimal after a certain value of C, which defines the appropriate number of clusters. This heuristic procedure is known as the "elbow plot" or the "L-curve". Although this procedure is very simple and straightforward, it suffers from some drawbacks. The major shortcoming is that it is not a formal statistical procedure, since there is no reference distribution to compare with it the curve of the observed within-cluster sum of squares versus the number of clusters C. Another disadvantage is that for different values of C, the within-cluster sum of squares are not normalized which makes the comparison unreliable. To overcome all the deficiencies of the L-curve, Tibshirani et al. (2001) has proposed a more formal procedure known as the "gap statistic".

The gap statistic, proposed by Tibshirani et al. (2001) is a very popular formal statistical method used to determine the optimal number of clusters to be used for any clustering method. The gap statistic compares the within-cluster sum of squares with the expected within-cluster sum of squares under a reference null distribution of no clustering. For data x_{ij}, $i = 1, \ldots, N$, $j = 1, \ldots, p$, where N and p are the number of independent observations and features measured for each observation, respectively; let $d_{ii'}$ denotes the distance between the observations i and i' calculated based on a chosen distance metric. The Euclidean distance is the most common metric for calculating the distances $d_{ii'}$. Now, suppose that the data x_{ij} can be classified into C clusters $\boldsymbol{\mathcal{C}} = \{\mathcal{C}_1, \ldots, \mathcal{C}_C\}$, where N_c is the number of observations in the cluster $\mathcal{C}_c$. Then, the sum of squares within a certain cluster $\mathcal{C}_c$ is defined as the sum of squared pairwise distances between all the points in this cluster as follows:

$$D_c = \sum_{i,i' \in \mathcal{C}_c} d^2_{i,i'},$$

and hence the within-cluster homogeneity is measured by the within-cluster sum of squares defined as:

$$W_C = \sum_{c=1}^{C} \frac{1}{2N_c} D_c.$$

The fundamental idea of the gap statistic is to compare the observed within-cluster sum of squares W_C with its expectation under an appropriate null reference distribution. Following from this, a reference distribution assuming no clustering structure in the data is needed (Tibshirani et al., 2001). The reference distribution can be generated from a uniform distribution where each reference feature is generated from a uniform distribution over the range of the observed values of that feature. Another choice for the reference distribution is to generate the reference features from a uniform distribution of the principal components of the data. The latter approach takes into account the shape of the data distribution and makes the procedure rationally invariant, see Tibshirani et al. (2001) for more details.

After generating the reference distribution using one of the above two approaches, B reference data sets are drawn from this reference distribution using Monte Carlo simulation; and the same clustering procedure applied to the observed data is applied for each reference data set. For each reference data and each number of clusters, the within-cluster sum of squares W^*_{Cb}, where C is the number of clusters and $b = 1, \ldots, B$ is the index of the reference data set, is calculated and the corresponding gap statistic is computed as follows:

$$\text{Gap}(C) = \frac{1}{B} \sum_{b=1}^{B} \Big[\log(W^*_{Cb}) - \log(W_C) \Big]. \tag{3.13}$$

According to Equation 3.13, the gap statistic is defined as the difference between the average within-cluster sum of squares for B reference data sets and the observed within-cluster sum of squares. The logs of the within-cluster sum of squares are considered to normalize the within-cluster homogeneity being compared. Based on the computed gap statistics, the estimate of the optimal number of clusters is the value of C that corresponds to the largest gap or in other words, the value of C at which the difference between the within-cluster dispersions of the observed data, known to have a clustering structure, and that of the reference data, assuming no clustering structure, is maximum. A tolerance of one standard error is used to account for the simulation error; and hence the estimated optimal number of clusters $\hat{C}$ is chosen as:

$$\hat{C} = \text{smallest } C \text{ such that Gap } (C) \geq \text{Gap}(C+1) - sd_{C+1}\sqrt{1 + 1/B},$$

where for a certain number of clusters C, $sd_C\sqrt{1 + 1/B}$ defines the one standard error tolerance and sd_C denotes the standard deviation of the within-cluster homogeneity of the B reference data sets given by:

$$sd_C = \sqrt{\frac{1}{B} \sum_{b=1}^{B} \Big\{ \log(W^*_{Cb}) - \frac{1}{B} \sum_{b=1}^{B} log(W^*_{Cb}) \Big\}}.$$

3.3.4 Comparing Clustering Results

Another point of interest is to compare the results of two clustering methods or measure the similarity between two classifications with different number of clusters. Milligan and Cooper (1985) has reviewed many indices used to measure the agreement between two partitionings. The Rand Index (RI), proposed by Rand (1971), is a very popular measure of agreement between two partitionings. To compute RI, suppose that $\boldsymbol{\mathcal{U}} = \{\mathcal{U}_1, \ldots, \mathcal{U}_R\}$ and $\boldsymbol{\mathcal{C}} = \{\mathcal{C}_1, \ldots, \mathcal{C}_C\}$ represent two partitions of the N observations in a data such that

$\bigcup_{r=1}^{R} \mathcal{U}_r = \bigcup_{c=1}^{C} \mathcal{C}_c$. Also, let d be the number of pairs of observations in the same class of $\boldsymbol{\mathcal{U}}$ and in the same class of $\boldsymbol{\mathcal{C}}$, e be the number of pairs of observations in the same class of $\boldsymbol{\mathcal{U}}$ but not in the same class of $\boldsymbol{\mathcal{C}}$, f be the number of pairs of observations not in the same class of $\boldsymbol{\mathcal{U}}$ but in the same class of $\boldsymbol{\mathcal{C}}$, and g be the number of pairs of observations in different classes of $\boldsymbol{\mathcal{U}}$ and different classes of $\boldsymbol{\mathcal{C}}$; then RI is computed by:

$$\text{RI} = \frac{d+g}{d+e+f+g}.$$

RI lies between 0 and 1; the two partitionings are identical when it is equal to 1. This index suffers from a number of shortcomings. Among these disadvantages is that it is highly dependent on the number of clusters. Another drawback is that the expected value of RI of two random partitions does not take a constant value (e.g. zero). To overcome these deficiencies, Hubert and Arabie (1985) proposed the Adjusted Rand Index (ARI) which assumes a generalized hypergeometric distribution as the model of randomness under the null hypothesis "H_o : the two clusterings $\boldsymbol{\mathcal{U}}$ and $\boldsymbol{\mathcal{C}}$ are drawn at random such that the number of clusters and the number of elements in each cluster are fixed". Let N_{rc} be the number of observations in the class $\mathcal{U}_r$ and the class $\mathcal{C}_c$ and let $N_{r.}$ and $N_{.c}$ be the number of observations in the class $\mathcal{U}_r$ and the class $\mathcal{C}_c$, respectively. Then, the ARI is the normalized difference between the RI and its expected value under the above null hypothesis defined as follows:

$$\text{ARI} = \frac{\sum_{rc}\binom{N_{rc}}{2} - \mathbb{E}\Big[\sum_{rc}\binom{N_{rc}}{2}\Big]}{\frac{1}{2}\Big[\sum_{r}\binom{N_{r.}}{2} + \sum_{c}\binom{N_{.c}}{2}\Big] - \mathbb{E}\Big[\sum_{rc}\binom{N_{rc}}{2}\Big]}.$$

The expected value of the ARI ranges between 0 for independent clustering and 1 for identical clustering.

In brief, functional data are infinitely dimensional curves, which makes clustering functional data a difficult task. A very popular approach, known as filtering, involves reducing the dimension of the data by approximating the curves into a finite basis of functions. Another data reduction technique is the functional principal component analysis, where the infinite-dimensional curves can be filtered into a finite and relatively smaller number of functional principal components scores (Jacques and Preda, 2014). After summarizing the functional data either by their basis coefficients or by their first principal component scores, standard clustering algorithms can be used to define the clusters. The key advantages for the use of functional principal components scores are: (i) the identification of the primary sources of variations in the daily patterns of EpCO$_2$ after adjusting for the average daily pattern and; (ii) the orthogonality and hence the independence between the functional principal components scores of the same smooth curve.

3.4 Application of FDA to the EpCO$_2$ data

This section aims at classifying the daily patterns of EpCO$_2$, across the whole study period (October 2003 - September 2006), into a set of clusters based on both mean level and shape. This should help in identifying the different hydrological and climatological conditions driving each of those daily patterns. As previously mentioned, the data cover three full hydrological years with 1095 days in total. In principle, the data are recorded every 15 minutes allowing for 96 observations to be available per day. Using a functional data analysis approach, the 96 (15-minute) observations within each day are considered as the discrete observations of a continuous smooth function. This view of the data allows the daily EpCO$_2$ patterns to be estimated using smooth curves without being concerned about the high-correlations between the 15-minute observations within the same day. The rest of this section describes the procedure of transforming the 15-min resolution discrete data to the 1-day resolution smooth continuous functions, and the clustering of these daily EpCO$_2$ functions using the above methodology.

3.4.1 Discrete Data to Functional Data

The initial challenging step in FDA is to estimate a smooth continuous function of the observed data for each day. The observations within each day are regularly spaced and have the same start time (00:00) and end time (23:45), which ensures a fair comparison of the EpCO$_2$ daily patterns. Generally, the EpCO$_2$ data are considered complete and regularly spaced across the whole study period, except some short periods of missings and irregularity. The hydrological year 2003/2004 witnessed 17 complete days of missings in June and one missing observation in February, while the hydrological year 2004/2005 encountered a period of inconsistently spaced observations with roughly one observation every 15 minutes. Maintaining the same level of smoothing across all days is essential to guarantee a fair comparison of the different daily curves. Therefore, the same smoothing parameter is used for each day and also the knots are placed such that the interval between each two knots contains the same quantity of data, otherwise some basis coefficients will be estimated more accurately than others.

The 17 complete days of missings in June 2004 have been excluded from the data, as the temporal order of days is not yet taken into account. Next, interpolating natural cubic splines have been fitted to the days with either a small number of missings or

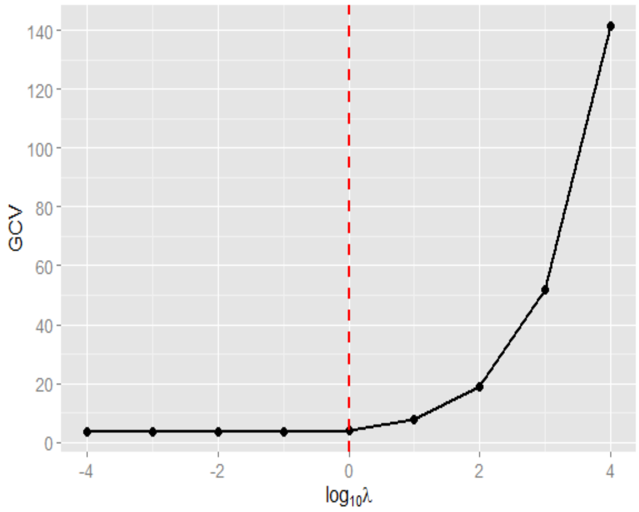


Figure 3.1: Plot of the GCV criterion against the corresponding smoothing parameter λ (in $\log_{10}$ scale) used to fit the daily $EpCO_2$ smooth curves using penalized cubic B-splines.

irregularly spaced observations. The estimated splines are then evaluated at the set of regularly spaced (15-minute) time points between 00:00 and 23:45 and the fitted values are used to replace the missing data at these time points. After that, the `fda` package in `R` was used to fit a smooth continuous curve to the observations within each day using penalized cubic B-splines. The order of cubic B-splines is 4 and the knots are placed at two hours intervals, suggesting that there are 11 interior knots over each day. Two hours was considered an appropriate interval since previous analysis (see Chapter 2) indicated that two hours is the extent of short-term dependence where the system does not seem to witness large variability. According to the rule specifying the number of basis functions as the sum of the number of interior knots and the order of basis functions, a total of 15 basis functions are used to fit the daily smooth functions. This number of basis functions is reasonable to map all the key features in the data and provides a good fit for the discrete data (see Figure 3.2).

As mentioned earlier in Chapter 2, in addition to the choice of the number of basis functions, the smoothing parameter λ is one of the key choices to be made to control the smoothness of a curve. Initially, GCV was used to select an appropriate smoothing parameter for the daily $EpCO_2$ curves. Figure 3.1 indicates that the GCV starts to stabilize at a smoothing parameter equal to 1 ($\log_{10}\lambda = 0$) and that using larger smoothing parameters tends to yield larger GCV. However, the GCV results here should be interpreted with caution as they tend to be unstable with correlated data. Therefore, a sensitivity analysis approach was also carried out to determine the most appropriate smoothing parameter by visually examining the effect of changing the smoothing parameter on the estimated smooth curves. A random sample of 12 days are used to illustrate

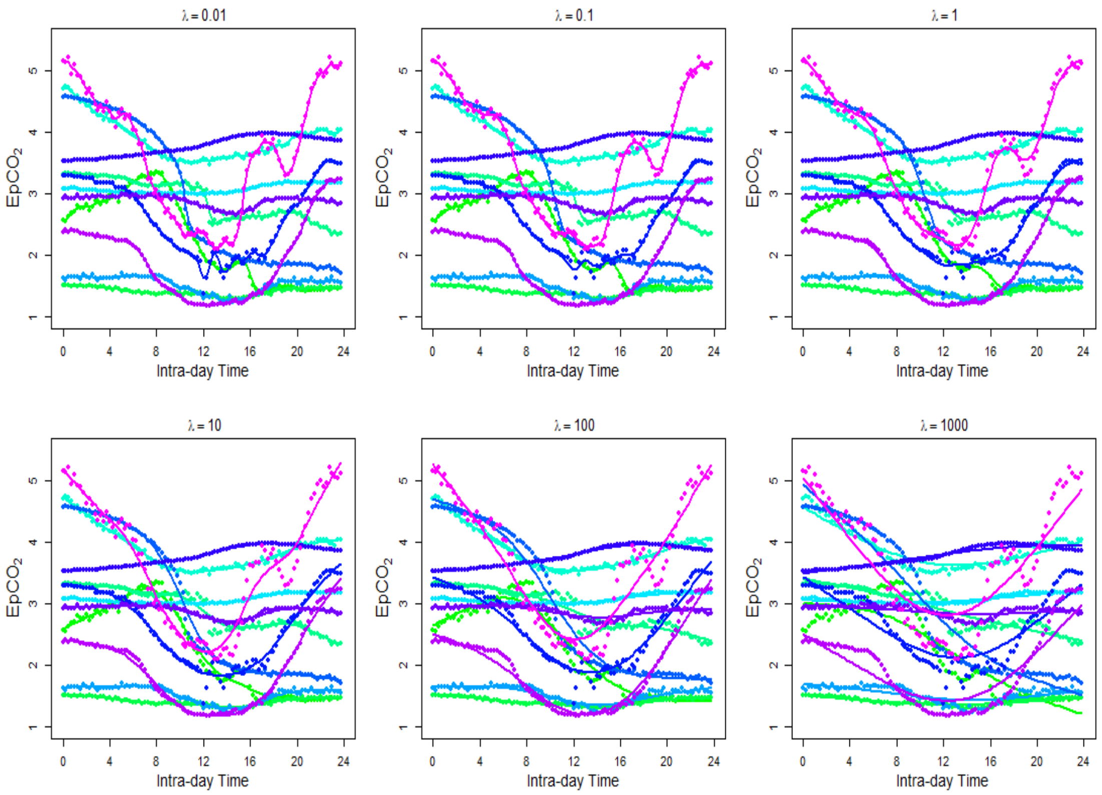


FIGURE 3.2: Plot of the observed (discrete) $EpCO_2$ data (points) for a random sample of 12 days and their corresponding smooth curves fitted with penalized cubic B-splines using different smoothing parameters λ.

the effect of using different smoothing parameters in Figure 3.2. The figure shows the smooth curves fitted using 15 cubic B-splines basis functions, for the randomly selected 12 days, combined with 6 different smoothing parameters $(10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3})$. The two smallest smoothing parameters 10^{-2} and 10^{-1} have similar effects on the level of smoothing; both are under-smoothing the data and tend to capture very local and fine variations that are not of interest. In contrast, the two largest smoothing parameters 10^{2} and 10^{3} over-smooth the data and miss some of the details and variability highlighting the daily patterns. Whereas, the smoothing parameters 1 and 10 provide a good compromise between capturing the main daily pattern of $EpCO_2$ and the particular features of each individual day. Using a smoothing parameter equals to 10 seems to miss some of the curvature in the data of some days which might be of interest to study. Therefore, it has been decided to smooth the daily $EpCO_2$ curves using 15 cubic B-spline basis functions combined with a roughness penalty term with a smoothing parameter equals to 1, which appears to be the best choice according to the results of both GCV and sensitivity analysis.

Figure 3.3 displays the mean daily curve of $EpCO_2$ and the ± 2 functional standard errors. The mean curve shows the average daily pattern of $EpCO_2$ that tends to be

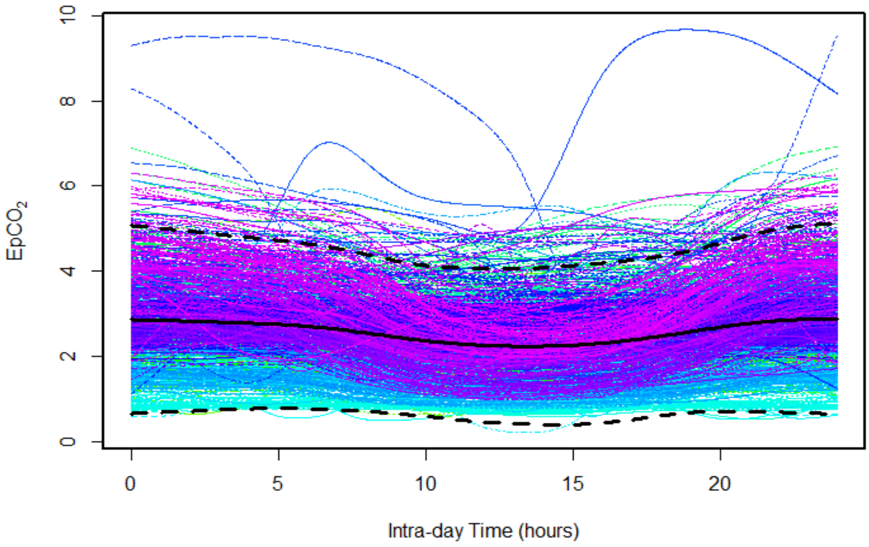
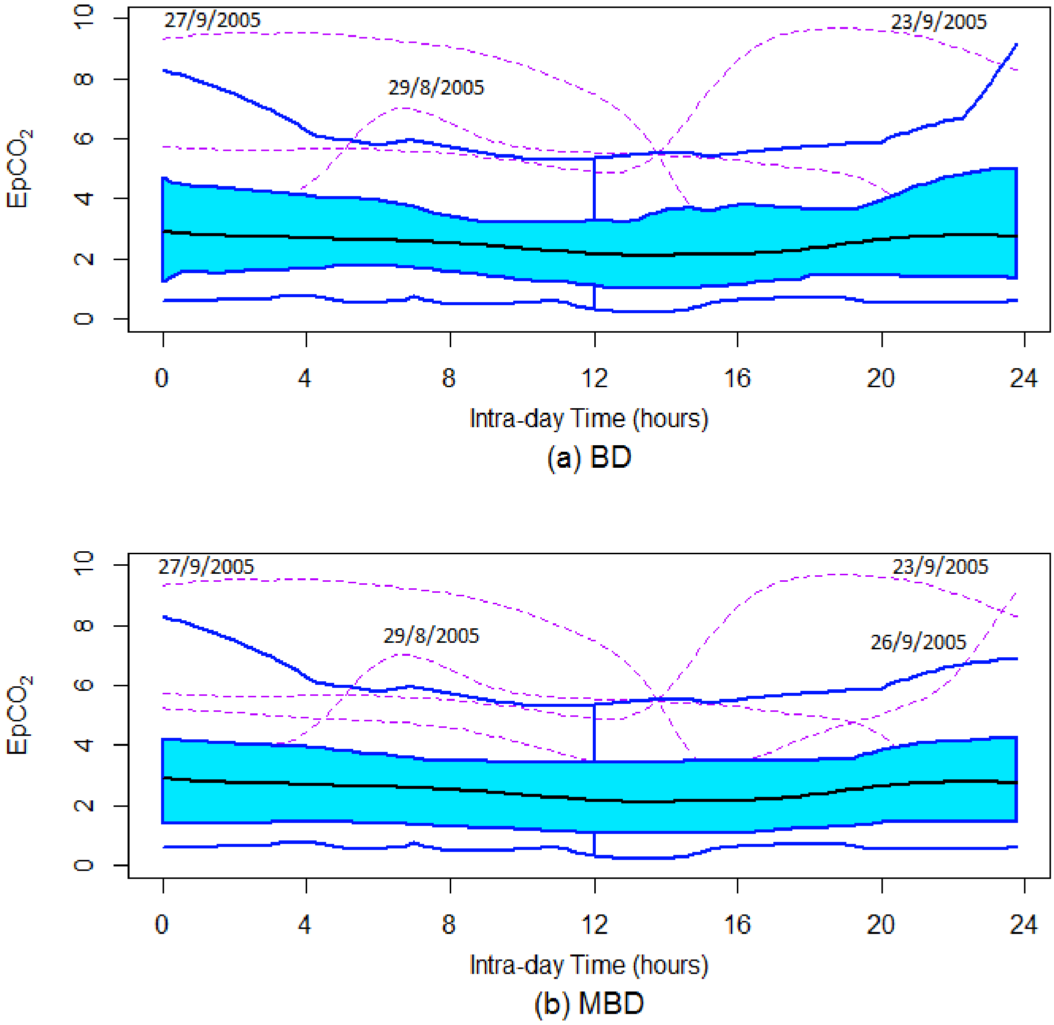


FIGURE 3.3: Plot of the daily $EpCO_2$ smoothed functions. The solid black curve is the mean function and the dashed black curves are the ± 2 standard errors functions.

lower during day-time reflecting the biological activity of consuming CO_2 during the light hours. Though, there is a quite large amount of variability between the daily curves of $EpCO_2$ in terms of both shape and level. It is then of interest to ecologists to study the different daily patterns to understand the biological and hydrological drivers of each of those patterns.

Functional box-plots have been considered to detect functional outliers in both shape and magnitude. The functional box-plots in Figure 3.4 constructed using the band depth (panel-a) and its modified version (panel-b) detected 29/8/2005, 23/9/2005 and 27/9/2005 as shape and magnitude functional outliers. The curve corresponding to 26/9/2005 was also flagged as a magnitude outlier according to its modified band depth value. As can be seen, the functional box-plots indicated that $EpCO_2$ was abnormal from the end of day 26/9/2005 up to 3:00 pm on 27/9/2005. It is unsurprising to get outliers in adjacent days since the daily curves are time dependent. The temporal correlation between the $EpCO_2$ daily curves affects the variance of the band depth (or its modified version) estimator but not its unbiasedness. In particular, accounting for the correlation between the curves would lead to larger variability (Sun and Genton, 2011a), and hence the number of potential outliers might decrease. However, 4 functional outliers represent only 0.4% of the total number of curves, and hence excluding these 4 curves before any further analysis should not affect the final results.



(a) BD

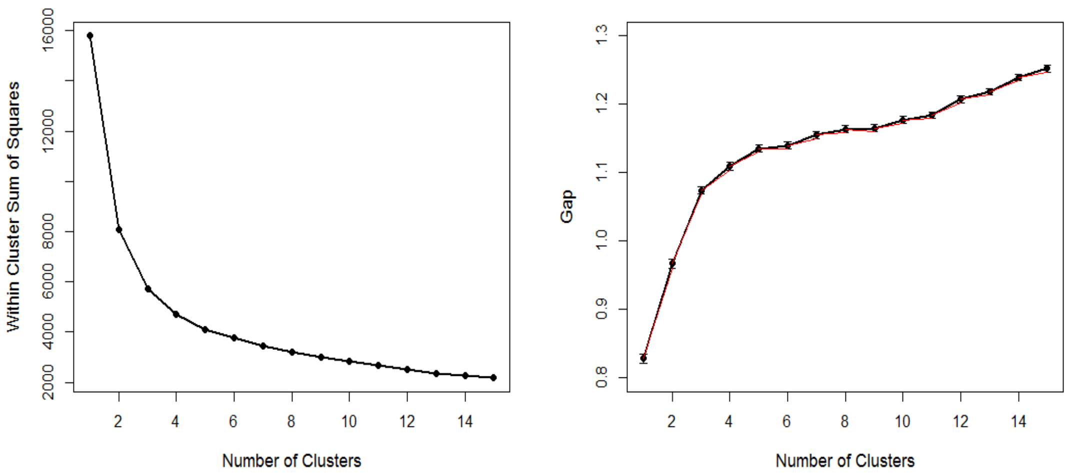
(b) MBD

FIGURE 3.4: Functional box-plots of the daily $EpCO_2$ curves constructed using (a) the band depth (BD) and (b) the modified band depth (MBD) measures.

3.4.2 Functional Clustering Using Basis Expansion Coefficients

After transforming the observed discrete data into continuous daily functions and removing the potential functional outliers, a functional clustering analysis approach is applied to visualize the similarities and the differences between the daily curves of $EpCO_2$ and to find the characteristics of each group of curves. As discussed earlier, one approach of functional clustering takes advantage of approximating the curves with a finite basis of functions and reducing the dimensionality of the data, then clusters the curves based on their basis coefficients using the standard tools of clustering such as K-means or hierarchical clustering.

The K-means clustering algorithm, presented in Section 3.3.2, was first applied to the B-splines coefficients of the daily functions of $EpCO_2$. The gap statistic and L-curve

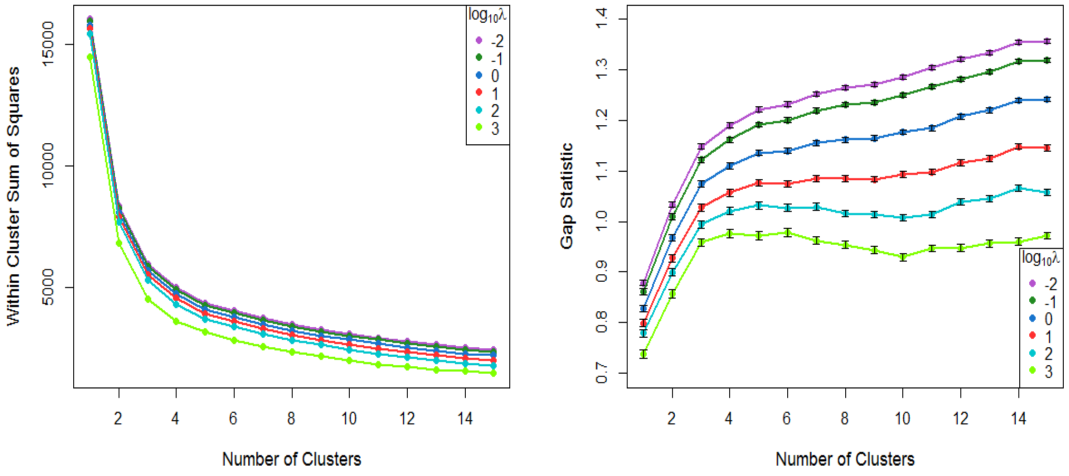


(a) (b)

Figure 3.5: (a) L-curve and (b) gap statistic plot for the K-means clustering of daily $EpCO_2$ curves approximated with their coefficients of cubic B-splines.

have been initially used to investigate the optimal number of clusters. For each potential number of clusters ranging from 1 to 15, 500 sets of reference data were generated using the principal components approach. There is no common rule for the number of reference data sets to be used; however using 500 data sets tended to produce consistent results when repeating the algorithm many times. A number of clusters ranging from 1 to 15 was considered to permit exploring a wide range of possible number of clusters. For each reference data set and each possible number of clusters, the K-means clustering algorithm was applied and the within-cluster sum of squares were calculated. Based on the sum of squares, the mean and standard deviation of the 500 reference data sets were computed. Then, the expected reference distribution and the corresponding gap statistic were calculated for each number of clusters ranging between 1 and 15. Figure 3.5(a) displays the elbow plot of the within-cluster sum of squares versus the potential number of clusters, also known as the L-curve. The optimal number of clusters, according to the L-curve, is the number at which the curve evidently flattens out. The L-curve, in Figure 3.5(a), seems to be inconclusive regarding the number of clusters. Instead, Figure 3.5(b) shows the plot of the gap statistic against the number of clusters, where the bars represent one standard error interval. According to the gap statistic, the optimal number of clusters is the number at which the gap is greater than the subsequent gap minus one standard error. That is, the optimal number of clusters is the first number at which the red line has a negative slope. For many runs, the gap statistic has consistently identified 5 as an appropriate number of clusters for the different daily patterns of $EpCO_2$ based on their basis coefficients.

To ensure that the optimal number of clusters identified using the gap statistic is not



(a) (b)

Figure 3.6: (a) L-curves and (b) gap statistic plots for the K-means clustering of daily $EpCO_2$ curves approximated with their basis expansion coefficients estimated using different smoothing parameters λ.

sensitive to the smoothing parameter used to smooth the daily curves, the gap statistic has been employed to select the optimal number of clusters for the curves fitted using different smoothing parameters. For different runs, the gap statistic has identified a very large number, 10 to 15, clusters for the curves estimated using a smoothing parameter less than 1, 4 to 5 optimal clusters for those fitted using a smoothing parameter larger than 1 and consistently 5 optimal clusters for a smoothing parameter equals to 1. Not surprisingly, a larger number of clusters is definitely needed to describe the different daily patterns if the estimated curves are capturing the very local and fine variations; and a smaller number of clusters would be optimal for over-smoothed curves only capturing the main daily pattern. Figure 3.6 displays the gap statistic curves as well as the L-curves corresponding to the clustering of curves fitted using different smoothing parameters. It is noticed from both plots that the differences between successive total within sum of squares become generally smaller after 5 clusters. Thus, 5 clusters seem to be a reasonable choice for the number of clusters describing the different daily patterns regardless of the chosen value of the smoothing parameter. It is also clear that the shape of the L-curves is identical for the different smoothing parameters and that there is no big differences between the curves in terms of the within sum of squares. This emphasizes that the grouping structure underlying the daily curves of $EpCO_2$ is not sensitive to the choice of the smoothing parameter and hence the fine and very local variability within the day captured or not captured by the different levels of smoothing. This is mainly because the clustering was carried out on using the K-means procedure, which classifies the data based on the mean level rather than any local features.

The following step involves applying a K-means clustering algorithm with 5 centers to the 15 splines coefficients defining the daily curves. The algorithm was repeated for 50 different random starts of centers to ensure the consistency of the resulting clusters. The K-means clustering results indicated that the grouping structure relies mostly on the estimated EpCO_2 mean level which differs from one class to another. This is unsurprising because the clusters are formed using the K-means algorithm which basically depends on the mean level (Figure 3.7). Another element of distinction between the 5 groups is the shape of the daily pattern of EpCO_2. For instance, some of the groups have a clear intra-daily cycle with a considerable drop in EpCO_2 level during day time while others have a fairly stable EpCO_2 level over the day.

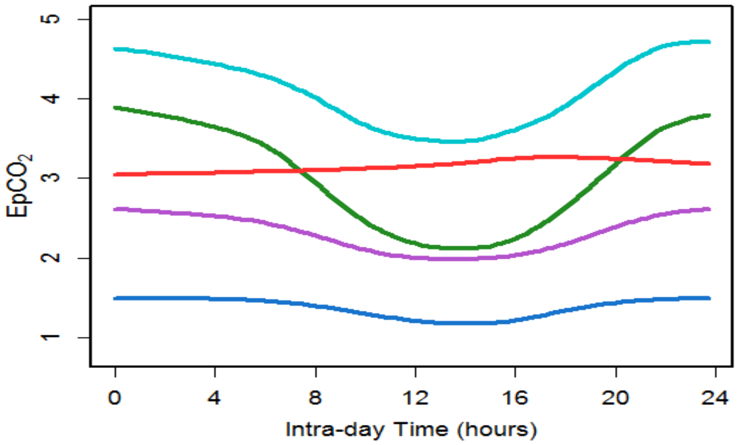


FIGURE 3.7: Mean EpCO_2 smoothed curves of the identified clusters.

Figure 3.8 shows the 5 classes of daily EpCO_2 curves. In each class, the majority of curves falls within ± 2 standard errors of the class mean curve and only a small proportion of the curves seems to exhibit peculiar patterns. Figures 3.8 and 3.9 indicate that the days in the green class have a relatively high level of EpCO_2 and a highly pronounced diel cycle and are mostly common in the dry periods of summer. Whereas, the turquoise curves, those characterized by a higher level of EpCO_2 and a moderately pronounced diel cycle, occur in summer especially in periods of a large drop in conductivity. This suggests that hydrological instability caused by flow events or other hydrological events tend to dilute or attenuate the diurnal cycle of EpCO_2. Whilst, the class of purple curves characterized by an average EpCO_2 level and a shallow trough during daytime underlies the days witnessing or following sudden drops in conductivity, concurrent with extremely high flow events periods. Yet, the red class, consisting of the flatter curves with a constant average level of EpCO_2 over the day, mostly corresponds to autumn and winter periods of heavy rainfall. In contrast, the blue daily curves with relatively low levels of $EpCO_2$ over the day characterize the hydrologically stable periods of winter and spring, coinciding with dry periods.

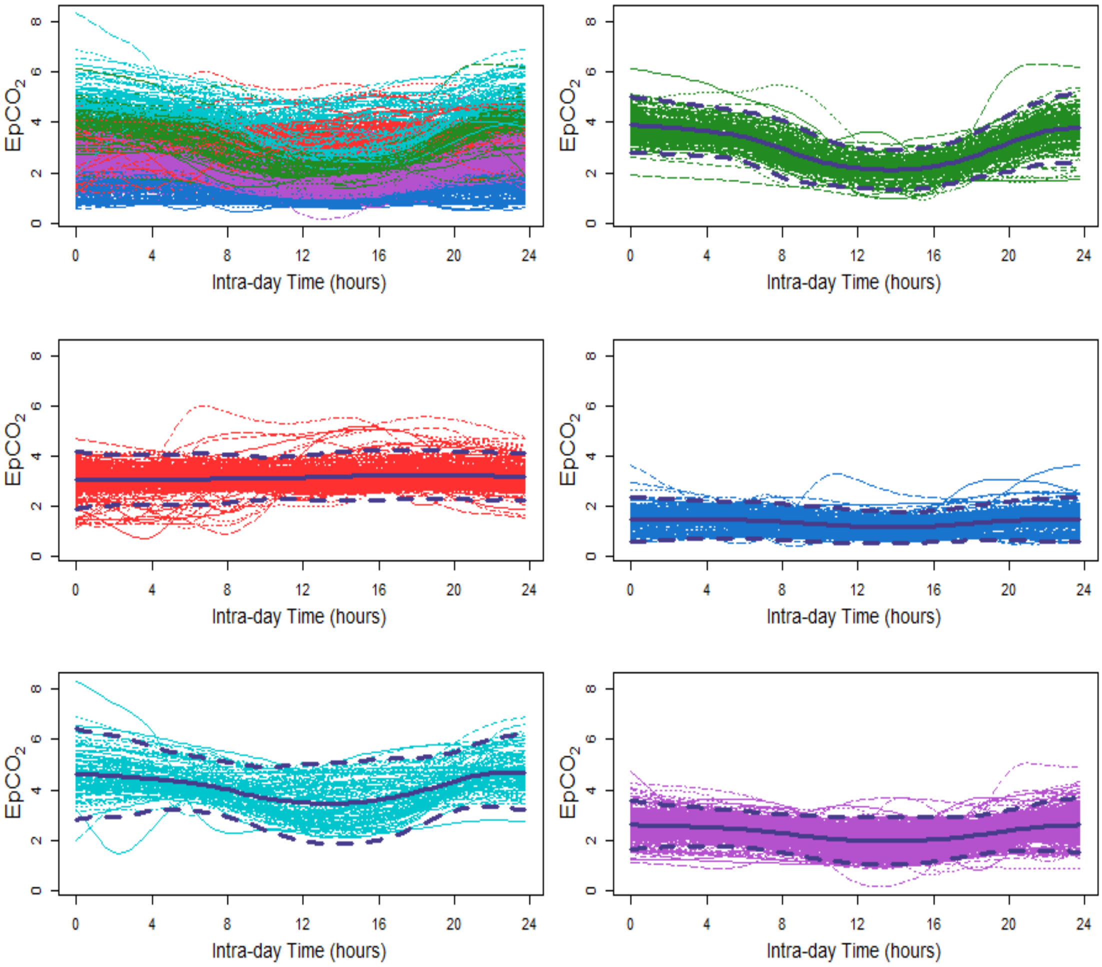


FIGURE 3.8: Daily $EpCO_2$ smoothed curves grouped using K-means of the curves' splines coefficients. The solid and dashed purple curves represent the functional mean and ± 2× standard deviation bands, respectively.

3.4.3 Functional Clustering Using FPCs

Another approach to cluster the daily curves of $EpCO_2$ is to use functional principal component analysis to first decompose the variability in the curves, then cluster the curves based on the scores of the first few FPCs. A key advantage of this approach is first reducing the data dimensionality through summarizing each curve by its first FPC scores. The number of principal components is selected according to the percentage of variance explained, which is the typical criterion in PCA. Clustering based on the FPC scores yields a grouping structure based on the main sources of variability in the daily curves. For each curve, the scores of the different FPCs are independent and hence classical clustering approaches including K-means can be easily applied.

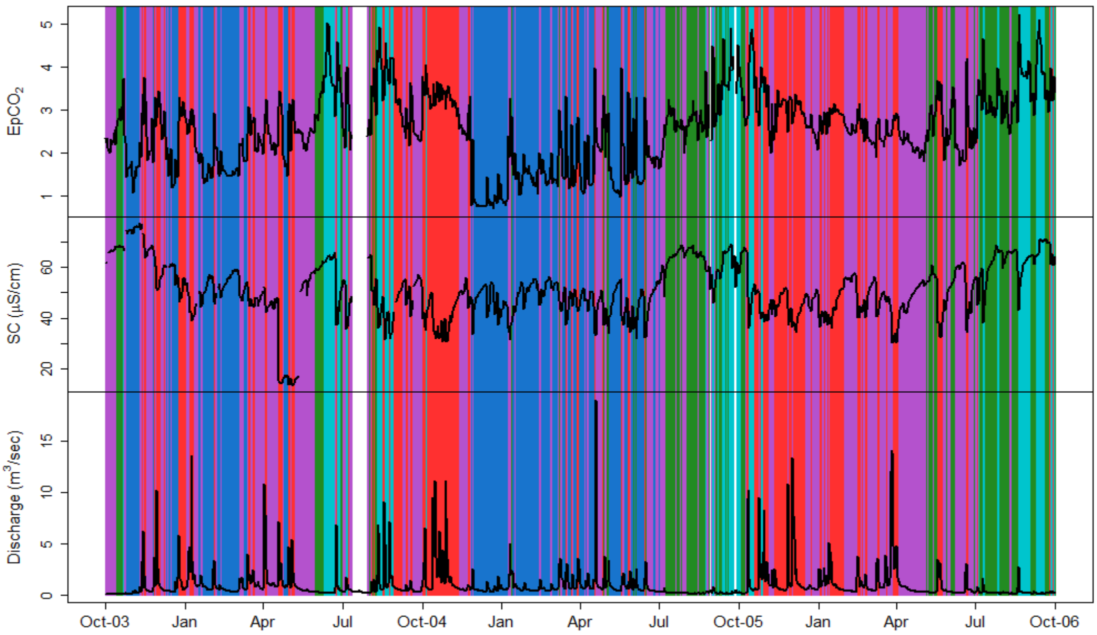


FIGURE 3.9: Class membership of days obtained using K-means clustering of the $EpCO_2$ curves' splines coefficients. The solid curves are the corresponding daily averaged $EpCO_2$ (top), SC (middle) and water discharge (bottom).

A FPCA was applied to the daily curves of $EpCO_2$ and the first 3 functional principal components were found to account for 97.4% of the variations around the mean $EpCO_2$ day curve, see the scree plot in Figure 3.10. The first 3 eigenvalues are 17.4, 2.3 and 1.4, respectively. Figure 3.11 presents the first three functional principal components, or alternatively eigenfunctions, by displaying the mean $EpCO_2$ curve and the effects of adding and subtracting a multiple or a small amount $\mathfrak{D}$ of each principal component curve. The constant $\mathfrak{D}$ is calculated as the root mean square difference between $\hat{\mu}(t)$ and its overall time average $\bar{\mu}$ as follows:

$$\mathfrak{D} = \sqrt{T^{-1}\sum_{t=1}^{T}[\hat{\mu}(t) - \bar{\mu}]^2}, \tag{3.14}$$

where $\bar{\mu}$ is obtained as $T^{-1}\sum_{t=1}^{T}\hat{\mu}(t)$ and T is the number of observed time points per day equals to 96 in our case. As can be seen from Figures 3.10 and 3.11, the first principal component accounts solely for 80% of the variation, reflecting a constant vertical shift in the mean curve of $EpCO_2$ across the 24 hours of a day. This large percentage indicates that this type of variation dominates all other types of variability. The first principal component function suggest that days for which the first FPC score is high will have higher $EpCO_2$ at night on average combined with higher levels during day time as well. The second eigenfunction accounts only for about 11% of the variability and seems to represent the contrast between day and night EpCO$_2$ levels. Whilst, the third component marginally contributes to the data variance with 6.5% and can be related to an intra-day trend effect.

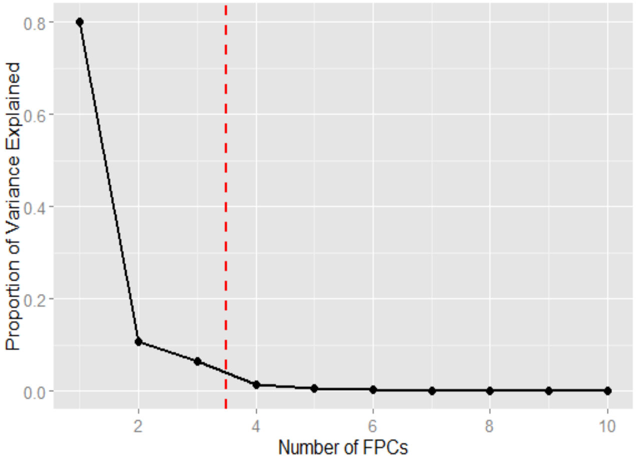


FIGURE 3.10: Scree plot of the number of FPCs versus the proportion of variance accounted for.

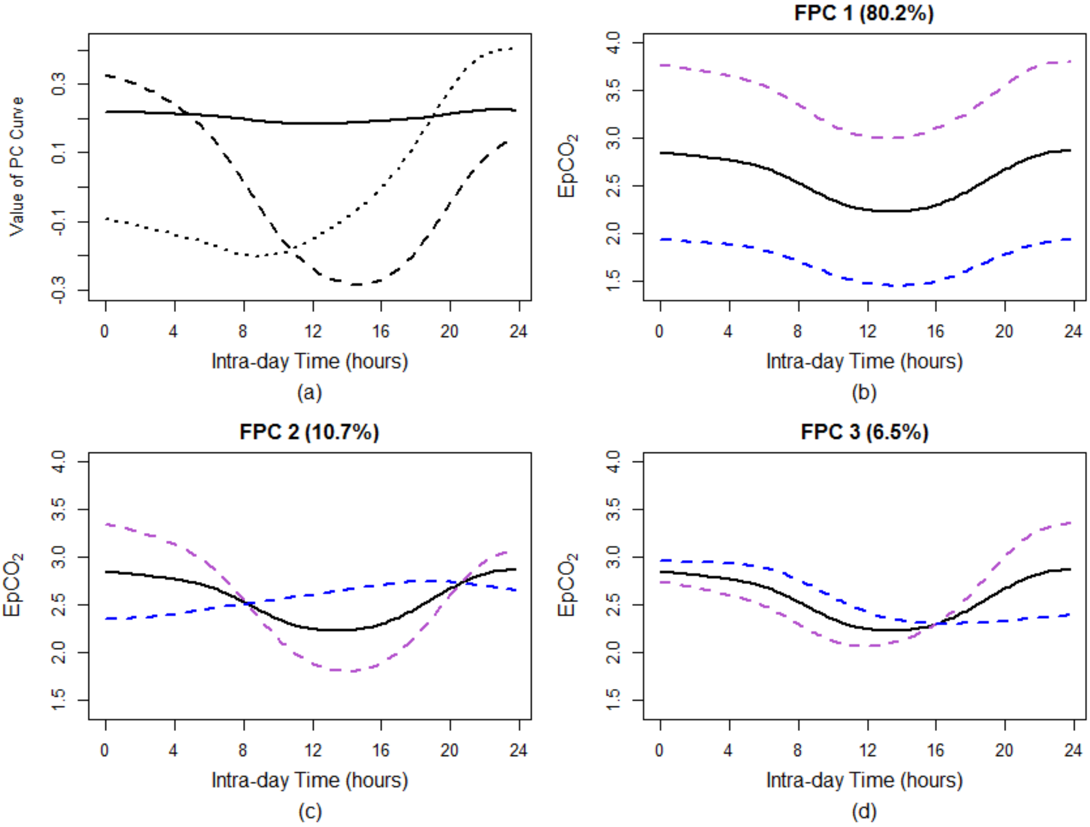


FIGURE 3.11: (a) Plot of the first FPC (solid line), second FPC (dashed line) and third FPC (dotted line). (b-d) Plots of the functional EpCO$_2$ mean (solid curve) and the effects of adding (- - -) and subtracting (- - -) a suitable multiple of each PC curve.

To obtain a more interpretable set of eigenfunctions, the VARIMAX rotation is computed and the results are displayed in Figure 3.12. The second rotated component accounts for 38.4% of the total variance and reveals variation that is strongest in daytime. While, the first and third components account for 34.5% and 24.5% of the variance, respectively, and capture primarily the variations occurring during the second half of night prior to sunlight and the variability occurring during the first half of night, respectively.

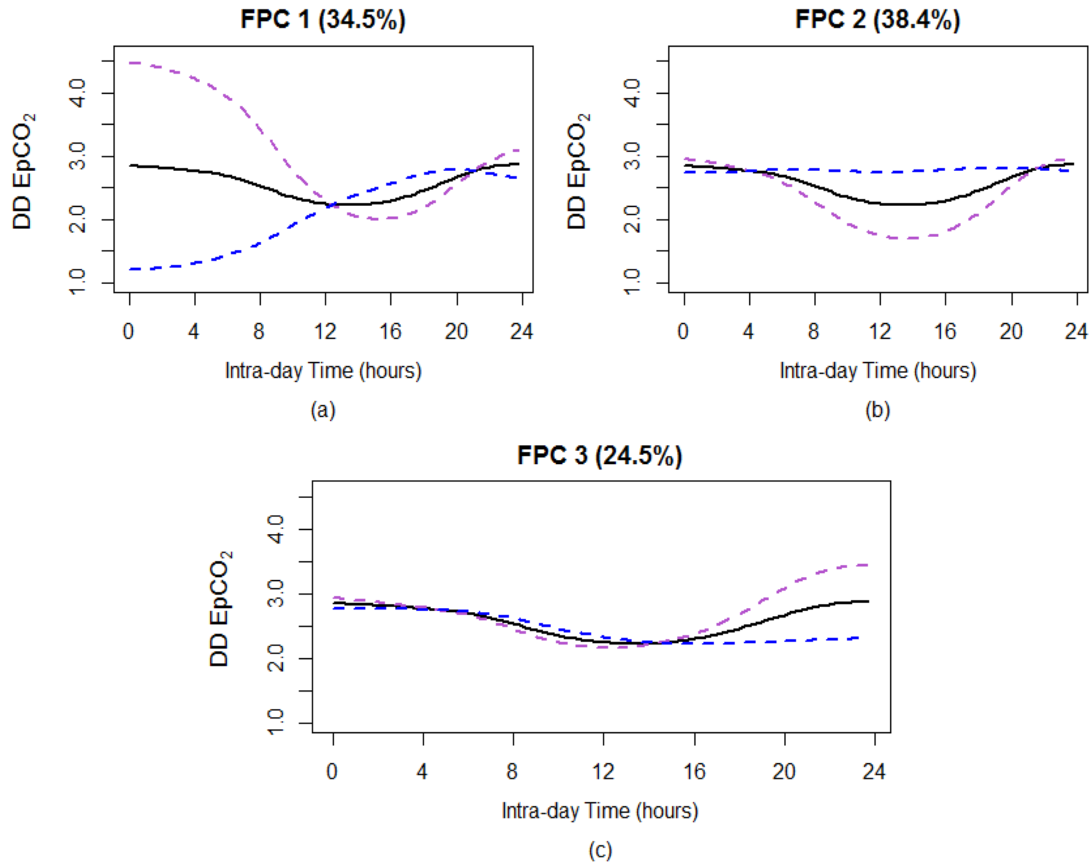
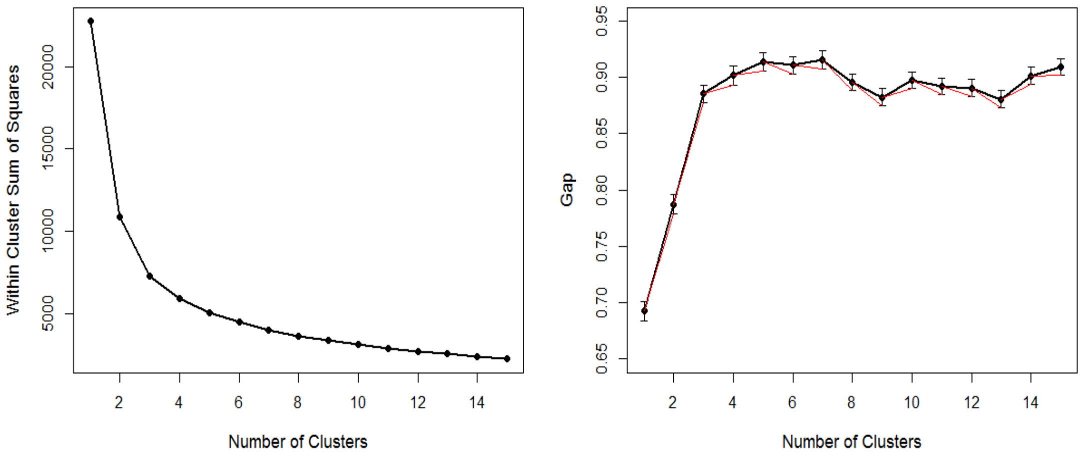


FIGURE 3.12: Plots of the functional $EpCO_2$ mean (solid curve) and the effects of adding (- - -) and subtracting (- - -) a suitable multiple of the (a) first, (b) second and (c) third VARIMAX rotated PC curves.

After identifying the 3 major directions of functional variations, each daily curve can be summarized by its corresponding 3 functional principal component scores. This reduces the dimensionality of the data to 3 attributes per day. Subsequently, the standard K-means algorithm is applied to cluster the daily patterns of $EpCO_2$ using either the scores of the original FPCs or the rotated FPCs defining the daily curves. Classification based on either ordinary principal components or the rotated components yielded similar results. This is because the rotated FPCs are as effective as their un-rotated counterparts in approximating the original curves. Therefore, all subsequent clustering analysis is based on the un-rotated version of FPCs.

The gap statistic has been initially used to identify the statistically optimal number of clusters. For each reference data out of 500 generated data sets, the K-means clustering algorithm was applied for a range of different number of clusters and the corresponding within-cluster sum of squares were obtained. The gap statistic was then calculated for each number of clusters ranging between 1 and 15. Figure 3.13(a-b) displays both the within class dispersion and the gap statistic against the number of clusters, when the classification is based on the FPC scores. The L-curve of the K-means clustering of the daily $EpCO_2$ based on the first 3 FPC scores is unclear regarding the correct number of clusters. However, for many runs, the gap statistic has consistently identified 5 clusters for the different daily patterns of $EpCO_2$ (the first number at which the red line has a negative slope). Accordingly, a K-means clustering with 5 centers is applied to the scores of the first three FPCs. The agreement between the classification based on the B-spline basis coefficients and that based on the FPC scores is quite high with an ARI of 0.86. The contingency table used to compare the results of both classifications is displayed below in Table 3.1. This is unsurprising since the principal component scores represent the coefficients of an orthonormal basis used to approximate the curves $z_i(t) = x_i(t) - \bar{x}(t)$; and hence this small number of coefficients represents a compact approximation for the coefficients of the 15 B-spline basis functions. In addition, the first 3 FPCs explain more than 97% of the total variation, providing a fairly good approximation to the original smooth curves. This means that the first 3 FPCs contribute most to the discrimination between the daily curves of $EpCO_2$.



(a) (b)

Figure 3.13: (a) L-curve and (b) gap statistic plot for the K-means clustering of daily $EpCO_2$ curves approximated with their first 3 FPC scores.

Figures 3.14 and 3.15 display the clustering results based on the principal component scores. Although the curves reconstructed using the first 3 functional principal components are smoother than the original curves, the underlying grouping structure is similar to that based on the spline coefficients. This is because the formation of groups is mainly based on the mean level and not the local variations. In addition to the dominant effect of mean level on the grouping structure, it is evident that the daily patterns within each group have also different shapes. For example, some groups have a clear intra-daily cycle with a considerable drop in $EpCO_2$ level during day-light hours, while others have a relatively stable $EpCO_2$ level over the day. This grouping structure suggests that the mixture of hydrological and climatological conditions affect the biological and hydrological activity within the catchment, which in turn contribute to the differences in the diurnal patterns of $EpCO_2$.

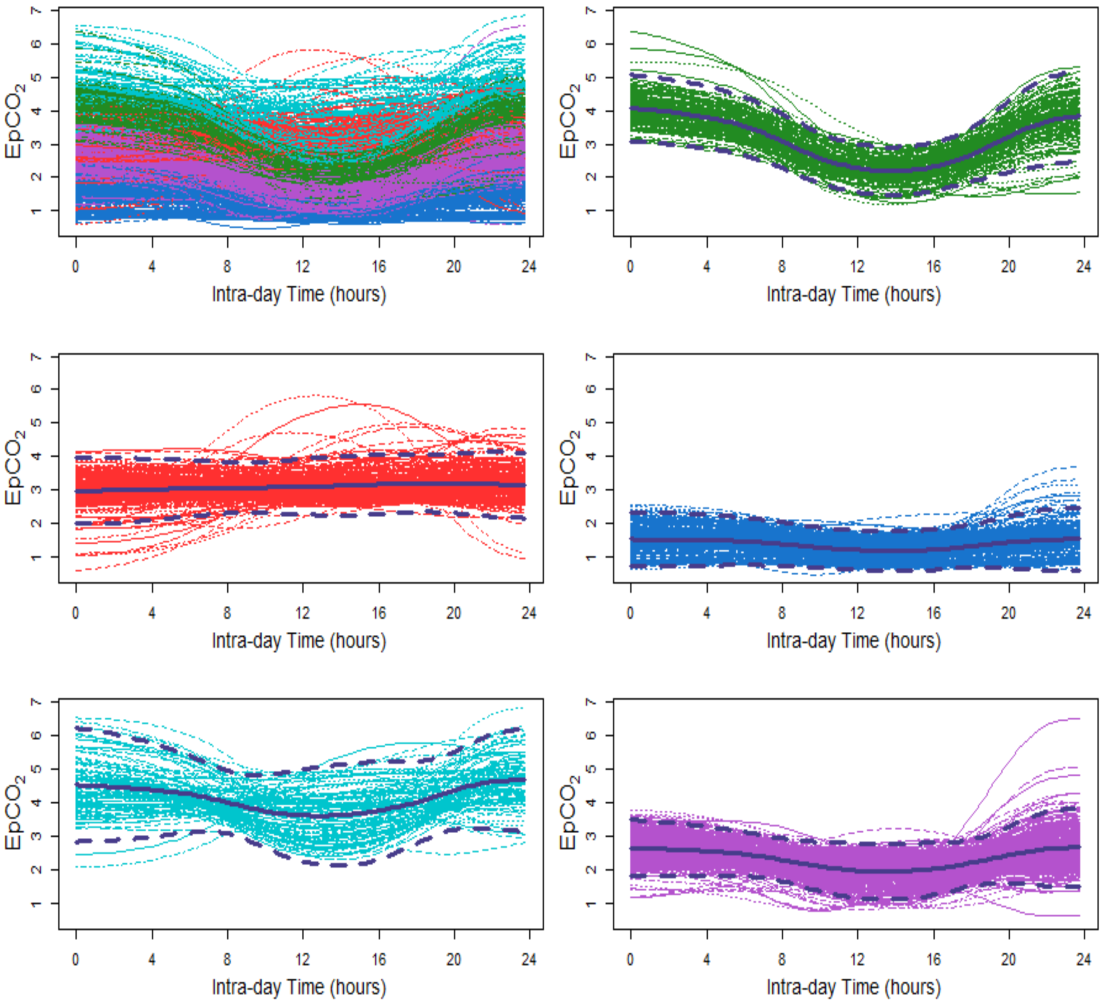


FIGURE 3.14: K-means clustering results of the daily $EpCO_2$ curves using their first 3 FPCs. The solid and dashed purple curves represent the mean curve and ± 2× standard deviation bands, respectively.

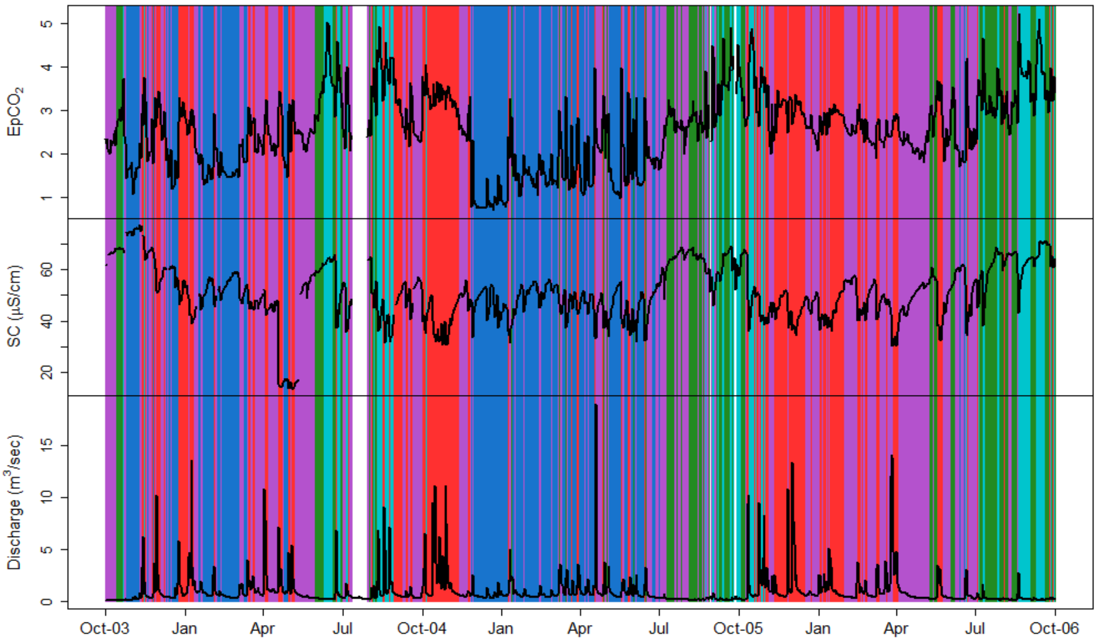


FIGURE 3.15: Class membership of days obtained using K-means clustering of the $EpCO_2$ curves' FPC scores. The solid curves are the corresponding daily averaged $EpCO_2$ (top), SC (middle) and water discharge (bottom).

	Basis coefficients' classes				
FPC scores' Classes	1 (Purple)	2 (Turquoise)	3 (Green)	4 (Red)	5 (Blue)
1 (Purple)	341	0	21	0	1
2 (Turquoise)	0	90	0	7	0
3 (Green)	1	9	138	1	0
4 (Red)	14	0	1	220	0
5 (Blue)	7	0	0	0	223

TABLE 3.1: Contingency table of the K-means classification based on the B-splines coefficients and that based on the functional principal components scores.

3.4.4 Effect of Time Components on Clustering

The previous functional data analysis has totally ignored the time component and the serial dependence between the daily curves. Performing either FPCA or FCA without taking into account the time effect may result in misleading conclusions and inefficient inference procedures. Although the previous analysis is purely descriptive, the lack of independence between the curves and ignorance of trend and seasonal effects make the resulting principal components and clustering structure not being the adequate dimension reduction for the data. The previous analysis has indicated the wide range of variability between the daily curves of $EpCO_2$ which could be, to some extent, attributed to seasonality. To avoid the effects of trend and seasonality on the mean level from dominating the clustering structure of the curves, the global trend as well as the

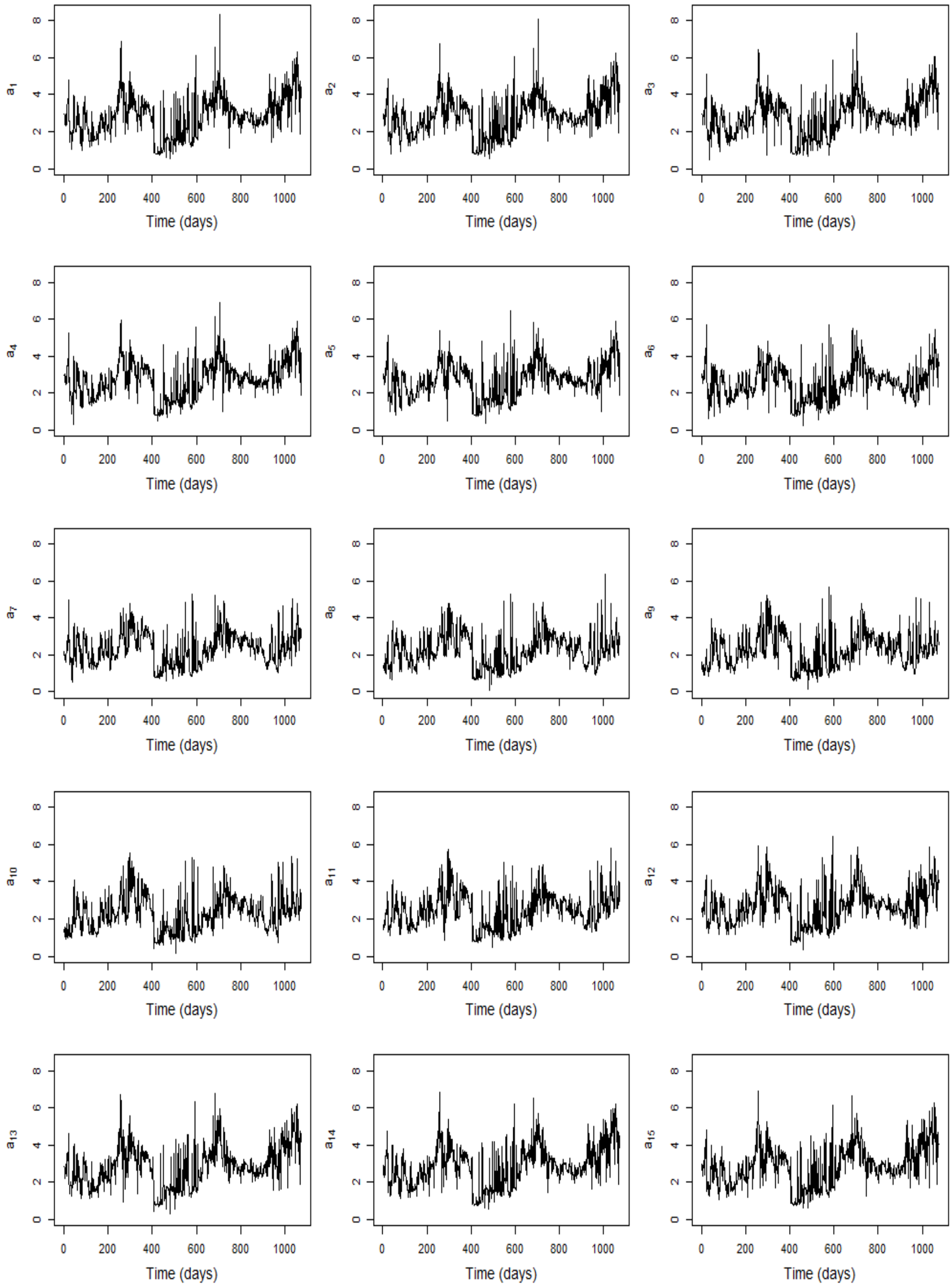


FIGURE 3.16: Time plots of the (15) B-spline basis coefficients ($\mathbf{a}_k, k = 1, \ldots, 15$) of the EpCO$_2$ daily curves.

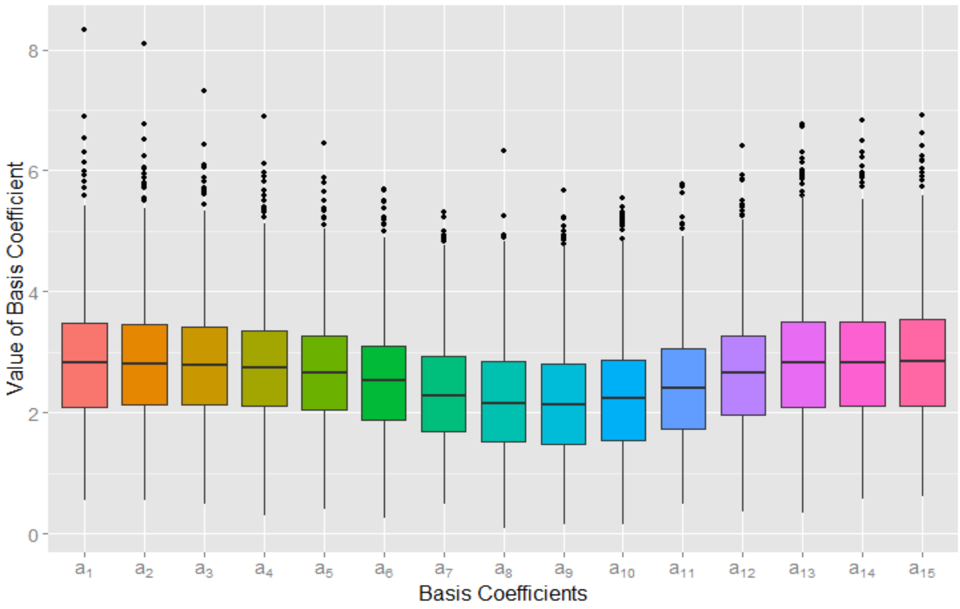


FIGURE 3.17: Box-plots of the B-spline basis coefficients of the $EpCO_2$ daily curves.

seasonal effects are estimated then subtracted from the discrete data before any further analysis. To determine the degree of smoothness of this estimated trend and seasonal cycle, the following analysis on the B-spline basis functions coefficients is performed.

As explained before, each daily curve is summarized by its corresponding 15 basis coefficients $\mathbf{a} = (a_1, \ldots, a_{15})^\top$. Figure 3.16 shows the plots of each basis coefficient over time. In all the panels, there is evidence for a global slightly increasing trend and a repeated seasonal cycle every year. The box-plots in Figure 3.17 indicate also that the average daily pattern of $EpCO_2$ has been successfully captured by the basis coefficients. To investigate the strength of the time dependence within the time series of each coefficient, a plot of the value of each coefficient at time $i+1$, $a_{i+1,k}$, versus its value at time i, $a_{i,k}$, is produced (see Figure 3.18). All the plots suggest a strong positive linear relationship implying the presence of a strong time-dependence between successive values of each basis coefficients. One way to statistically assess the time-dependence structure in time series is to estimate the autocorrelation function (ACF). The ACF of each of the coefficients, in Figure 3.19, decays slowly and becomes almost insignificant after 3 months (90 lags). Nevertheless, each ACF appears to decay with a periodic pattern of approximately 1 year, if the plot of the ACF is extended to cover a larger number of lags (see Figure 3.20). It is noticed that the ACF of the boundary coefficients is more persistent and decays with a slower rate than that of the middle coefficients, implying that the measurements during the day-time are less correlated than the night measurements.

The above analysis of the basis coefficients suggested the evidence of a significant correlation between the daily curves that lasts up to 3 months with a repeated cycle every

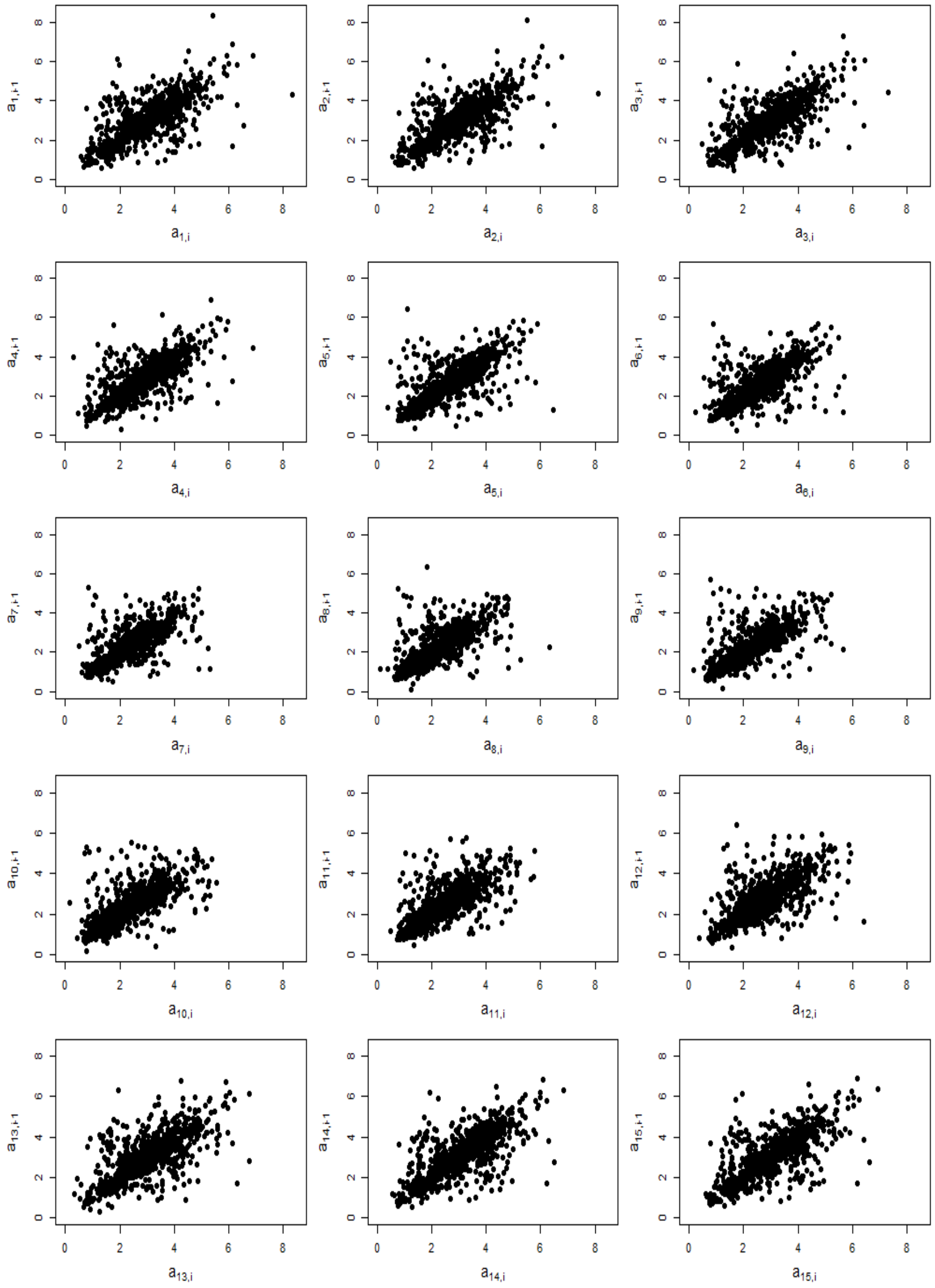


FIGURE 3.18: Scatter plots of the value of each B-spline basis coefficients ($\mathbf{a}_k, k = 1, \ldots, 15$) of the EpCO$_2$ daily curves at time i versus its value at time $i-1$.

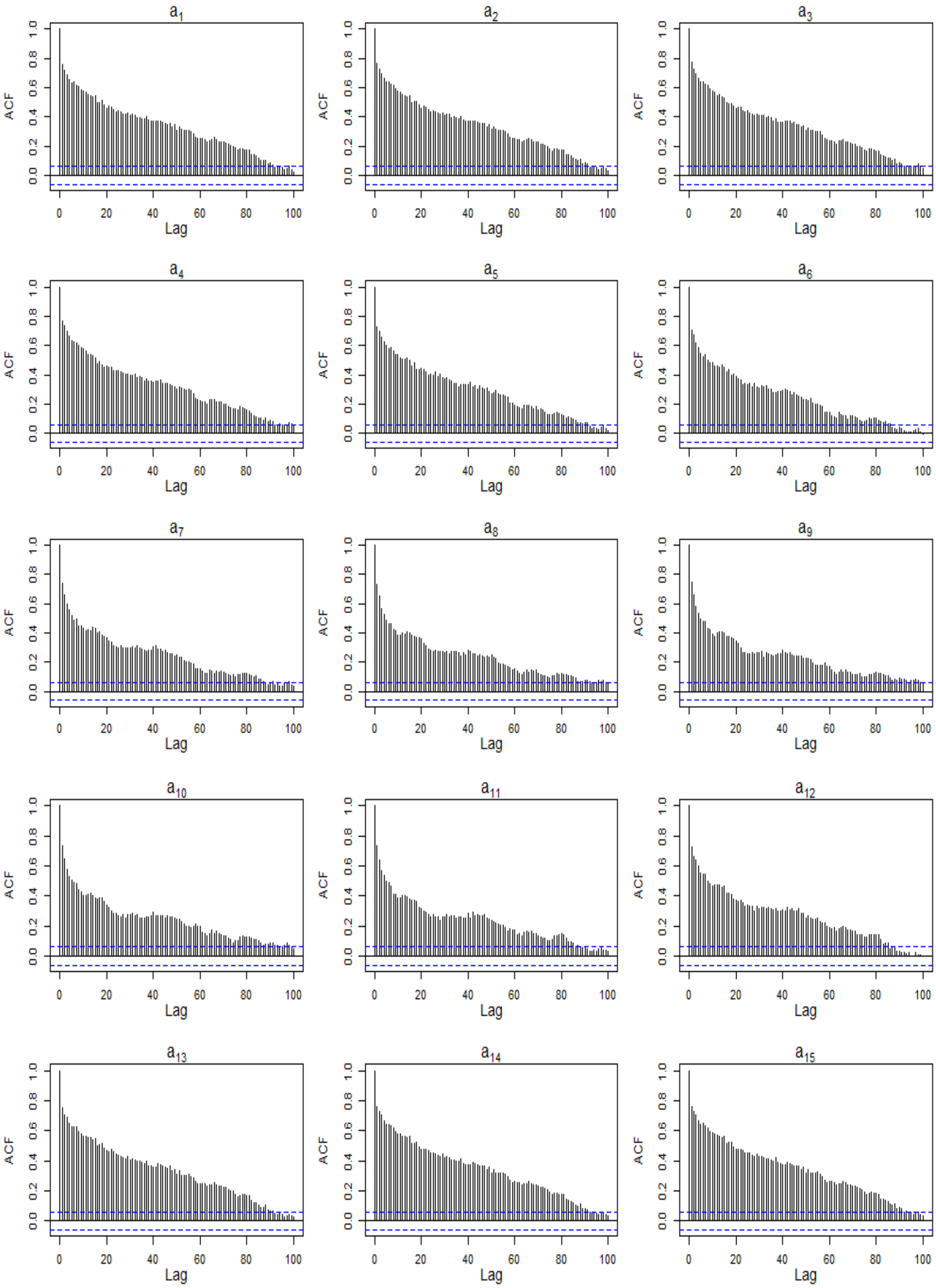


FIGURE 3.19: ACF of the B-spline basis coefficients ($\mathbf{a}_k, k = 1, \ldots, 15$) of the EpCO$_2$ daily curves (maximum number of lags = 100).

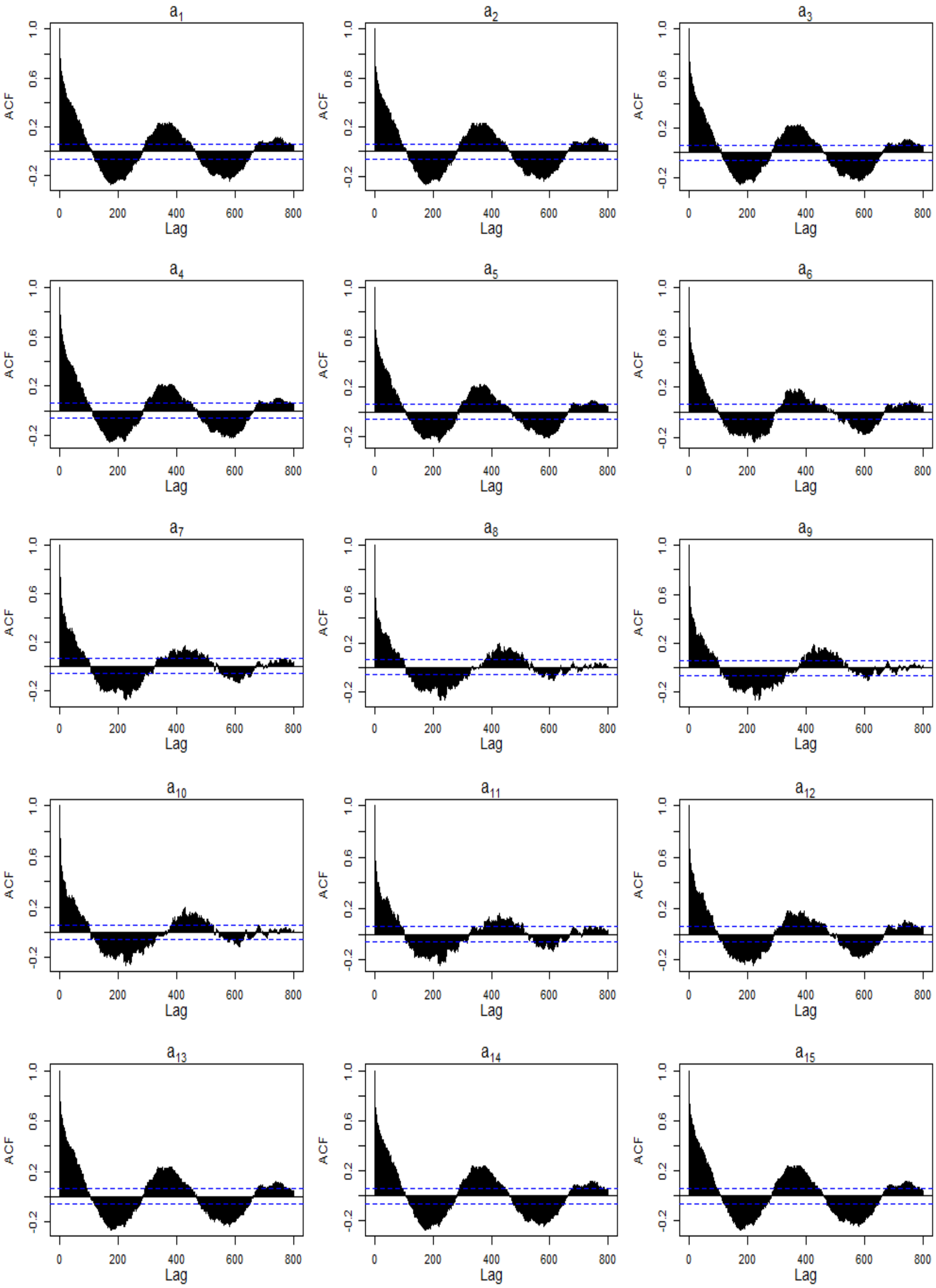


FIGURE 3.20: ACF of the B-spline basis coefficients ($\mathbf{a}_k, k = 1, \ldots, 15$) of the EpCO$_2$ daily curves (maximum number of lags =800).

year. For this reason, it was decided to estimate a global mean $\hat{\mu}_t^G = \hat{g}(X_t^{\text{Time}})$, for the overall trend and the seasonal effect, using cubic regression splines with a knot placed every 3 months. This estimated global mean function (Figure 3.21) is then evaluated at the set of regularly spaced 15-minute time points from 1^{st} October 2003 at 00:00 up to 30^{th} September 2006 at 23:45, and the corresponding fitted values are obtained. Next, the detrended and deseasonalized (DD) time series of EpCO$_2$ is obtained by subtracting the fitted values from their corresponding observed values. FDA is then applied to the resulting DD series of EpCO$_2$, where the discrete DD series are transformed to continuous smooth daily curves. Each of these smooth daily curves is fitted using 15 cubic B-spline basis functions with a knot placed every 2 hours and a roughness penalty with λ equal 1, for the same reasons mentioned before. The estimated auto-correlation functions of the new basis coefficients in Figure 3.22 suggest that although the coefficients still exhibit serial dependence, the magnitude of this dependence has dramatically dropped after removing the trend and seasonal effects. This implies that a large share of the autocorrelation is mainly attributed to trend and seasonality. Using a higher number of basis functions to estimate the seasonal pattern in addition to the global trend, introduces an artificial auto-correlation pattern in the data. For example, see the ACF, in Figure 3.23, of a_1 after removing the average trend estimated using cubic regression splines with a knot placed every 3 months, 2 months, 1 month and ≈ 21 days.

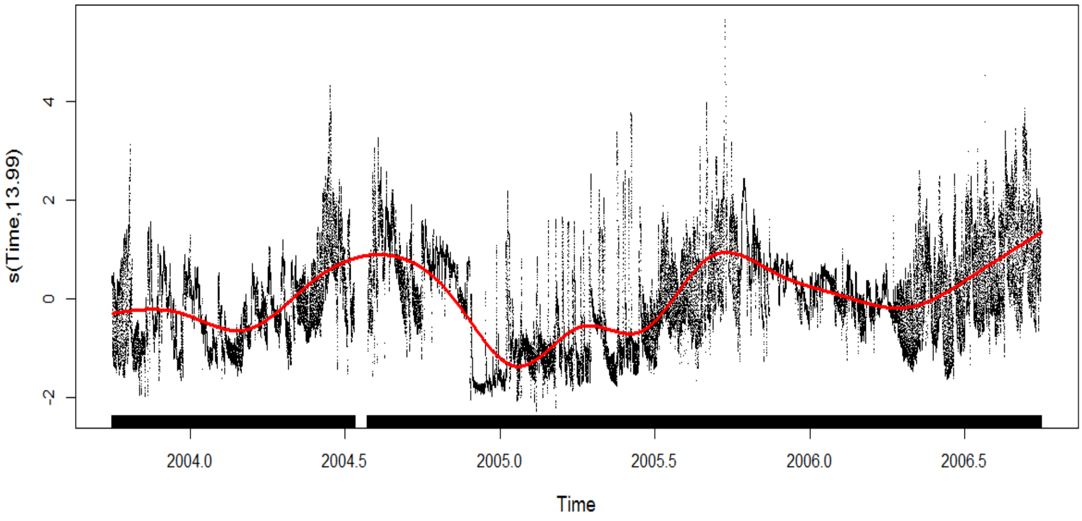


Figure 3.21: Estimated overall trend and seasonal effects using cubic regression splines of the continuous time variable with a knot placed every 3 months. The black dots are the model residuals.

After removing the global mean representing the overall trend and seasonality effect, the first 3 FPCs are found to account for 95.7% of the total variability in the daily curves of EpCO$_2$, see Figure 3.24. Since the main interest lies in the daily pattern of EpCO$_2$, the FPCA was performed without centering the daily DD curves at their mean curve $\mu(t)$. Figure 3.25 displays the first 3 FPCs and the effects of adding and subtracting each

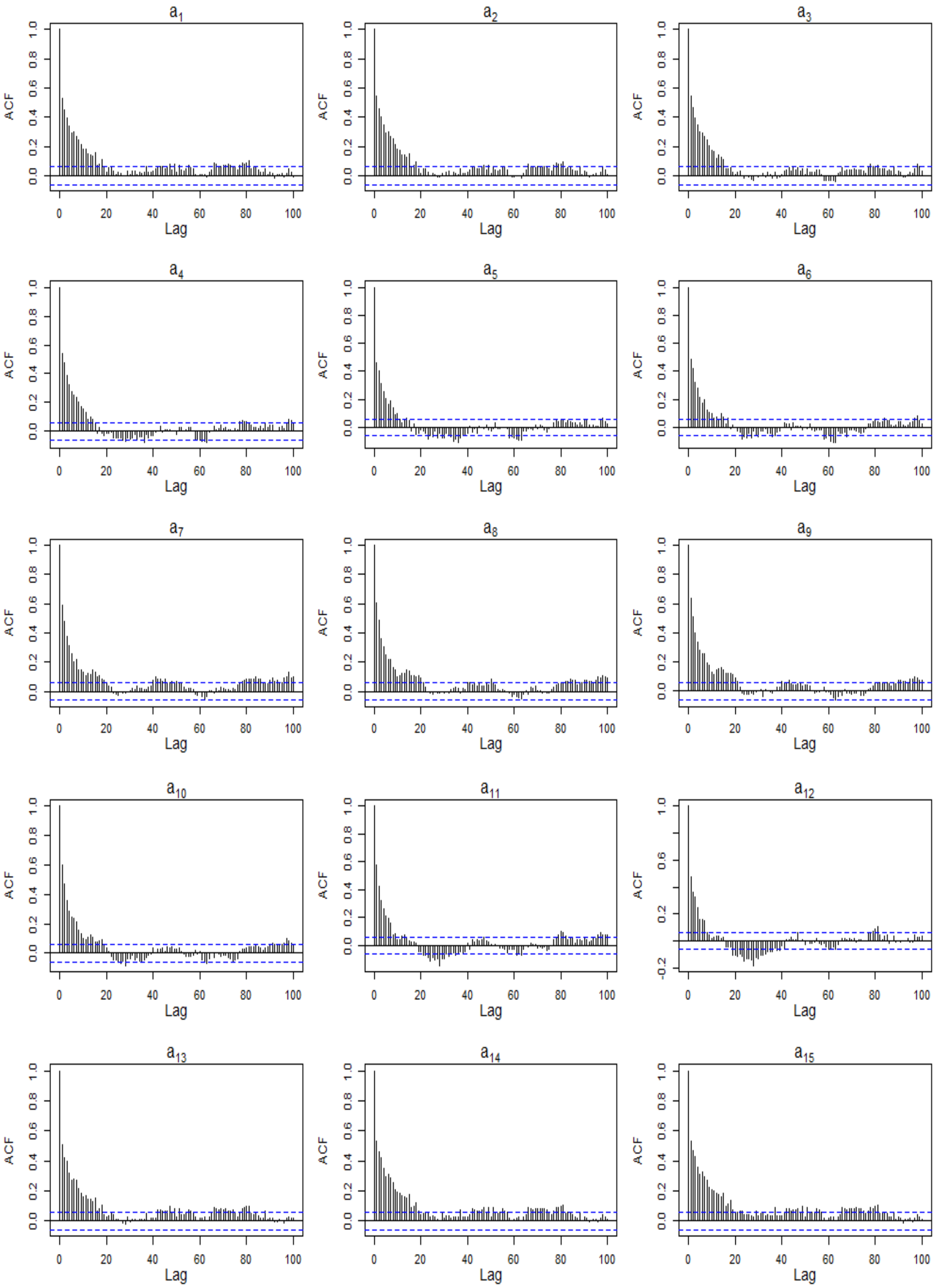


FIGURE 3.22: ACF of the B-spline basis coefficients ($\mathbf{a}_k, k = 1, \ldots, 15$) of the EpCO$_2$ daily curves after removing the average trend and seasonality from the original EpCO$_2$ data.

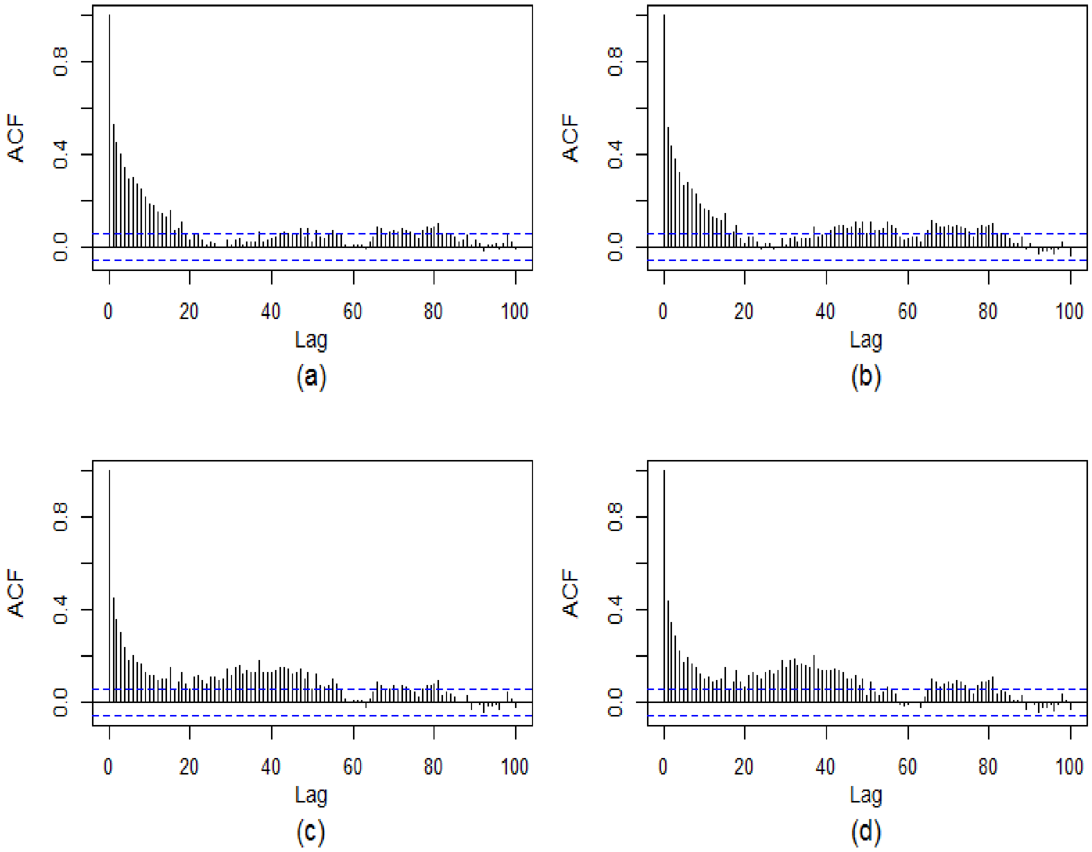


FIGURE 3.23: ACF of the first basis coefficient ($\mathbf{a}_1$) of the $EpCO_2$ daily curves after removing the average trend and seasonality from the original $EpCO_2$ data fitted using cubic regression splines with (a) a knot every 3 months, (b) a knot every 2 months, (c) a knot every 1 month and (d) a knot every $\approx$ 21 days.

principal component curve to the mean DD $EpCO_2$ curve. It is evident that the major modes of variability in the data, even after removing the effect of trend and seasonal cycle, are still the same. However, the first FPC accounts only for 58% of the variations compared with 80% before removing the global mean. This reflects the drop in the effect of the mean on the variations in the $EpCO_2$ curves. In contrast, the effect of discrepancy between day and night, reflected by the second FPC, on the variability is inflated and accounted for 26.5% of the variance. Similarly, the percentage of variance accounted for by the third eigenfunction, related to the intra-day trend effect, has increased from 6.5% to $\approx$ 11%.

Since the first 3 FPCs account for almost 96% of the variance in the data, each daily curve of the DD data can be efficiently summarized by the corresponding scores of those 3 FPCs, which are orthogonal at a fixed time i. Then, a conventional K-means clustering procedure can be applied to these scores. For each possible number of clusters ranging from 1 to 15, 500 reference data sets are generated and the gap statistic is calculated.

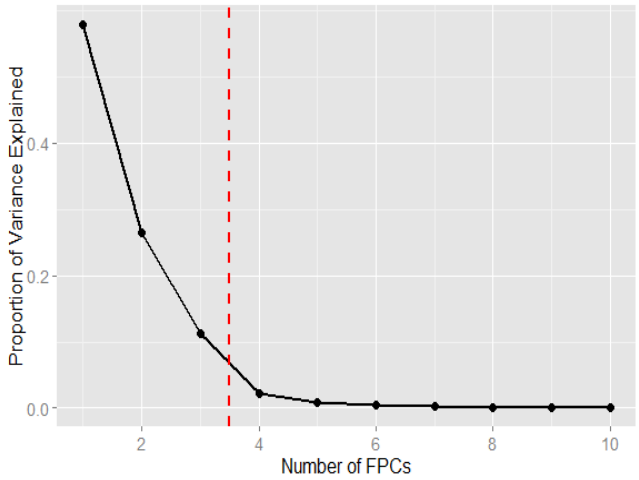


FIGURE 3.24: Scree plot of the number of FPCs versus the proportion of variance accounted for, after removing the global trend and seasonality.

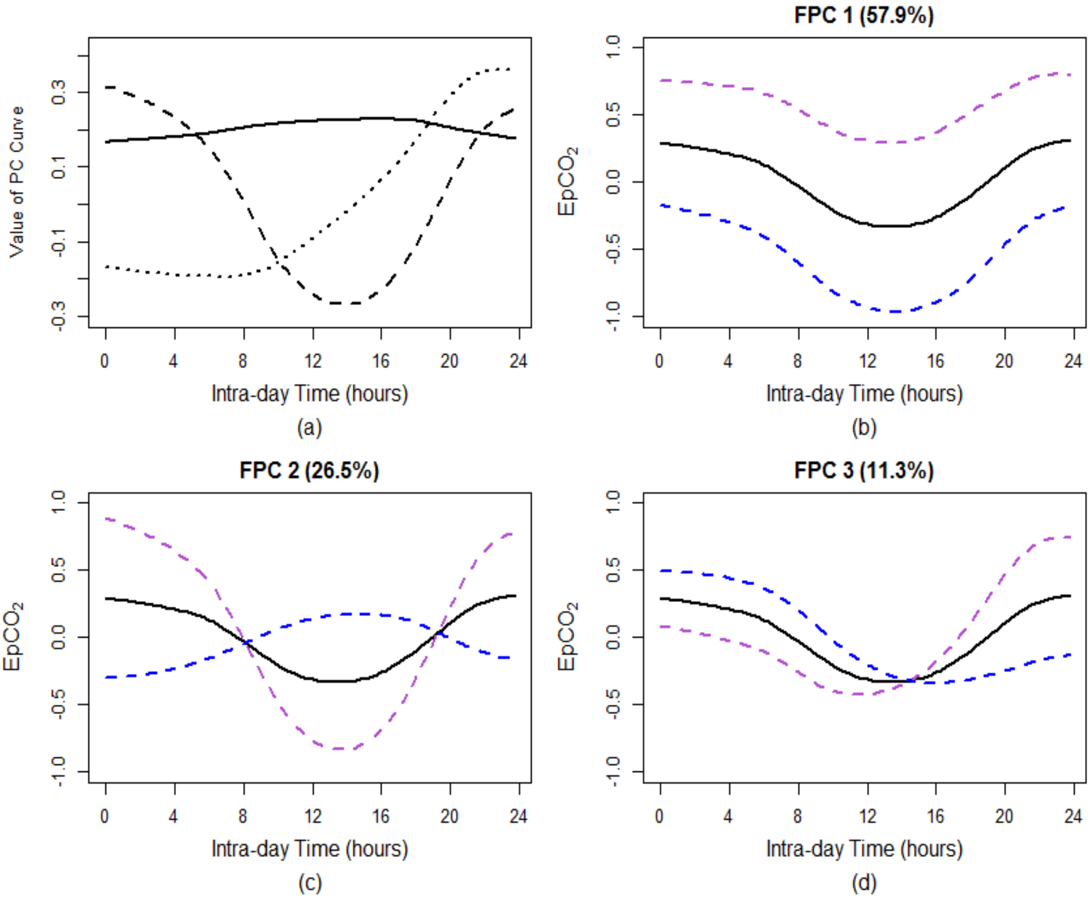


FIGURE 3.25: (a) Plot of the first FPC (solid line), second FPC (dashed line) and third FPC (dotted line). (b-d) Plots of the mean EpCO$_2$ curve and the effects of adding (- - -) and subtracting (- - -) a suitable multiple of each PC curve of the DD curves.

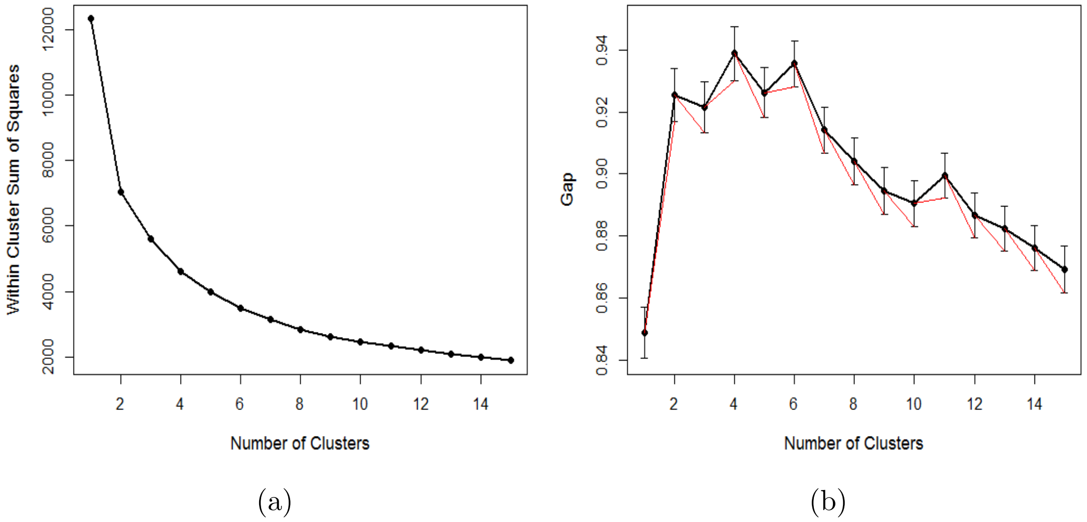


FIGURE 3.26: (a) L-curve and (b) gap statistic plot for the K-means clustering of daily DD EpCO$_2$ curves approximated with their first 3 FPC scores.

The gap statistic (Figure 3.26) has optimally identified 2 clusters for the daily curves after removing the effect of trend and seasonality. These 2 clusters of curves are displayed in Figure 3.27, where the discrepancy between the two groups of curves is mainly based on the average level across the day. It is also noticed that the class of turquoise curves is more chaotic and variable relative to the blue class. The variability bands (dashed purple lines) show that some of the curves classified to the turquoise cluster exhibit peculiar daily patterns and do not particularly follow the main features of the cluster.

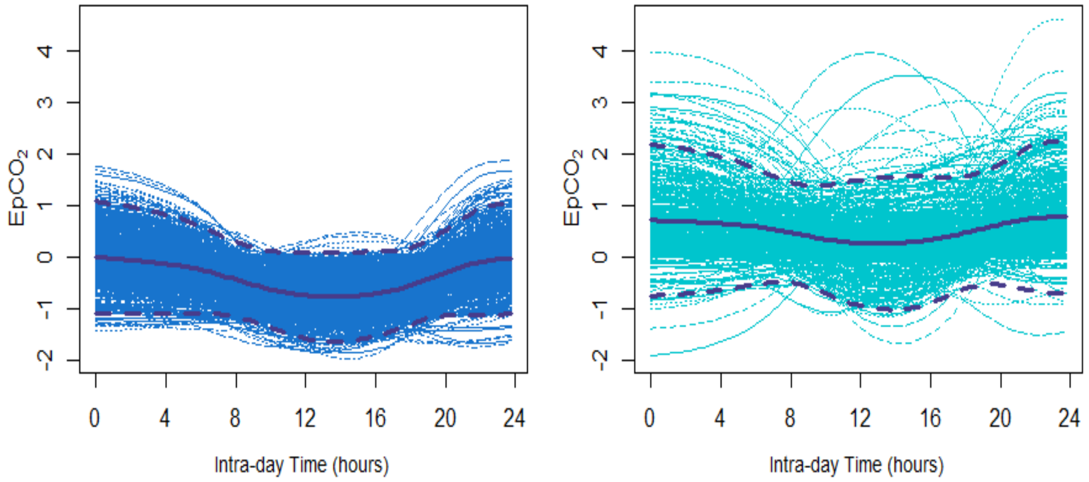


FIGURE 3.27: K-means clustering results of the daily DD EpCO$_2$ curves smoothed using the first 3 FPCs. The solid and dashed purple curves represent the mean curve and ± 2× standard deviation bands, respectively.

Figure 3.28 shows that by limiting the effects of trend and seasonality on the clustering

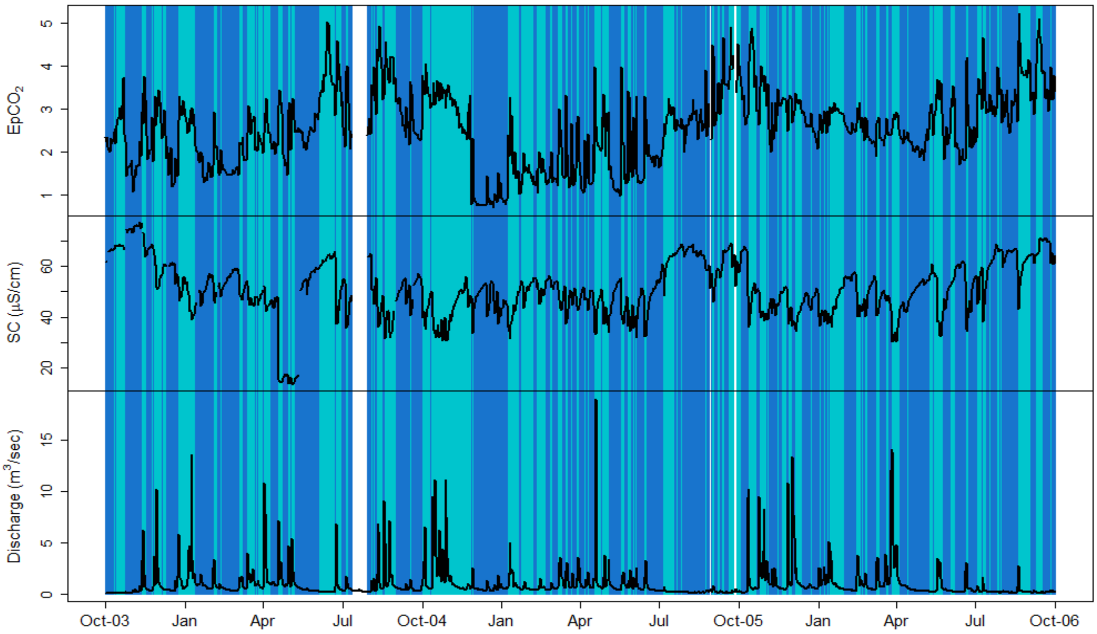


FIGURE 3.28: Class membership of days obtained using K-means clustering of the DD EpCO$_2$ curves' FPC scores. The solid curves are the corresponding daily averaged EpCO$_2$ (top), SC (middle) and water discharge (bottom).

structure of the curves, the hydrological conditions, to a great extent, become responsible for the resulting classification. Regardless of some minor exceptions, the blue curves correspond to days with relatively stable hydrological regime reflected by stable levels of conductivity. Whereas, the turquoise curves often underlie days with unstable surface water environment induced by drops in conductivity which are mostly attributed to heavy flow events.

3.5 Summary and Discussion

Functional data analysis has proven to be an efficient and effective tool in analyzing high-frequency environmental time series. Analyzing the EpCO$_2$ high-frequency data using an FDA approach has reduced the data dimensionality from 15-min resolution discrete observations to 1-day resolution functional observations. That is, the dimension of the data has dropped by 96 times. FDA has also allowed the high-frequency time series to be analyzed without being concerned about the high-correlations between the (high-frequency) measurements within the same functional unit (day). The FDA techniques, including FPCA and FCA, have also been very useful in exploring the intra-daily pattern of EpCO$_2$ over time and in identifying the different patterns and the sources of variability between them.

Using penalized cubic B-splines to fit the smooth daily curves has mapped all the key features in the underlying data in a very flexible and computationally efficient way. The primary exploratory FDA has shown that the daily curves are quite variable from one day to another with few outliers in shape and magnitude. More advanced FDA techniques, specifically FPCA, has indicated that the main sources of variability between the curves are their deviations from the mean level in addition to the contrast effect between day and night that varies across the year. The initial functional clustering whether based on the spline coefficients or the primary FPC scores of the curves suggested that the underlying mixture of hydrological and climatological conditions are the main drivers of the different daily patterns. A certain climate combined with a certain hydrological state affect the biological and hydrological activity of the catchment in a way that induces a particular daily pattern of $\mathrm{EpCO_2}$. For example, in a very hot and dry period the biological activity becomes more dominant and more $\mathrm{CO_2}$ tends to be consumed during the day; whereas in a wet period the heavy rainfall dilutes the intra-daily cycle of $\mathrm{EpCO_2}$.

Despite the temporal dependence structure between the daily curves of $\mathrm{EpCO_2}$, the resulting grouping structure suggests that the classification is quite rough (not smooth) and heavily relies on the differences between the mean level. This is attributed to the ignorance of both the time component and the effect of existing trend and seasonality. Therefore, a smooth global mean, for both the trend and seasonal effects, has been estimated and used to obtain a detrended and deseasonalized time series. Based on the detrended and deseasonalized functional data, only 2 clusters have been identified for the daily patterns. This clustering structure was believed to be driven mainly by hydrology, where one of the classes mostly contains days following sudden and heavy flow events and the other seems to highlight the more stable periods.

In the above analysis, the time dependence between the daily curves has been neglected and all potential information carried by nearby observations is discounted. This in turn could lead to inadequate dimensionality reduction and misleading conclusions. Therefore, taking the time dependence between the curves into account is the main focus of the next chapter, where a new frequency domain approach is considered.

Chapter 4

Accounting for the temporal correlation in FDA

Until now, the FDA methods, including FPCA and FCA, have considered the analysis of the consecutive daily curves of $\mathrm{EpCO_2}$ assuming complete independence between curves. However, the nature of the data as well as the previous analysis indicate the evidence of strong time dependence between the daily $\mathrm{EpCO_2}$ curves, even after removing the global trend and seasonality effect. This temporal correlation has to be accounted for appropriately in both the FPCA and FCA, as failure to account for this temporal correlation could lead to inappropriate dimension reduction and hence misleading results.

Most of the existing techniques in FDA have been developed for independent functional observations. This is a major shortcoming, as in many applications the functional data are either spatially or temporally dependent. Examples include water quality trends at a range of spatially correlated sites (Haggarty et al., 2015), annual temperatures measured at different locations (Giraldo et al., 2012) and daily patterns of environmental data (Hormann et al., 2014). Discarding the dependence between functional observations not only results in inefficient inference but also inappropriate dimension reduction, as information carried by nearby observations is totally ignored. Therefore, recent developments in FDA have been devoted to both spatially and temporally correlated functional data. Numerous methods for clustering correlated functional data have been proposed. Giraldo et al. (2012) have considered the spatial correlation between annual temperature curves measured across Canada in a hierarchical functional clustering approach, through a spatially weighted dissimilarity matrix. While, Haggarty et al. (2015) have proposed a similar approach to cluster spatially correlated functional water quality data, based on the stream distance between monitoring sites. Univariate and multivariate functional clustering models have been also developed by Haggarty et al. (2012b) to investigate the spatio-temporal structure of water quality variables in a number of Scottish lakes. Other approaches have been introduced for the functional principal component analysis of correlated functional data. One suggestion was to obtain the smooth functional data, model their corresponding basis coefficients over time using vector auto-regressive models, then remove the estimated temporal structure and apply FPCA on the residuals of the coefficients which are less correlated (Jaimungal and Eddie, 2007). A clear disadvantage of this method is that by removing the temporal structure in the data, we might loose valuable information and existing temporal patterns in the data which are of interest to the analyst. Another way to account for spatial correlation in FPCs is considered by Liu et al. (2016), who propose a spatial principal component analysis using a conditional expectation framework to explicitly estimate spatial correlations and reconstruct individuals curves based on estimated scores. Recently, alternative approaches that account for the temporal correlation in the frequency domain have been developed for FPCA by Hormann et al. (2014) and Panaretos and Tavakoli (2013). In this thesis, we will only consider the method proposed by Hormann et al. (2014), which has proven to outperform the traditional FPCA in case of temporally correlated functional observations.

In functional time series, the data are viewed as the realization of a functional stochastic process $(X_i(t) : i \in \mathbb{Z})$, where i is a discrete time parameter and t is a continuous time parameter. For example, in daily functional observations, the curve $x_i(t)$ denotes the observation on day i with intra-day time parameter t. Traditional FPCA assumes that each sample (curve) in the data is drawn independently from a stationary distribution and hence the PCs are invariant to the permutation of the sampled curves. However, in a functional time series, the i.i.d. requirements of FPCA are violated and hence the traditional FPCs will not provide the optimal dimension reduction, although they are consistently estimated (Hormann and Kokoszka, 2010). This is mainly because the traditional FPCA does not take into account the potential information carried by the past values of the functional observations. In particular, a FPC with a small eigenvalue and a negligible effect on $x_i(t)$ might have a major influence on $x_{i+1}(t)$. In addition, FPCs are treated as multivariate time series, where they are cross-sectionally independent at a fixed time i but still exhibit lagged cross-correlations. For this reason, the resulting FPCs cannot be analyzed componentwise and have to be considered as vector time series, which are more difficult to analyze and interpret. These limitations motivate the use of the dynamic FPCA, recently proposed by Hormann et al. (2014). The dynamic FPCA is a frequency domain version of the FPCA, in which the FPCs are considered as vector time series that can be analyzed componentwise, as the individual components are mutually independent at all lags and leads and account for most of the variations in the original process. This work has been inspired by Brillinger (1981) who founded the theory of PCA in the frequency domain.

This chapter mainly focuses on the analysis of functional time series taking into consideration the temporal dependence between the curves. The chapter first outlines the methodology of the newly available dynamic FPCA, mainly drawn from Hormann et al. (2014), then proposes an extension of the methodology to functional data approximated using all types of basis functions. Afterwards, the application of this dynamic FPCA to the EpCO$_2$ daily curves and the corresponding results are presented in Section 4.2. Next in Section 4.3, a novel clustering approach based on these dynamic FPCs is proposed and illustrated using the EpCO$_2$ data. Finally, the results of clustering are assessed using functional ANOVA in addition to some permutation F-tests in Sections 4.4 and 4.5 and the hydrological drivers of each cluster are studied in Section 4.6.

4.1 Dynamic Functional Principal Component Analysis

Consider a functional time series $(X_i : i \in \mathbb{Z})$, such that X_i belongs to the Hilbert space of complex-valued square integrable functions on $[0, 1]$ denoted by $L^2_H([0, 1])$. That is, $X_i = (X_i(t) : t \in [0, 1])$ with $\int_0^1 |X_i(t)|^2 dt < \infty$, where $|X(t)|$ is the modulus of $X(t)$ defined by $\sqrt{X(t)X^\bullet(t)}$ and $X^\bullet(t)$ denotes the complex conjugate of $X(t)$. The complex vector space definition is useful for the spectral methods being used to obtain the dynamic FPCs. The Hilbert space defined above is accompanied with the inner product $\langle x, y\rangle = \int_0^1 x(t)y^\bullet(t)dt$ and the norm $\|x\| = \sqrt{\langle x, x\rangle} = \sqrt{\int_0^1 x(t)x^\bullet(t)dt}$. Let $X \in L^p_H$ indicates that for some $p > 0$, $\mathbb{E}[\|X\|^p] < \infty$. Thus, any $X \in L^2_H$ has a mean curve $\mu = (\mathbb{E}[X(t)] : t \in [0, 1])$ and a covariance kernel operator $\mathcal{V}$ defined by:

$$\mathcal{V}(x)(t) = \int_0^1 V(t, s)x(s)ds, \quad \text{where } s, t \in [0, 1],$$

where $V(t, s) = \text{COV}\{X(t), X(s)\} = \mathbb{E}[(X(t) - \mathbb{E}[X(t)])(X(s) - \mathbb{E}[X(s)])^\bullet]$.

The process $(X_i : i \in \mathbb{Z})$ is said to be weakly stationary if, for all i,

1. $X_i \in L^2_H$;
2. $\mathbb{E}[X_i] = \mathbb{E}[X_0]$;
3. for all $h \in \mathbb{Z}$ and $s, t \in [0, 1]$:

$$\text{COV}\{X_{i+h}(t), X_i(s)\} = \text{COV}\{X_h(t), X_0(s)\} = V_h(t, s). \tag{4.1}$$

Note that $\mathcal{V}_h, h \in \mathbb{Z}$ is the operator corresponding to the auto-covariance kernel V_h and hence $\mathcal{V}_0 = \mathcal{V}$. The following subsection briefly revisits the construction of traditional FPCs and the corresponding Karhunen-Loève expansion and the subsequent subsections explain the theory and practical implementation of dynamic FPCA to functional data approximated using any type of basis functions. Afterwards, this dynamic FPCA is applied to the daily curves of EpCO$_2$ and the corresponding results are discussed.

4.1.1 Karhunen-Loève Expansion

Following Section 3.2, the covariance operator $\mathcal{V}$ of a zero mean process $X \in L^2_H([0,1])$ admits the following eigen-decomposition:

$$\mathcal{V}(x) = \sum_{m=1}^{\infty} \gamma_m \langle x, \xi_m \rangle \xi_m. \tag{4.2}$$

$(\gamma_m : m \geq 1)$ are the eigenvalues of $\mathcal{V}$ and $(\xi_m : m \geq 1)$ are the corresponding eigenfunctions such that $\|\xi_m\|^2 = \int_0^1 \xi_m(t)\xi_m^{\bullet}(t) = \int_0^1 \xi_m(t)\xi_m(t) = 1$. If $\mathcal{V}$ has full rank, then the sequence $(\xi_m : m \geq 1)$ forms an orthonormal basis of $L^2([0,1])$ and the process X can be reconstructed as follows:

$$X = \sum_{m=1}^{\infty} \langle X, \xi_m \rangle \xi_m, \tag{4.3}$$

where $\langle X, \xi_m \rangle$ defines the orthogonal projection of X onto the subspace of all linear combinations of the orthonormal basis $(\xi_m : m \geq 1)$. This representation is known as the Karhunen-Loève expansion of X. The eigenfunctions ξ_m are the static or traditional FPCs previously defined in Section 3.2 and the inner products $\langle X, \xi_m \rangle$ are the corresponding scores S_m obtained by $\int_0^1 X(t)\xi_m(t)dt$ assuming a zero mean process (see Equation 3.5); and hence Equation 4.3 is an equivalent representation to Equation 3.6. As mentioned before, this orthonormal basis $(\xi_m : m \geq 1)$ is optimal in representing X in the sense that if there exists another orthonormal basis of L^2_H, say $(w_m : m \geq 1)$, then:

$$\mathbb{E}\Big[\|X - \sum_{m=1}^{q} \langle X, \xi_m \rangle \xi_m\|^2\Big] \leq \mathbb{E}\Big[\|X - \sum_{m=1}^{q} \langle X, w_m \rangle w_m\|^2\Big], \quad \forall q \geq 1. \tag{4.4}$$

This property also indicates that a finite number of FPCs, say q, can be used to approximate the function X with a minimum loss of instantaneous information. The major concern regarding these traditional FPCs is their static nature which does not take into account the serial time dependence between the curves x_i. Therefore, Hormann et al. (2014) proposed a dynamic version of FPCs based on a frequency domain approach, in which the approximation of X involves lagged observations and is based on the whole family of covariance $(\mathcal{V}_h : h \in \mathbb{Z})$ and not only $\mathcal{V}_0$. This approach requires first introducing the spectral density operator, which contains the full information on the family of operators $(\mathcal{V}_h : h \in \mathbb{Z})$.

4.1.2 Spectral Density Operator

For every signal over time, $x(t)$, there exists a corresponding frequency domain function, $\mathbb{F}(\theta)$, describing the frequency content generating the signal $x(t)$. The time domain representation is viewed as a regression of the present on the past values, while the frequency domain representation is regarded as a regression of the present on the periodic sines and cosines (Shumway and Stoffer, 2011). The Fourier transform specifically studies this relationship between the time domain and the corresponding frequency domain functions. The Fourier transform of a signal $x(t)$ that belongs to the space $L^2([0,1])$ is defined as the decomposition of that signal into its frequency components θ as follows:

$$\mathbb{F}(\theta) = \frac{1}{\sqrt{2\pi}} \int_0^1 x(t)\exp(-\mathrm{i}\theta t)dt, \qquad \theta \in [-\pi, \pi]; \tag{4.5}$$

whereas the inverse Fourier transform from the frequency domain to the time domain used to reconstruct the original signal is given by:

$$x(t) = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} \mathbb{F}(\theta)\exp(\mathrm{i}\theta t)d\theta. \tag{4.6}$$

Equations 4.5 and 4.6 are known as the Fourier transform pair[1].

The *spectral density* of a stochastic process, denoted by F_θ, is the square of the magnitude of the Fourier transform of the signal $|\mathbb{F}(\theta)|^2$. By definition, the auto-covariance function V of a stationary process depends only on the time lag h and not the time t, i.e. $V(t, t-h) = V_h$. Thus, the spectral density is defined as the Fourier transform of the auto-covariance function:

$$F_\theta = \frac{1}{2\pi} \int_0^1 V_h \exp(-\mathrm{i}\theta h)dh, \qquad \theta \in [-\pi, \pi],$$

[1] Other authors define the Fourier transform pair, for $t \in [0,1]$ and $\theta \in [-\pi, \pi]$ in one of the following ways:

1. $\mathbb{F}(\theta) = \frac{1}{2\pi}\int_0^1 x(t)\exp(-\mathrm{i}\theta t)dt$ and $x(t) = \int_{-\pi}^{\pi} \mathbb{F}(\theta)\exp(\mathrm{i}\theta t)d\theta$
2. $\mathbb{F}(\theta) = \int_0^1 x(t)\exp(-\mathrm{i}\theta t)dt$ and $x(t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathbb{F}(\theta)\exp(\mathrm{i}\theta t)d\theta$

where the auto-covariance function V_h can be regenerated by the inverse Fourier transform of the spectral density as follows:

$$V_h = \int_{-\pi}^{\pi} F_\theta \exp(\mathrm{i}\theta h) d\theta.$$

In practice, the signal is only observed at a set of discrete time points $x_t : t = 1, 2, \dots, N$ and its Fourier transform is obtained by:

$$\mathbb{F}^N(\theta) = \frac{1}{\sqrt{2\pi N}} \sum_{t=1}^{N} x_t \exp(-\mathrm{i}\theta t), \qquad \theta \in [-\pi, \pi],$$

which is exactly the same as the DFT previously defined by Equation 1.8 in Chapter 1, but based on an angular frequency of $2\pi\theta$. Accordingly, the sampled version of the spectral density, known as the *periodogram*, is defined as:

$$\begin{aligned} F_\theta^N &= \left|\mathbb{F}^N(\theta)\right|^2 = \frac{1}{2\pi N}\left|\sum_{t=1}^{N} x_t \exp(-\mathrm{i}\theta t)\right|^2, \\ &= \frac{1}{2\pi}\sum_{h\in\mathbb{Z}} V_h^N \exp(-\mathrm{i}\theta h), \qquad \theta \in [-\pi, \pi], \end{aligned}$$

where V_h^N denotes the sampled auto-covariance obtained by $N^{-1}\sum_{t=1}^{N-h}(x_t - \bar{x})(x_{t+h} - \bar{x})$; and $\bar{x}$ is the sample mean of the series. For more details, see Shumway and Stoffer (2011).

By analogy, let $X = (X(t) : t \in [0, 1])$ be a zero mean stationary functional process, then the kernel of its *spectral density operator* $\mathcal{F}_\theta^X$ at frequency θ is given by:

$$F_\theta^X(t, s) = \frac{1}{2\pi}\sum_{h\in\mathbb{Z}} V_h(t, s)\exp(-\mathrm{i}h\theta), \qquad \theta \in [-\pi, \pi].$$

To ensure the convergence of the series defining F_θ^X, the following condition is imposed:

$$\sum_{h\in\mathbb{Z}}\left\{\int_0^1\int_0^1 |V_h(t, s)|^2 dtds\right\}^{1/2} < \infty,$$

which can be rewritten as:

$$\sum_{h\in\mathbb{Z}} \|\mathcal{V}_h\|_{\mathcal{S}} < \infty,$$

such that $\|.\|_{\mathcal{S}}$ denotes a Hilbert-Schmidt norm[2]. Thus, for every frequency θ, the

[2] For a linear operator A between 2 Hilbert spaces H and H', $A : H \to H'$, the Hilbert Schmidt norm of A denoted by $\|A\|_{\mathcal{S}}$ is defined as:

$$\|A\|_{\mathcal{S}} = \Big(\sum_{j=1}^{\infty} \|A(e_j)\|^2\Big)^{1/2},$$

where $\{e_j\}_{j=1}^{\infty}$ is any orthonormal basis of H. This definition is independent of the choice of the orthonormal basis. The linear operator A such that $\|A\|_{\mathcal{S}} < \infty$ is called a Hilbert-Schmidt operator.

operator $\mathcal{F}_\theta^X$ is a non-negative, self-adjoint Hilbert-Schmidt operator. Hence, analogous to the covariance operator decomposition in Equation 4.2, $\mathcal{F}_\theta^X$ admits the following dynamic Karhunen-Loève expansion:

$$\mathcal{F}_\theta^X = \sum_{m=1}^{\infty} \gamma_m(\theta)\langle x, \varphi_m(\theta)\rangle \varphi_m(\theta) \qquad \forall\theta, \tag{4.7}$$

where $\gamma_m(\theta)$ and $\varphi_m(\theta)$ are called the dynamic eigenvalues and eigenfunctions, respectively such that $\gamma_1(\theta) \geq \gamma_2(\theta) \geq \ldots \geq 0$ for all $\theta \in [-\pi, \pi]$ and $\|\varphi_m(\theta)\| = 1$ for all m and $\theta \in [-\pi, \pi]$.

4.1.3 Construction Theory of Dynamic FPCs

The spectral density operator is used to obtain the functional filters which are the essential building blocks of the dynamic FPCs. The definition of a filter is already given in Section 1.5.2, whereas a *functional filter* is a sequence of linear operators $\mathbf{\Phi} = (\Phi_l : l \in \mathbb{Z})$ between the Hilbert space of square integrable functions $L^2([0,1])$ and the q dimensional real coordinate space $\mathbb{R}^q$. Accordingly, the filtered series Y_i has the form $Y_i = \sum_{l\in\mathbb{Z}} \Phi_l(X_{i-l})$, where $\Phi_l = (\phi_{1l}, \ldots, \phi_{ql} \in L^2([0,1]))$.

Now, let the functional filters $(\phi_{ml} : 1 \leq m \leq q$ and $l \in \mathbb{Z})$ for which the sequences $(\sum_{l=-L}^{L} \phi_{ml}(t)\exp(\mathrm{i}l\theta) : L \geq 1)$ converge in $L^2([0,1] \times [-\pi, \pi])$ be chosen such that:

$$\lim_{L\to\infty} \int_{-\pi}^{\pi} \int_0^1 \left\{ \sum_{l=-L}^{L} \phi_{ml}(t)\exp(\mathrm{i}l\theta) - \varphi_m(t|\theta) \right\}^2 dt d\theta = 0,$$

where the function $\varphi_m(t|\theta)$ is jointly measurable in t and θ and standardized to unit length:

$$\int_{-\pi}^{\pi} \int_0^1 \varphi_m^2(t|\theta) dt d\theta = 2\pi.$$

Thus, we can write $\varphi_m(t|\theta) = \sum_{l\in\mathbb{Z}} \phi_{ml}(t)\exp(\mathrm{i}l\theta)$ or shortly $\varphi_m(\theta) = \sum_{l\in\mathbb{Z}} \phi_{ml}\exp(\mathrm{i}l\theta)$, where the functional filters ϕ_{ml} are defined as the Fourier expansion of the eigenfunctions φ_m as follows:

$$\phi_{ml}(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \varphi_m(t|\theta)\exp(-\mathrm{i}l\theta) d\theta \qquad \forall t \in [0,1]. \tag{4.8}$$

Following from this and assuming that X_i is a zero-mean stationary process with values in L^2, the series $\sum_{l\in\mathbb{Z}} \Phi_l(X_{i-l})$ used to define the dynamic FPC scores converges in mean square to a limit Y_i, such that the m^{th} dynamic FPC score of X_i is given by:

$$Y_{mi} = \sum_{l\in\mathbb{Z}} \langle X_{i-l}, \phi_{ml}\rangle, \qquad i \in \mathbb{Z} \quad \text{and} \quad m \geq 1. \tag{4.9}$$

Note that if $\mathbb{E}[X_i] = \mu \neq 0$, the dynamic FPC scores Y_{mi} are obtained by replacing X_i with $X_i - \mu$ in Equation 4.9. The functional filters $(\phi_{ml} : l \in \mathbb{Z})$ are called the m^{th} dynamic FPC filters and they play the role of the eigenfunctions of the static FPCs. For a finite number of dynamic FPCs, say q, the spectral density matrix of the filtered series Y_i has the following form:

$$\mathcal{F}_\theta^Y = \begin{pmatrix} \langle \mathcal{F}_\theta^X\{\varphi_1(\theta)\}, \varphi_1(\theta)\rangle & \dots & \langle \mathcal{F}_\theta^X\{\varphi_q(\theta)\}, \varphi_1(\theta)\rangle \\ \vdots & \ddots & \vdots \\ \langle \mathcal{F}_\theta^X\{\varphi_1(\theta)\}, \varphi_q(\theta)\rangle & \dots & \langle \mathcal{F}_\theta^X\{\varphi_q(\theta)\}, \varphi_q(\theta)\rangle \end{pmatrix} \qquad \forall\theta,$$

which, according to the dynamic Karhunen-Loève expansion given by Equation 4.7 where $\langle \varphi_m, \varphi_m' \rangle = 0$ for all $m \neq m'$ and $\|\varphi_m\| = 1$ for all m, reduces to:

$$\mathcal{F}_\theta^Y = \mathrm{diag}\{\gamma_1(\theta), \dots, \gamma_q(\theta)\} \qquad \forall\theta.$$

This implies that the filtered processes Y_i are uncorrelated at all leads and lags, with diagonal auto-covariance matrices at all lags h. That is, $\mathrm{COV}(Y_{mi}, Y_{m'j}) = 0$ for all i, j and $m \neq m'$.

The resulting dynamic FPC scores (Y_{mi}) have a number of desirable properties. First, they are real since the eigenfunctions $\varphi_m(\theta)$ are Hermitian. Second, the dynamic FPC scores coincide with the static FPC scores if $\mathcal{V}_h = 0$ for $h \neq 0$. Third, the dynamic FPC scores Y_{mi} and $Y_{m'j}$ are uncorrelated for all i and j and $m \neq m'$. Fourth, the long-run covariance matrix of the dynamic FPC score vector process (Y_i) is obtained by $\lim_{N\to\infty} \frac{1}{N}\mathrm{var}(Y_1 + \ldots + Y_N) = 2\pi\mathrm{diag}\{\gamma_1(0), \dots, \gamma_q(0)\}$.

According to the dynamic analogue of the static Karhunen-Loève expansion, the original process $(X_i(t) : i \in \mathbb{Z}, t \in [0,1])$ can be recovered from the scores $(Y_{mi} : i \in \mathbb{Z}, m \geq 1)$ as follows:

$$X_i(t) = \sum_{m=1}^{\infty} X_{mi}(t) \qquad\qquad X_{mi}(t) = \sum_{l\in\mathbb{Z}} Y_{m,i+l}\phi_{ml}(t). \tag{4.10}$$

Similar to the traditional FPCs, the original process $X_i(t)$ can be approximated using the first q-dimensional reconstructions $\sum_{m=1}^{q} X_{mi}(t)$, $q \geq 1$, which involve only the q time series $(Y_{mi} : i \in \mathbb{Z}, 1 \leq m \leq q)$. The reconstruction based on the functional filters $(\phi_{ml} : 1 \leq m \leq q)$ proves to be optimal in the sense that if ϕ_{ml} in expressions 4.9 and 4.10 are replaced by alternative sequences, say ω_{ml} and τ_{ml}, such that $\tilde{Y}_{mi} = \sum_{l\in\mathbb{Z}} \langle X_{i-l}, \omega_{ml}\rangle$ and $\tilde{X}_{mi} = \sum_{l\in\mathbb{Z}} \tilde{Y}_{m,i+l}\tau_{ml}$, then:

$$\mathbb{E}\Big[\|X_i - \sum_{m=1}^{q} X_{mi}\|^2\Big] = \sum_{m>q} \int_{-\pi}^{\pi} \gamma_m(\theta) d\theta \leq \mathbb{E}\Big[\|X_i - \sum_{m=1}^{q} \tilde{X}_{mi}\|^2\Big] \qquad \forall q \geq 1.$$

This inequality is regarded as the dynamic version of the one given by Equation 4.4.

4.1.4 Practical Implementation

In real life applications, the data are recorded discretely and the curve $x(t)$ is observed at a finite number of points $0 \leq t_1 < t_2 < \ldots < t_T \leq 1$. In high-frequency data, T is often a large value. As previously mentioned, the first step in FDA is to construct the smooth functional curves from the discrete observations and represent each curve in terms of a finite number of basis functions $(\psi_k : 1 \leq k \leq p)$, such that $x(t) = \sum_{k=1}^{p} a_k \psi_k(t)$. Equivalently, let x belong to the span $H_p = \bar{\text{sp}}(\psi_k : 1 \leq k \leq p)$ of $\psi_1, \ldots, \psi_p$, then x can be expressed in vector-matrix notation as $\boldsymbol{\psi}^\top \mathbf{a}$, where $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_p)^\top$ and $\mathbf{a} = (a_1, \ldots, a_p)^\top$. The basis functions $\psi_1, \ldots, \psi_p$ are assumed to be linearly independent. Thus, any statement about the curve x can be expressed equivalently in terms of the coefficients $\mathbf{a}$. That is, if $A : H_p \to H_p$ is a linear operator, then for $x \in H_p$:

$$A(x) = \sum_{k=1}^{p} a_k A(\psi_k) = \sum_{k=1}^{p} \sum_{k'=1}^{p} a_k \langle A(\psi_k), \psi_{k'} \rangle \psi_{k'} = \boldsymbol{\psi}^\top \mathfrak{A} \mathbf{a},$$

where $\mathfrak{A}$ is the corresponding $p \times p$ matrix of A obtained such that $\mathfrak{A}^\top = (\langle A(\psi_k), \psi_k' \rangle : 1 \leq k, k' \leq p)$. If A and B are linear operators on H_p with corresponding matrices $\mathfrak{A}$ and $\mathfrak{B}$, respectively; and $\otimes$ denotes a tensor product operator. Then:

I. For any $\alpha, \beta \in \mathbb{C}$, the corresponding matrix of $\alpha A + \beta B$ is $\alpha \mathfrak{A} + \beta \mathfrak{B}$;

II. Letting $A = \sum_{k=1}^{p} \sum_{k'=1}^{p} g_{kk'} \psi_k \otimes \psi_{k'}$ such that $G = (g_{kk'} : 1 \leq k, k' \leq p)$ is a $p \times p$ matrix with entry $g_{kk'} \in \mathbb{C}$ in row k and column k' and $\mathbf{W} = (\langle \psi_k, \psi_{k'} \rangle : 1 \leq k, k' \leq p)$ is a $p \times p$ matrix with entry $\langle \psi_k, \psi_{k'} \rangle$ in row k and column k', the corresponding matrix of A is $\mathfrak{A} = G \mathbf{W}^\top$.

The spectral density operator $\mathcal{F}_\theta^X$ of the process X clearly depends on the auto-covariance operators $\mathcal{V}_h^X$ at all lags h. Thus, to obtain the corresponding matrix of the spectral density operator $\mathcal{F}_\theta^X$, first express the zero-mean stochastic process X_i as $\sum_{k=1}^{p} a_{ik} \psi_k = \boldsymbol{\psi}^\top \mathbf{a}_i$, then express $\mathcal{V}_h^X$ as follows:

$$\mathcal{V}_h^X = \mathbb{E}[X_h \otimes X_0] = \sum_{k=1}^{p} \sum_{k'=1}^{p} \mathbb{E}[a_{hk} a_{0k'} \psi_k \otimes \psi_{k'}].$$

Now, let $V_h^{\mathbf{a}} = \mathbb{E}[\mathbf{a}_h \mathbf{a}_0^\top]$, then $V_h^{\mathbf{a}} \mathbf{W}^\top$ is the corresponding matrix of $\mathcal{V}_h^X$ according to II, such that $\mathbf{W} = \mathbf{W}^\top = \int \boldsymbol{\psi}(t) \boldsymbol{\psi}(t)^\top dt$ is the symmetric matrix of inner products between the basis functions; and following the linearity property in I, the corresponding matrix of the spectral density operator $\mathcal{F}_\theta^X$ is:

$$\mathfrak{F}_\theta^X = \frac{1}{2\pi}\Big\{\sum_{h\in\mathbb{Z}} V_h^{\mathbf{a}}\exp(-\mathrm{i}h\theta)\Big\}\mathbf{W}^\top. \tag{4.11}$$

Suppose that $\gamma_m(\theta)$ is the m^{th} largest eigenvalue of $\mathfrak{F}_\theta^X$ and $\varphi_m(\theta)$ is the corresponding eigenvector. Then, $\gamma_m(\theta)$ is also an eigenvalue for the operator $\mathcal{F}_\theta^X$ and $\boldsymbol{\psi}^\top\varphi_m(\theta)$ is its corresponding eigenfunction. In case of non-orthogonal basis functions ($\mathbf{W} \neq \mathbf{I}$), $\mathfrak{F}_\theta^X$ is not necessarily Hermitian and hence the corresponding eigenvalues might not be real. This case has not been discussed by Hormann et al. (2014), as they were focusing on the use of orthogonal basis functions - like the Fourier basis - to estimate the functional data. However, in practice, functional data can be approximated using any type of basis functions, e.g. B-splines basis functions. One example is the daily curves of EpCO$_2$ which are approximated here using penalized B-splines basis functions. Such basis functions are not often orthogonal to allow for other desirable features, see Chapter 2.

To get real eigenvalues $\gamma_m(\theta)$ and corresponding Hermitian eigenvectors $\varphi_m(\theta)$, I suggest defining $\varphi_m^\star(\theta) = \mathbf{W}^{1/2}\varphi_m(\theta)$ as the eigenvectors of the following Hermitian matrix:

$$\frac{1}{2\pi}\mathbf{W}^{1/2}\Big\{\sum_{h\in\mathbb{Z}} V_h^{\mathbf{a}}\exp(-\mathrm{i}h\theta)\Big\}\mathbf{W}^{1/2\top},$$

and obtain the eigenvectors $\varphi_m(\theta)$ by $(\mathbf{W}^{1/2})^{-1}\varphi_m^\star(\theta)$. $\mathbf{W}^{1/2}$ is the Cholesky decomposition matrix of $\mathbf{W}$ such that $\mathbf{W} = \mathbf{W}^{1/2\top}\mathbf{W}^{1/2}$. Subsequently, the dynamic FPCs are obtained via the Fourier expansion of those eigenfunctions and the functional filters ϕ_{ml} are calculated as:

$$\phi_{ml} = \frac{\boldsymbol{\psi}^\top}{2\pi}\int_{-\pi}^{\pi}\varphi_m(s)\exp(-\mathrm{i}ls)ds = \boldsymbol{\psi}^\top\tilde{\boldsymbol{\phi}}_{ml},$$

where $\tilde{\boldsymbol{\phi}}_{ml}$ are considered as the basis coefficients of ϕ_{ml} and are known as the filter coefficients. Then, the dynamic FPC scores y_{mi} are obtained as:

$$Y_{mi} = \sum_{l\in\mathbb{Z}}\int_0^1 \mathbf{a}_{i-l}^\top\boldsymbol{\psi}(t)\boldsymbol{\psi}^\top(t)\tilde{\boldsymbol{\phi}}_{ml}dt = \sum_{l\in\mathbb{Z}}\mathbf{a}_{i-l}^\top\mathbf{W}\tilde{\boldsymbol{\phi}}_{ml}. \tag{4.12}$$

In practice and according to Equation 4.11, the spectral density matrix $\mathfrak{F}_\theta^X$ is replaced by the spectral density matrix of the coefficient sequence $(\mathbf{a}_i)$ given by:

$$\mathfrak{F}_\theta^{\mathbf{a}} = \frac{1}{2\pi}\sum_{h\in\mathbb{Z}} V_h^{\mathbf{a}}\exp(-\mathrm{i}h\theta).$$

Using multivariate techniques and assuming that the data are centered, the auto-covariances $V_h^{\mathbf{a}}$ are first estimated for $|h| < N$ as follows:

$$\begin{aligned} \hat{V}_h^{\mathbf{a}} &= \frac{1}{N} \sum_{i=1}^{N-h} \mathbf{a}_i \mathbf{a}_{i+h}^\top, && h \geq 0 \\ \hat{V}_h^{\mathbf{a}} &= (\hat{V}_{-h}^{\mathbf{a}})^\top, && h < 0. \end{aligned}$$

The next step is to estimate the spectral density of the basis coefficients sequence. There are numerous methods for estimating the spectral density. One method involves dividing the data sequence into a number of segments of the same length and compute the periodogram for each segment, then get the average of periodograms across all segments (Bartlett, 1950). As the number of segments increases, the variance of the estimate decreases; but the number of points within each segment decreases simultaneously, which in turn decreases the spectral resolution. To reduce the variance of the spectral density estimate, Blackman and Tukey (1958) has proposed another approach to smooth the periodogram obtained from the entire N point data sequence using a window function. The smoothed periodogram is obtained by convolving the periodogram with a spectral window function, which is equivalent to the Fourier transform of the product of the auto-covariance function and a lag window function in the time domain. Welch (1967) has also proposed a modified version of the Bartlett's method to estimate the spectral density more efficiently. According to Welch (1967), the data sequence is divided into a number of segments of equal length then a smooth version of the periodogram is computed within each segment separately and finally the spectral density estimate is calculated as the average of smoothed peridogorams across all segments. In this thesis, the spectral density is always estimated using the second approach proposed by Blackman and Tukey (1958). Thus, the spectral density $\mathfrak{F}_\theta^{\mathbf{a}}$ estimated by using a lag window estimator is given by:

$$\hat{\mathfrak{F}}_\theta^{\mathbf{a}} = \frac{1}{2\pi} \sum_{|h| \leq Q} w\left(\frac{h}{Q}\right) \hat{V}_h^{\mathbf{a}} \exp(-\mathrm{i}h\theta),$$

where w is an appropriate lag window function regarded as the inverse Fourier transform of a spectral window function, $Q = Q_N \to \infty$ and $Q_N/N \to 0$. There are various choices for the weighting function w and the tuning parameter Q_N. The most common choice for w is the traditional Bartlett kernel $w(\frac{h}{Q}) = 1 - |\frac{h}{Q}|$ with bandwidth $Q = \lfloor N^{1/2} \rfloor$. The bandwidth Q determines the extent of the data to be observed through the window.

After estimating the spectral density matrix $\mathfrak{F}_\theta^{\mathbf{a}}$ of the basis coefficients, set $\hat{\mathfrak{F}}_\theta^X = \hat{\mathfrak{F}}_\theta^{\mathbf{a}} \mathbf{W}^\top$ and compute the estimates of the eigenvalues $\gamma_m(\theta)$ and eigenvectors $\varphi_m(\theta)$, denoted by $\hat{\gamma}_m(\theta)$ and $\hat{\varphi}_m(\theta)$, respectively. In case of non-orthogonal basis functions, $\hat{\mathfrak{F}}_\theta^{\mathbf{a}} \mathbf{W}^\top$ is not necessarily a Hermitian matrix. Therefore, I suggest a new re-parameterization where I define $\varphi^{\star}_m(\theta) = \mathbf{W}^{1/2}\varphi_m(\theta)$ and use the Hermitian matrix eigenvalue routine to obtain the required (real) eigenvalues $\hat{\gamma}_m(\theta)$ and (Hermitian) eigenvectors $\hat{\varphi}^{\star}_m(\theta)$ of $\mathbf{W}^{1/2}\hat{\mathfrak{F}}^{\mathfrak{a}}_{\theta}\mathbf{W}^{1/2\top}$, then compute $\hat{\varphi}_m(\theta) = (\mathbf{W}^{1/2})^{-1}\hat{\varphi}^{\star}_m(\theta)$. Afterwards, the functional filters are estimated by:

$$\hat{\phi}_{ml} = \frac{\boldsymbol{\psi}^{\top}}{2\pi}\int_{-\pi}^{\pi} \hat{\varphi}_m(\theta)\exp(-\mathrm{i}l\theta)d\theta.$$

$\hat{\varphi}_m(\theta)$ does not have an analytic form, and therefore $\hat{\phi}_{ml}$ are obtained by numerical integration as follows:

$$\hat{\phi}_{ml} = \frac{\boldsymbol{\psi}^{\top}}{2\pi(2N_\theta+1)}\sum_{j=-N_\theta}^{N_\theta} \hat{\varphi}_m\Big(\frac{\pi j}{N_\theta}\Big)\exp\Big(-\mathrm{i}l\frac{\pi j}{N_\theta}\Big) = \boldsymbol{\psi}^{\top}\hat{\hat{\boldsymbol{\phi}}}_{ml}, \qquad N_\theta >> 1. \tag{4.13}$$

The approximation enhances as N_θ gets larger; but the choice of N_θ simultaneously depends on the available computing power.

After estimating the functional filters $\hat{\phi}_{ml}$, the dynamic FPC scores Y_{mi} are estimated by substituting $\hat{\phi}_{ml}$ in Equation 4.12 and replacing the infinite sum with a rolling window:

$$\hat{Y}_{mi} = \sum_{l=-L}^{L} a^{T}_{i-l}\mathbf{W}\hat{\hat{\boldsymbol{\phi}}}_{ml} \qquad \text{for } i \in \{L+1,\dots,N-L\}, \tag{4.14}$$

while for $1 \leq i \leq L$ or $N-L+1 \leq i \leq N$, set $X_{-L+1} = \ldots = X_0 = X_{N+1} = \ldots = X_{N+L} = \mathbb{E}[X_1] = 0$. Note that $\mathbb{E}[X_1] = \mu \neq 0$ if the functions are not centered. Padding the ends of the series with the mean value creates bias on the boundary of the observation period. These boundary problems arise in the filtering process as in DWT, earlier discussed in Chapter 1. Regarding the choice of L, it is natural to choose L such that $\sum_{-L\leq l\leq L}\|\hat{\phi}_{ml}\|^2 \geq 1-\epsilon$, for some small threshold ϵ, since we have $\sum_{l\in\mathbb{Z}}\|\hat{\phi}_{ml}\|^2 = 1$. As $L = L(N) \to \infty$, $\hat{Y}_{mi}$ converges in probability to Y_{mi} as $N \to \infty$, but the number of biased scores on the boundary simultaneously increases. Thus, the choice of L is a trade-off between getting consistent estimators for the scores Y_{mi} as $\epsilon \to 0$ and getting more biased scores at the boundary of the series as $L \to \infty$.

Finally, the original curves can be approximated using the q-term dynamic Karhunen Loève expansion as follows:

$$\hat{X}_i = \sum_{m=1}^{q}\sum_{l=-L}^{L}\hat{Y}_{m,i+l}\hat{\phi}_{ml}, \qquad \hat{Y}_{mi} = 0,\ \ i \in \{-L+1,\dots,0\}\cup\{N+1,\dots,N+L\}. \tag{4.15}$$

Note that if the functional data are not centered from the beginning then $\hat{Y}_{mi}$ is set equal to $\bar{\hat{Y}}_{m.}$ for $i \in \{-L+1,\dots,0\}\cup\{N+1,\dots,N+L\}$.

As in any standard PCA, the appropriate number of components, q, is chosen based on the proportion of variance explained by those components. In dynamic FPCA, the proportion of variance explained by the first q dynamic FPCs is calculated by:

$$\mathrm{PV}_{\mathrm{dyn}}(q) = \frac{\pi}{N_\theta} \sum_{m=1}^{q} \sum_{j=-N_\theta}^{N_\theta} \hat{\gamma}_m\Big(\frac{\pi j}{N_\theta}\Big) \Big/ \frac{1}{N} \sum_{i=1}^{N} \|X_i(t)\|^2. \tag{4.16}$$

The quantity $1-\mathrm{PV}_{\mathrm{dyn}}(q)$ can then be used as a measure of the information loss resulting from approximating the curves in a reduced dimension q, or alternatively one can use the normalized mean squared error given by:

$$\mathrm{NMSE}(q, L) = \sum_{i=2L+1}^{N-2L} \|X_i(t) - \hat{X}_i(t)\|^2 \Big/ \sum_{i=2L+1}^{N-2L} \|X_i(t)\|^2. \tag{4.17}$$

In general, the 2 quantities, $1-\mathrm{PV}_{\mathrm{dyn}}(q)$ and $\mathrm{NMSE}(q)$, do not coincide but converge to the same limit. Although the NMSE seems to be less practical as it depends on L, the NMSE is a more honest estimate that is highly recommended since the approximated curves $\hat{X}_i$ are based on the choice of both q and L. However, in practice, the NMSE is calculated by ignoring the effect of L as follows:

$$\mathrm{NMSE}(q) = \sum_{i=1}^{N} \|X_i(t) - \hat{X}_i(t)\|^2 \Big/ \sum_{i=1}^{N} \|X_i(t)\|^2. \tag{4.18}$$

In brief, the dynamic FPCs are essentially based on the spectral decomposition of the spectral density operator, which contains full information on the family of covariance operators $(\mathcal{V}_h : h \in \mathbb{Z})$ not only $\mathcal{V}_0$. The initial step in obtaining the dynamic FPCs is to express each curve in terms of its basis coefficients $\mathbf{a}$ then estimate the auto-covariance of the basis coefficients sequence, $\hat{V}_h^{\mathbf{a}}$, for all lags h. Next, estimate the spectral density matrix of the basis coefficients sequence, $\hat{\mathfrak{F}}_\theta^{\mathbf{a}}$, and obtain the corresponding spectral density matrix of the process (X_i), $\hat{\mathfrak{F}}_\theta^{X}$, by setting $\hat{\mathfrak{F}}_\theta^{X} = \hat{\mathfrak{F}}_\theta^{\mathbf{a}}\mathbf{W}^\top$. After computing the dynamic eigenvalues $\hat{\gamma}_m(\theta)$ and eigenfunctions $\hat{\varphi}_m(\theta)$ of $\hat{\mathfrak{F}}_\theta^{X}$, the functional filters $\hat{\phi}_{ml}$, which are the building blocks of the dynamic FPCs, are obtained via the Fourier expansion of those eigenfunctions and finally the dynamic FPC scores are estimated. Those dynamic FPC scores can then be used to approximate the original curves using a more compact representation.

4.2 Application of Dynamic FPCA to the EpCO$_2$ Data

As mentioned before, a weakness of the analysis presented in Chapter 3 is that it ignores the temporal dependence between the curves. Accordingly, the obtained FPCs might not be the appropriate dimension reduction representation for the data. In addition, the clustering procedure of the curves does not benefit from the information provided by nearby observations and hence the resulting classification is not properly smooth. This chapter focuses on (i) the identification of the main modes of variability in the the daily curves of EpCO$_2$ and (ii) the classification of these daily curves into clusters such that the days within each cluster share similar characteristics of the EpCO$_2$ daily patterns, taking into account the temporal dependence structure in the data. To achieve these objectives, dynamic FPCA is employed.

As in Chapter 3, the global trend and seasonal effect are first removed from the discrete data and the DD series is obtained. Second, the 15-min. regularly spaced DD data within each day are used to estimate a continuous smooth function for each day, using penalized cubic B-splines with a knot every 2 hours and a smoothing parameter equal to 1. Note that in order to compute the dynamic FPCs, the discrete data of the 17 complete missing days in July 2004 have been first imputed using linear interpolation as in Chapter 1, to preserve the time order and lag between the curves across the whole time series. As the discrete data have been already centered at the global smooth mean $\hat{\mu}_t^G$, the daily curves of the DD series need not to be centered at their empirical mean curve $\hat{\mu}(t)$ before applying the dynamic FPCA. This is mainly because we are interested in both the average pattern of the daily curves of EpCO$_2$ and the deviations from this mean pattern. After estimating the B-splines basis coefficients of the daily smooth curves, the covariance matrix of the basis coefficients sequence is computed for each lag h. Next, the traditional Bartlett kernel $w(\frac{h}{Q}) = 1 - |\frac{h}{Q}|$, with bandwidth $Q = \lfloor \sqrt{N} \rfloor = 32$, is used to obtain an estimator for the spectral density operator, $\hat{\mathfrak{F}}_\theta^{\mathbf{a}}$, as explained in Section 4.1.4. Subsequently, the dynamic eigenfunctions are obtained and the corresponding functional filters, $\hat{\phi}_{ml}$, are estimated.

According to the scree plot in Figure 4.1, displaying the number of dynamic FPCs q to be retained against the corresponding proportion of variance explained computed by Equation 4.18, it has been decided to retain the first 2 dynamic FPCs explaining about 88% using Equation 4.18 (and 93.5% using Equation 4.16) . The first and second dynamic FPCs account for 74% and 14% of the total variability, respectively. For the choice of L, Figure 4.2 displays the sum of squared norm $\sum_{-L}^{L} \|\hat{\phi}_{ml}\|^2$ for $m = 1, 2$ versus versus a range of lags L. The figure assures that for $L > 30$, additional lags seem to have trivial contribution to the norm sum of squares. In particular, $\sum_{l=-30}^{30} \|\hat{\phi}_{1l}\|^2 \approx 0.993$ and $\sum_{l=-30}^{30} \|\hat{\phi}_{2l}\|^2 \approx 0.993$, which justify the choice of L to be 30 for the calculation

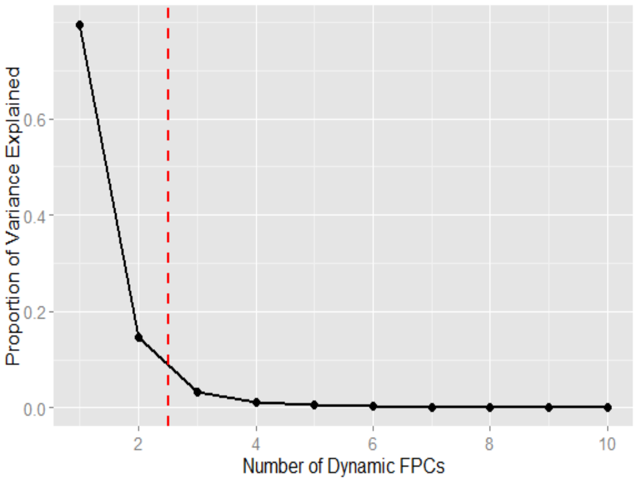


FIGURE 4.1: Scree plot of the number of dynamic FPCs versus the proportion of variance accounted for.

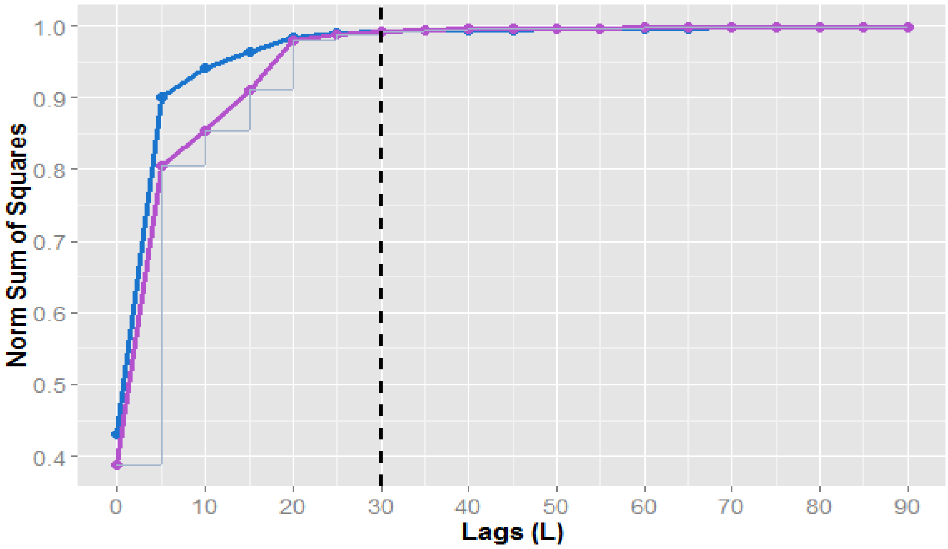


FIGURE 4.2: Plot of the sum of squared norm of the functional filters $\hat{\phi}_{1l}$ (blue) and $\hat{\phi}_{2l}$ (purple), $l = -L, \dots, L$ against the lag L.

of the dynamic FPC scores. The estimated functional filters $\hat{\phi}_{1l}(t)$ and $\hat{\phi}_{2l}(t)$, $l = -30, \dots, 30$ are plotted in Figure 5.16, from which they appear to fade down and converge to zero quite rapidly. The 3 central filters of the first 2 dynamic FPCs, $\hat{\phi}_{1l}(t)$ and $\hat{\phi}_{2l}(t)$ for $l = -1, 0, 1$, which tend to be the most influential filters are displayed solely in Figure 4.4.

After obtaining the functional filters, the dynamic FPC score sequences $(\hat{y}_{1i}, 1 \leq i \leq 1095)$ and $(\hat{y}_{2i}, 1 \leq i \leq 1095)$ are estimated; 1095 is the number of daily curves in 3 hydrological years (365×3). The score sequences of the first 2 dynamic FPCs are displayed in the bottom panels of Figure 4.5(a-b). To calculate the dynamic FPC scores

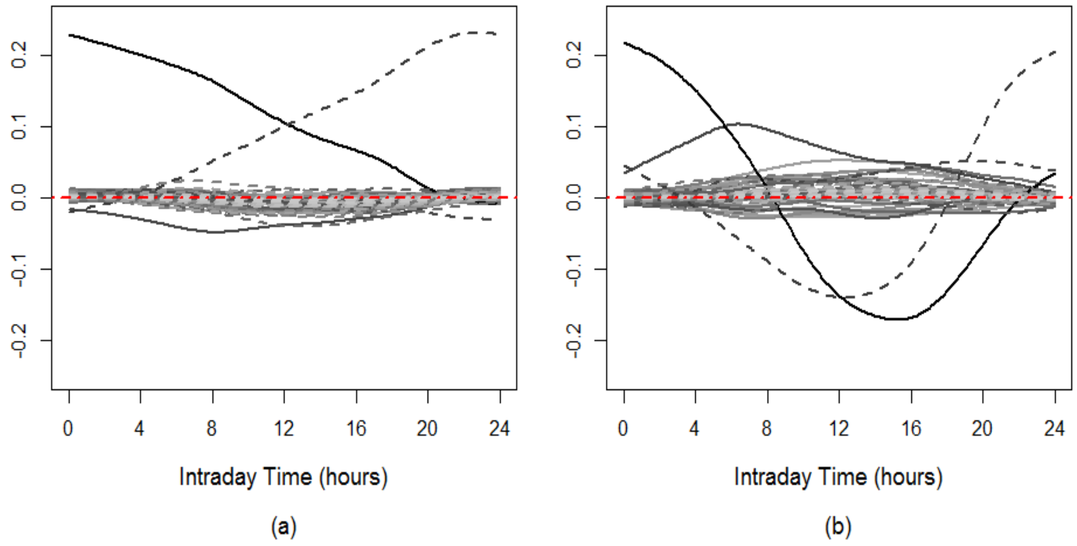


FIGURE 4.3: The functional filters $\hat{\phi}_{ml}(t)$ (or shortly $\hat{\phi}_{ml}, l = -30, \ldots, 30$) of the (a) first and (b) second dynamic FPCs. For $m = 1, 2$: the filters $\hat{\phi}_{ml}(t)$ for $1 \leq l \leq 30$ are dashed and the filters $\hat{\phi}_{ml}(t)$ for $-30 \leq l \leq 0$ are solid. The larger $|l|$, the lighter the curve.

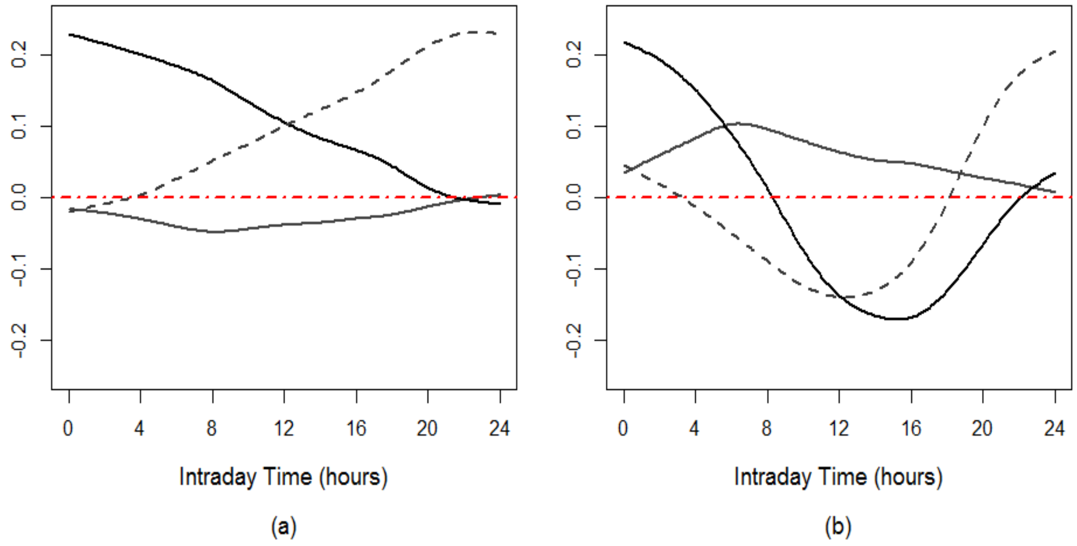


FIGURE 4.4: The 3 central functional filters $\hat{\phi}_{ml}$, $l = -1, 0, 1$, of the (a) first and (b) second dynamic FPCs. For $m = 1, 2$: the filters $\hat{\phi}_{m,-1}(t)$ are solid grey, $\hat{\phi}_{m,0}(t)$ are solid black and $\hat{\phi}_{m,1}(t)$ are dashed grey.

according to Equation 4.14, each score $\hat{Y}_{mi}$ requires L functional observations to be available before and after time i. To obtain the dynamic FPC scores at the boundaries for $1 \leq i \leq 30$ and $1066 \leq i \leq 1095$, the ends of the series are padded with the sample functional mean, by setting $x_{-29} = \ldots = x_0 = \hat{\mu}(t)$ and $x_{1095} = \ldots = x_{1125} = \hat{\mu}(t)$ (Hormann et al., 2014). This creates some bias in the scores estimated at the boundary of the observation period. The boundary areas with biased scores are marked

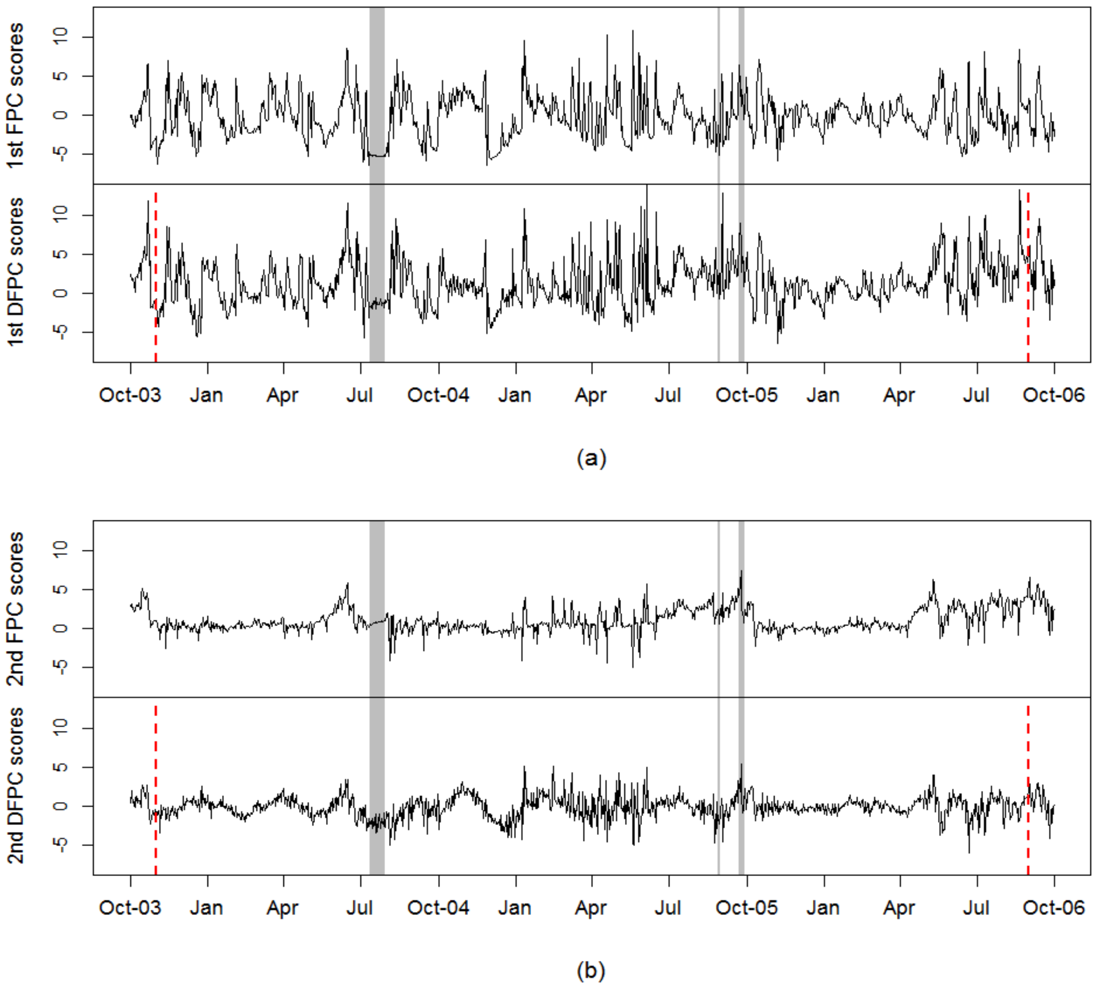


FIGURE 4.5: Time plots of the (a) first static (top) and dynamic (bottom) FPC score sequences, and (b) second static (top) and dynamic (bottom) FPC score sequences. The red dashed vertical lines indicate the boundary areas of biased scores. The grey lines highlight the missing days.

by red dashed vertical lines in Figure 4.5. The top panels of the same figure display the corresponding score sequences of the first 2 traditional FPCs. The static and dynamic scores are originally based on entirely different methodology and are loading different functions, and hence comparing them componentwise is not appropriate. However, it can be observed that they both have similar patterns over time with different magnitude. The dynamic scores seem to be more variable as they do not just involve the present observation.

Taking advantage of the fact that all functional filters $\hat{\phi}_{ml}$ for $m = 1, 2$ and $|l| > 1$ are close to zero and that the 3 central filters of the first 2 dynamic FPCs are the most important filters with $\sum_{l=-1}^{1} \|\hat{\phi}_{1l}\|^2 = 0.88$ and $\sum_{l=-1}^{1} \|\hat{\phi}_{2l}\|^2 = 0.76$, the approximation

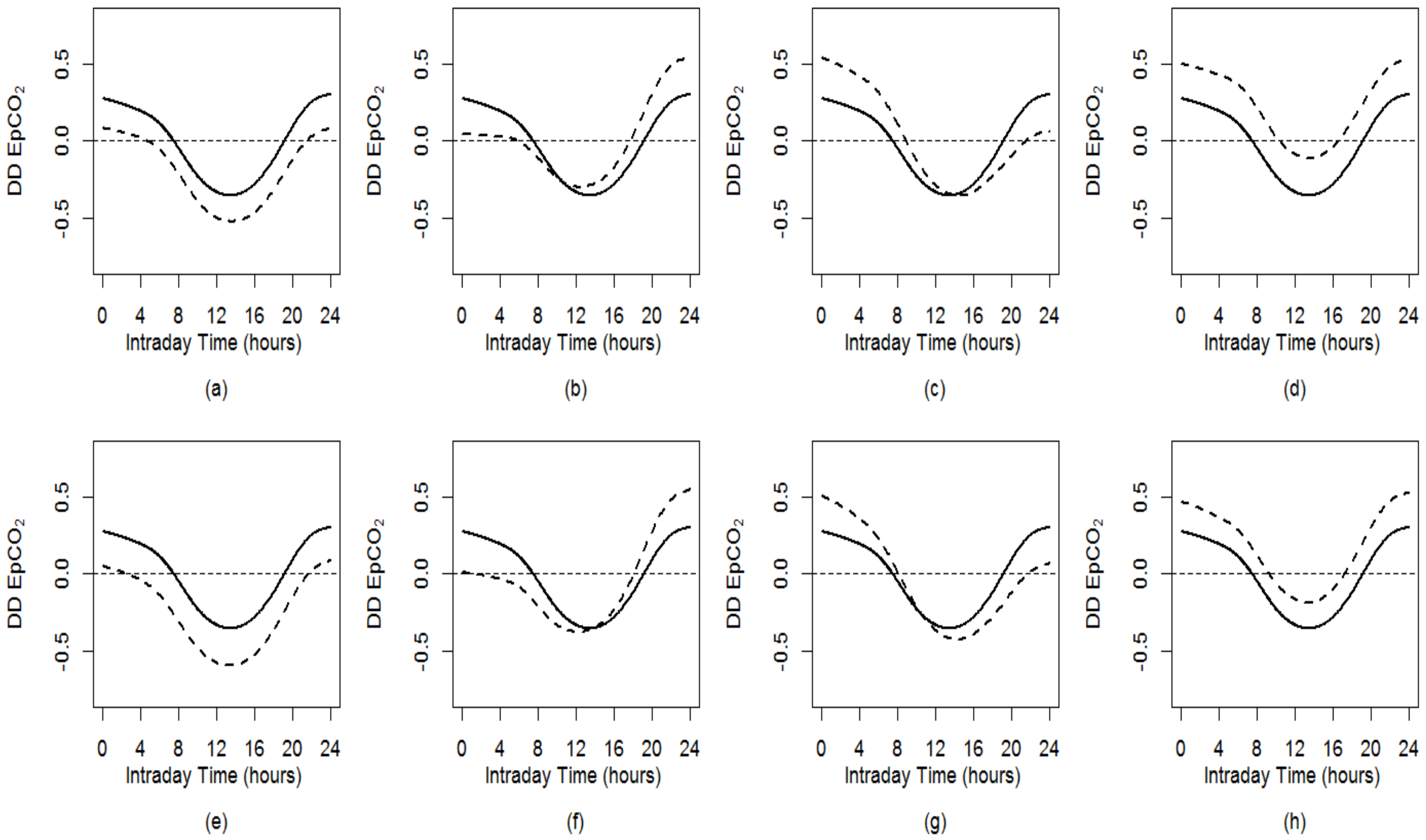


FIGURE 4.6: The functional mean $\hat{\mu}(t)$ of the DD EpCO$_2$ daily curves (solid line) and $\hat{\mu}(t) + \text{eff}(\delta_{-1}, \delta_0, \delta_1)$ with $\delta_l = \pm 1$ of $\hat{\phi}_{1l}$(dashed line)(a) $(\delta_{-1}, \delta_0, \delta_1) = (-1, -1, -1)$, (b) $(\delta_{-1}, \delta_0, \delta_1) = (-1, -1, 1)$, (c) $(\delta_{-1}, \delta_0, \delta_1) = (-1, 1, -1)$, (d) $(\delta_{-1}, \delta_0, \delta_1) = (-1, 1, 1)$, (e) $(\delta_{-1}, \delta_0, \delta_1) = (1, -1, -1)$, (f) $(\delta_{-1}, \delta_0, \delta_1) = (1, -1, 1)$ (g) $(\delta_{-1}, \delta_0, \delta_1) = (1, 1, -1)$, (h) $(\delta_{-1}, \delta_0, \delta_1) = (1, 1, 1)$

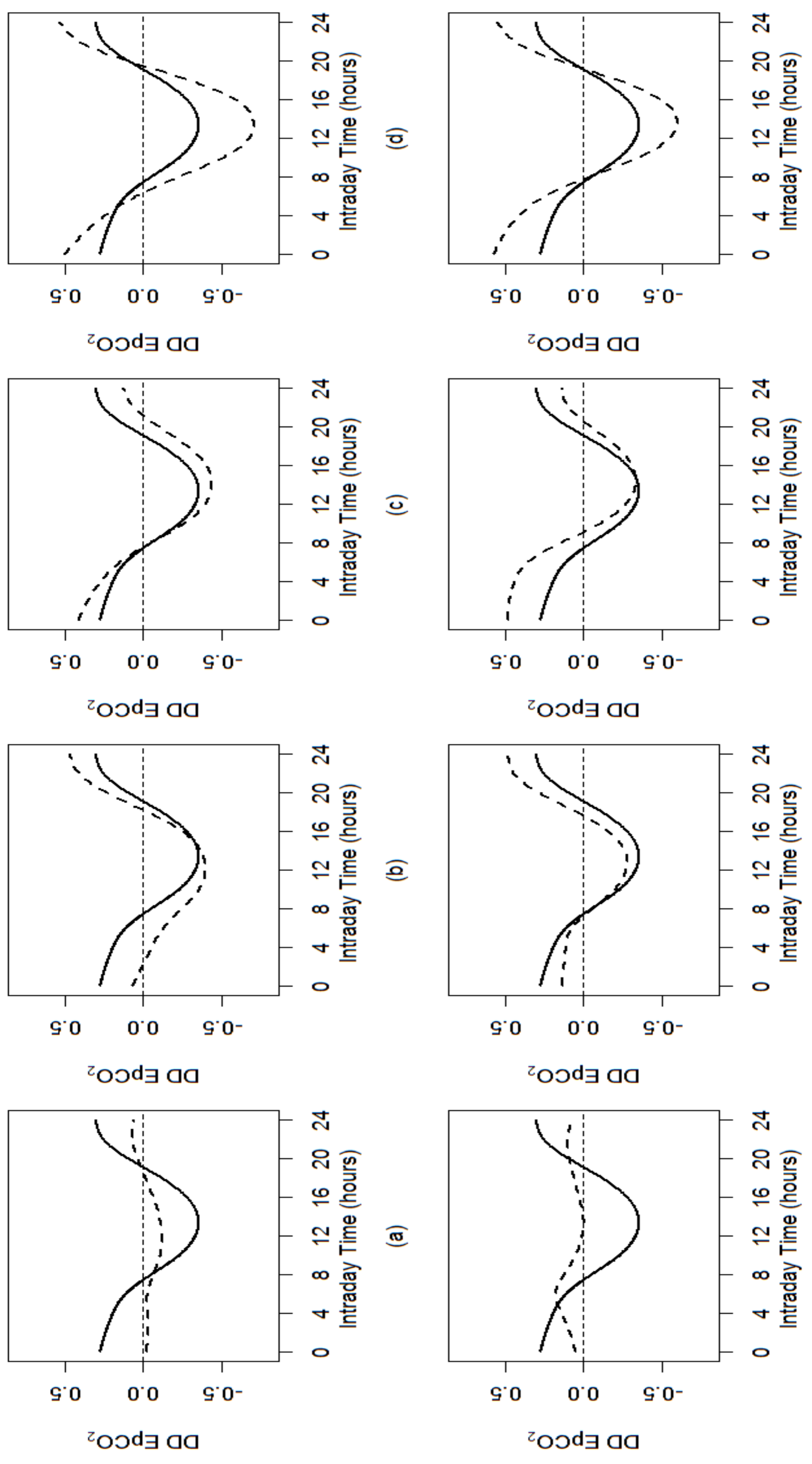


FIGURE 4.7: The functional mean $\hat{\mu}(t)$ of the DD EpCO$_2$ daily curves (solid line) and $\hat{\mu}(t) + \text{eff}(\delta_{-1}, \delta_0, \delta_1)$ with $\delta_l = \pm 1$ of $\hat{\phi}_{2l}$(dashed line)(a) $(\delta_{-1}, \delta_0, \delta_1) = (-1, -1, -1)$, (b) $(\delta_{-1}, \delta_0, \delta_1) = (-1, -1, 1)$, (c) $(\delta_{-1}, \delta_0, \delta_1) = (-1, 1, -1)$, (d) $(\delta_{-1}, \delta_0, \delta_1) = (-1, 1, 1)$, (e) $(\delta_{-1}, \delta_0, \delta_1) = (1, -1, -1)$, (f) $(\delta_{-1}, \delta_0, \delta_1) = (1, -1, 1)$ (g) $(\delta_{-1}, \delta_0, \delta_1) = (1, 1, -1)$, (h) $(\delta_{-1}, \delta_0, \delta_1) = (1, 1, 1)$

by a two-term dynamic Karhunen-Loève expansion is roughly given by:

$$\hat{x}_i(t) \approx \sum_{l=-1}^{1} \hat{y}_{1,i+l}\hat{\phi}_{1l}(t) + \sum_{l=-1}^{1} \hat{y}_{2,i+l}\hat{\phi}_{2l}(t). \tag{4.19}$$

This approximation suggests studying the sequential effect of the three consecutive scores of the first dynamic FPC $(\hat{y}_{1,i-1}, \hat{y}_{1,i}, \hat{y}_{1,i+1})$ and the second dynamic FPC $(\hat{y}_{2,i-1}, \hat{y}_{2,i}, \hat{y}_{2,i+1})$ on the EpCO$_2$ of day i. This effect can be studied by adding the following functions to the overall sample mean curve $\hat{\mu}(t)$:

$$\text{eff}_m(\delta_{m,-1}, \delta_{m,0}, \delta_{m,1}) = \sum_{l=-1}^{l} \delta_{ml}\hat{\phi}_{ml}(t), \qquad \delta_l = \pm 1, \quad m = 1, 2.$$

For instance, Figure 4.6(a) shows the effect of three consecutive small scores of the first dynamic FPC $\hat{\mu}(t) + \text{eff}_m(-1, -1, -1)$, which results in a negative shift in the mean curve. If two small scores are followed by a larger score, the EpCO$_2$ level increases and exceeds the mean as we approach the end of the day (Figure 4.6(b)). Since a large value of the score $\hat{y}_{1,i+1}$ implies a large value of EpCO$_2$ on day $i + 1$, and since the EpCO$_2$ curves are highly correlated at the transition from day i to day $i + 1$, a higher value of EpCO$_2$ is expected to occur towards the end of day i. Similarly, Figure 4.7 shows the effect of triples $(\hat{y}_{2,i-1}, \hat{y}_{2,i}, \hat{y}_{2,i+1})$ on the EpCO$_2$ of day i. Figure 4.7(a) shows that three consecutive small scores of the second DFPC result in an almost flat curve over the day with no clear difference between day and night. In contrast, Figure 4.7(h) indicates that three consecutive large dynamic scores result in a clear contrast between day and night on day i. This implies that the contrast between day and night levels of EpCO$_2$ is higher during periods of high EpCO$_2$.

Based on the above results and figures, the first 2 dynamic FPCs appear to take over a mixture of the roles of the first 3 traditional/static FPCs, presented earlier in Section 3.4.4, in approximating the daily curves of EpCO$_2$. As mentioned before in Section 3.4.4, the first static FPC is interpreted as the average deviation of the curves $x_i(t)$ from the global mean $\hat{\mu}_t^G$, while the second and third static FPCs correspond to the contrast effect between day and night and an intra-day trend effect, respectively. Accordingly, if the second static score of day i is large (or small), the difference between day and night EpCO$_2$ values is respectively large (or small) over the 24 hours of the day; and if the third static score of day i is large (or small), the curve $x_i(t)$ is respectively decreasing (or increasing) over the 24 hours of the day. However, the EpCO$_2$ daily curves are sequentially time dependent, and hence information about what is happening over the day can be derived from future and past nearby observations. Thus, we can roughly say that:

$$\sum_{m=1}^{2}\sum_{l=-1}^{1}\hat{y}_{m,i+l}\hat{\phi}_{ml}(t) \approx \sum_{m=1}^{3}\hat{S}_{mi}\hat{\xi}_m(t). \tag{4.20}$$

That is, each daily curve of $\mathrm{EpCO_2}$ can be summarized either by the first 3 static FPC scores ($x_i(t) \approx \sum_{m=1}^{3}\hat{S}_{mi}\xi_m(t)$), or alternatively by the scores corresponding to the 3 central (most influential) filters of the first 2 dynamic FPCs ($x_i(t) \approx \sum_{m=1}^{2}\sum_{l=-1}^{1}\hat{y}_{m,i+l}\hat{\phi}_{ml}(t)$).

Figure 4.8(a-c) shows a sample of the DD daily curves of $\mathrm{EpCO_2}$ smoothed using penalized cubic B-splines and the approximation of the same curves using the first 2 static FPCs and the first 2 dynamic FPCs, respectively. The difference between the two approximations in panels (b) and (c) is notable. The reconstruction based on the static FPCs in panel (b) is relatively smoother, merely provides an average level and misses the local variations in the daily curves. Whereas, the approximation based on the dynamic FPCs in panel (c) retrieves the overall day pattern in addition to the intra-day evolution and the local features in the daily curves that vary considerably from one curve to another. This illustrates that the dynamic FPCs provide a better approximation to the original curves by accounting for the temporal correlation between the curves and taking advantage of the information carried by the future and past observations. Needless to say that this approximation improves and the local features are better retrieved as the number of retained dynamic FPCs increases. Following from Equation 4.20, Figure 4.8(d) displays the approximation of the 10 sampled curves based only on the scores corresponding to the 3 central filters of the first 2 dynamic FPCs. It is noticed that although the reconstruction of the curves in panel(c) is superior to that in panel(d), the overall daily pattern and the relationships between the daily curves is still preserved.

4.3 Clustering Functional Data using Dynamic FPCs

Another goal of the chapter is to classify the daily curves of $\mathrm{EpCO_2}$ into clusters of different day regimes, taking into account the temporal correlation between the curves. To make the classification of the daily curves benefit from the information carried by nearby past and future observations, I suggest clustering the curves based on their corresponding dynamic FPC scores which have proven to provide a more appropriate dimension reduction representation than their static counterparts.

Following this suggestion and according to Equation 4.20, we propose clustering the daily curves, using any standard clustering technique, based on the sequences of the current, previous and future scores of the first 2 dynamic FPCs ($\hat{y}_{1,i-1}, \hat{y}_{1,i}, \hat{y}_{1,i+1}, \hat{y}_{2,i-1}, \hat{y}_{2,i}, \hat{y}_{2,i+1}$)

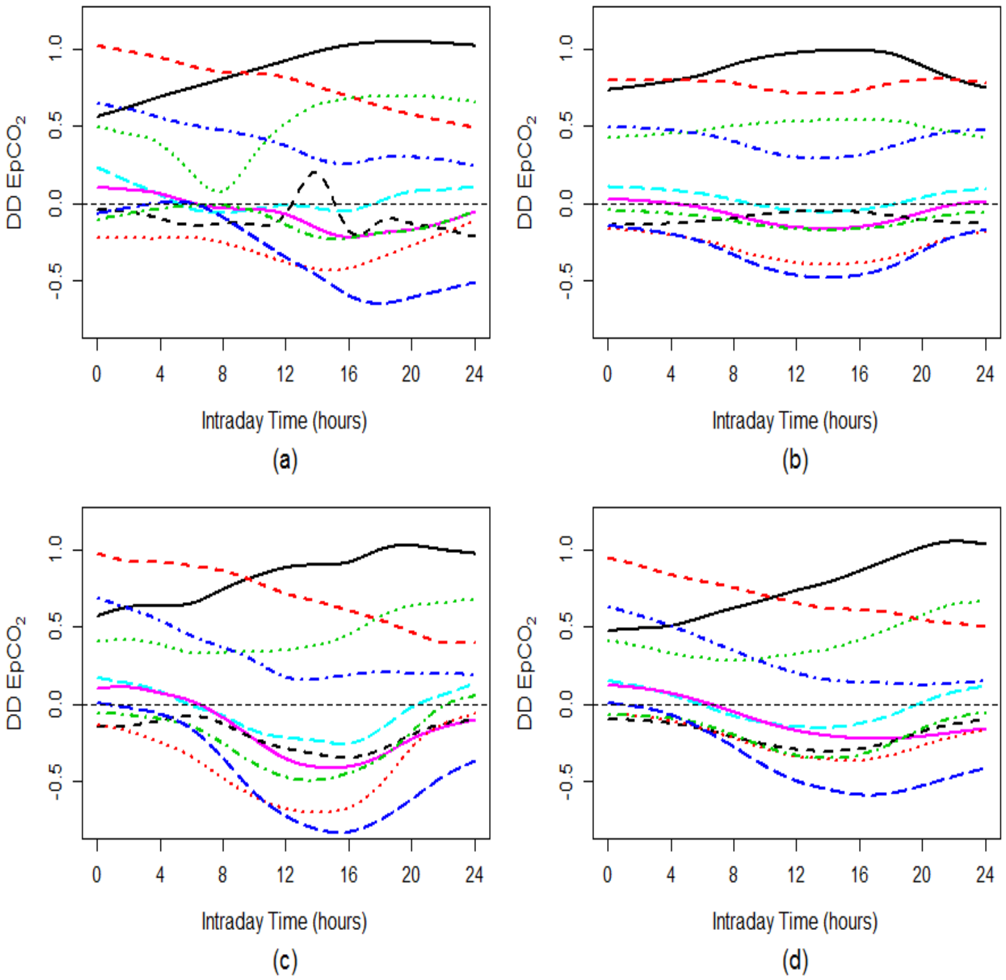
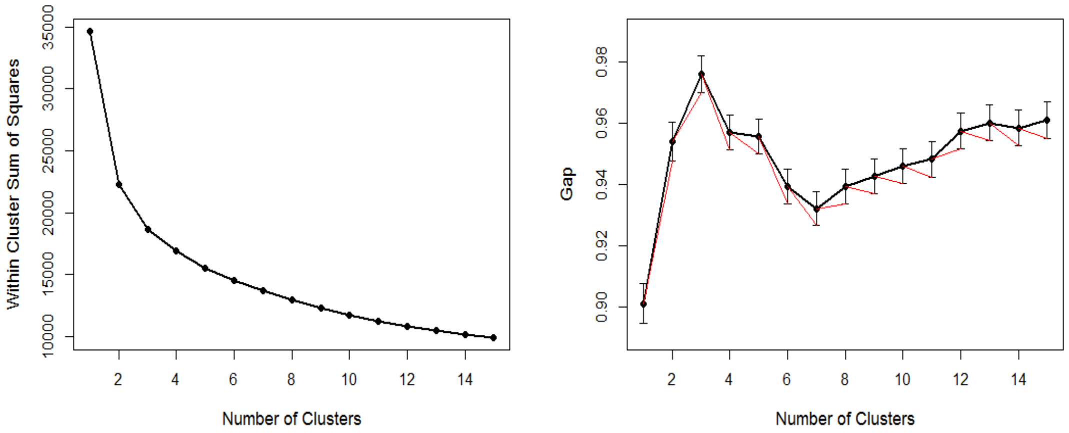


Figure 4.8: (a) 10 successive daily smooth curves, (b) the corresponding static Karhunen-Loève expansion based on 2 static FPCs, (c) the dynamic Karhunen Loève expansion based on the filters $\hat{\phi}_{ml}(t), l = -30, \ldots, 0, \ldots, 30; m = 1, 2$ and (d) the dynamic Karhunen Loève expansion based on the filters $\hat{\phi}_{ml}(t), l = -1, 0, 1; m = 1, 2$.

which are used collectively to approximate the original daily curves. Here, we have simply used the K-means clustering algorithm.

The gap statistic is first used to identify the statistically optimal number of clusters, where 500 reference data sets were generated and the K-means algorithm was applied to each data set individually for a range of different number of clusters. Afterwards, the gap statistic is calculated for each number of clusters based on the average and standard deviation of the within-cluster sum of squares calculated for each data set. The gap statistic consistently identified 3 clusters for the EpCO$_2$ daily patterns (see Figure 4.9). Following from this, a K-means clustering with 3 centers is applied to the sequences of dynamic FPC scores and the results are displayed in Figures 4.10 and 4.11. It is quite obvious that the grouping structure based on the dynamic scores differs from that based on the static scores (Figure 3.27). One reason is that the static



(a) (b)

FIGURE 4.9: (a) L-curve and (b) gap statistic plot for the K-means clustering of daily DD $EpCO_2$ curves approximated with their first 2 dynamic FPC scores.

and dynamic FPCs are based on different methodology and are loading entirely different functions. The former is based on a time domain approach ignoring the time dependence structure between the curves. Whilst, the dynamic version is based on a frequency domain approach that takes advantage of the information carried by the past and future observations and accounts for the temporal correlation between the days. Clustering based on the dynamic version of FPCs is more preferred in this situation as it provides a more appropriate dimension reduction for the data. Moreover, the dynamic FPC scores are cross-sectionally uncorrelated at a fixed day i as well as being uncorrelated at any time lag h unlike the static FPC scores which exhibit lagged cross-correlations and hence cannot be analyzed componentwise.

After confining the effects of trend and seasonality from dominating the grouping structure of the daily $EpCO_2$ curves, it is evident from Figure 4.10 that the discrepancy between the groups is fundamentally based on the overall average daily pattern, taking into account the temporal correlation between the curves. Interpreting this grouping structure is not an easy task, since the clustering is not only based on the current observational score but also the scores of the preceding and following days, which themselves are linear combinations of the previous and future observations. However, roughly speaking, it can be concluded from Figure 4.11 that the purple class underlies the days with a relatively higher $EpCO_2$ average and a fairly obvious diel cycle, possibly corresponding to short-term changes in the river hydrology often associated with short lived flow events, reflected by short-term conductivity drops. Whereas, the blue class highlights the days with average $EpCO_2$ level and a relatively shallower diel cycle, characterizing periods of longer term hydrological instability or extremely high flow events. Such high-flow events

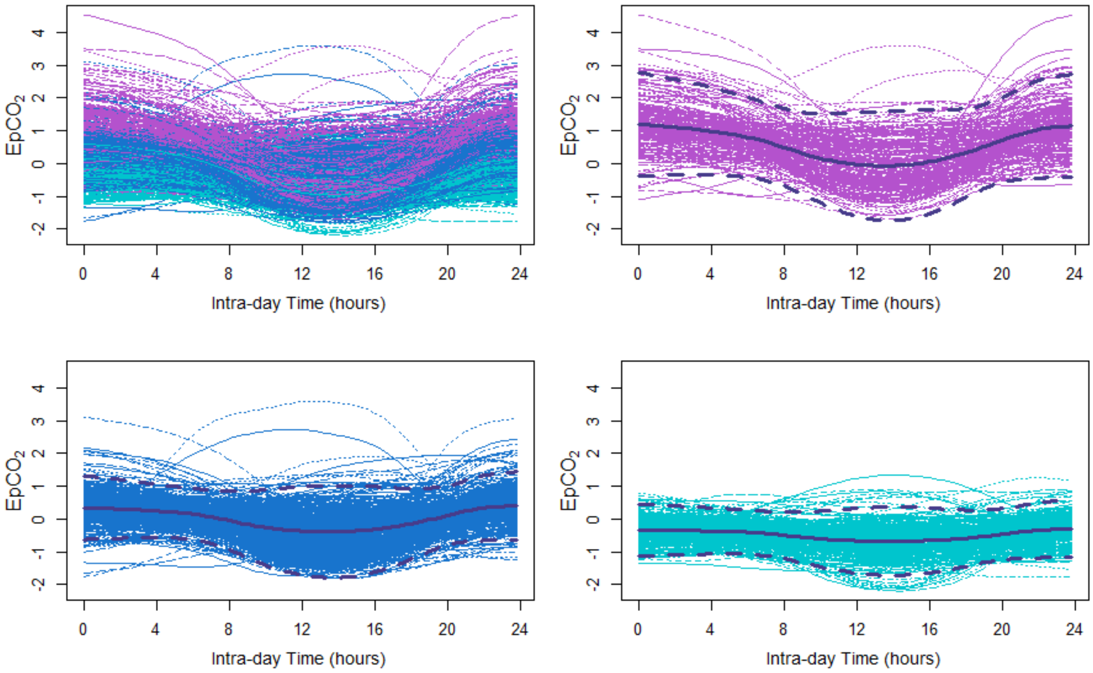


FIGURE 4.10: K-means clustering results of the daily DD EpCO$_2$ curves based on the dynamic FPC scores $(\hat{y}_{1,u-1}, \hat{y}_{1,u}, \hat{y}_{1,u+1}, \hat{y}_{2,u-1}, \hat{y}_{2,u}, \hat{y}_{2,u+1})$. The curves displayed here are recovered using the first 2 dynamic FPCs. The solid and dashed purple curves represent the mean curve and ± 2× standard deviation bands, respectively.

dilute the EpCO$_2$ and its diel cycle and lead to relatively large drops in conductivity that require a longer time to recover. In contrast, the turquoise curves generally coincide with days characterized by relatively lower EpCO$_2$ and a negligible diel cycle, with considerably stable conductivity levels following an event recovery period, in addition to some wintery days with short lived flow events.

The above clustering procedure has been repeated for different numbers of dynamic FPCs (1,3 and 4) and they all result in 3 optimal clusters with almost perfect agreement between their grouping structures (ARI ≈ 1). It is just noticed that using 2 dynamic FPCs is better than 1 dynamic FPC in terms of the stability of the results. This emphasizes that the first 2 dynamic FPCs together form the directions of variability that are most responsible for the discrepancy between the daily curves.

Note that the approximation of the curves using the dynamic Karhunen-Loève expansion improves and more local features are retrieved as more functional filters ϕ_{ml} of the first 2 dynamic FPCs are involved in Equation 4.19, see Figure 4.8(c-d). Consequently, the optimal number of clusters underlying the daily curves, approximated using their dynamic FPCs, will increase as the number of lagged filters increases. However, our main goal here is to classify the EpCO$_2$ curves into clusters based on their overall daily patterns and not their local and fine features, taking into account the time dependence

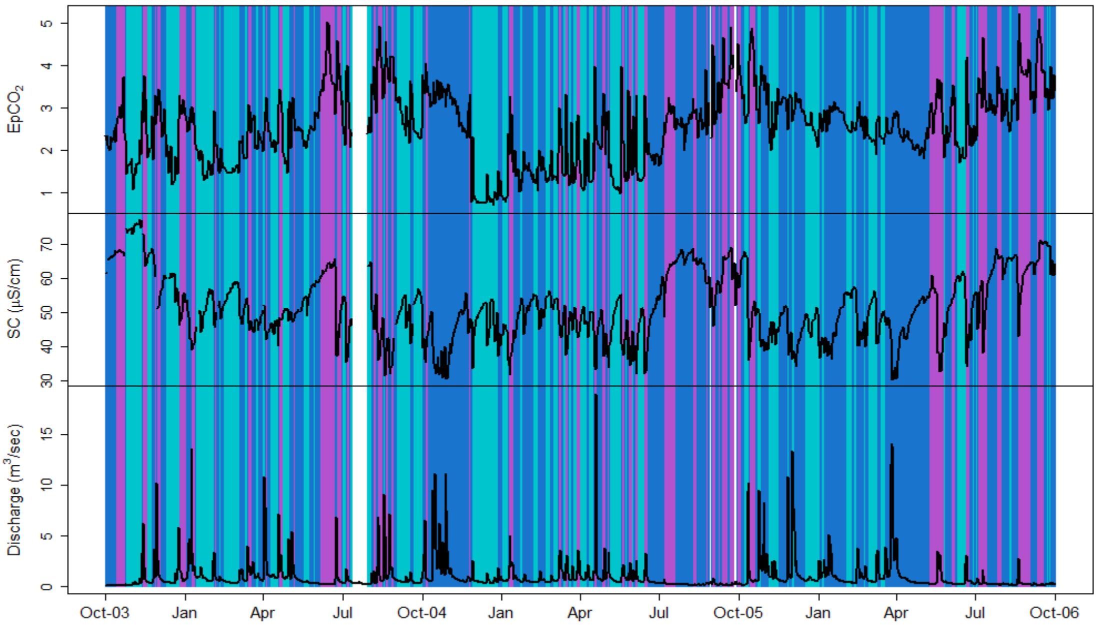


FIGURE 4.11: Class membership of days obtained using K-means clustering of the DD EpCO$_2$ curves' dynamic FPC scores. The solid curves are the corresponding daily averaged EpCO$_2$ (top), SC (middle) and water discharge (bottom).

between consecutive days. For this reason, we have decided to carry out the clustering based only on $\{\hat{y}_{m,i+l}\}, m = 1, 2; l = -1, 0, 1$, which mainly retrieves the overall EpCO$_2$ daily patterns while considering the temporal correlation between the curves.

To better understand the resulting clustering structure based on the dynamic FPCs, it is of interest to investigate the effect of the clustering structure on the daily pattern of EpCO$_2$ and evaluate the discrimination between the identified clusters. One-way functional ANOVA (fANOVA) is devoted to determine how the clusters of the curves are significantly different, while further permutation F-tests are used to assess the significant differences between the clusters. Both the fANOVA and the permutation F-tests are fitted using the `fda` package in `R`. The following sections present the theory results of fANOVA and permutation F-tests, and discuss the results of fANOVA and permutation F-tests in the assessment of the clustering results.

4.4 Functional Analysis of Variance

With analogy to the classical linear models, functional linear regression and analysis of variance are useful techniques for explaining the variability in an observed variable in terms of other observed quantities. A linear model is considered functional if the response variable is functional and the explanatory variables are scalar, or if the response variable is scalar and one or more explanatory variables are functional, or if both the response and one or more explanatory variables are functional. In all these cases, the regression coefficients, say β_j, are no longer scalar but functions, denoted by $\beta_j(t)$. One-way functional ANOVA is equivalent to standard one-way ANOVA, in which the variability in a functional response is decomposed into functional effects using a scalar matrix.

Let $x_1(t), \ldots, x_N(t)$ be N response functional observations divided into G mutually exclusive groups, such that each group g, $g = 1, \ldots, G$, has N_g observations. Then, the fANOVA model for the i^{th} functional observation in the g^{th} group can be expressed as:

$$x_{ig}(t) = \beta_o(t) + \beta_g(t) + \epsilon_{ig}(t), \tag{4.21}$$

where $\beta_o(t)$ is the overall mean function, $\beta_g(t)$ is the specific effect of group g measuring the deviations of group g from the overall mean and $\epsilon_{ig}(t)$ is the residual function accounting for the unexplained variation specific to the i^{th} individual in group g. $\epsilon_{ig}(t)$ is assumed to be an independent zero mean Gaussian stochastic process.

According to the fANOVA model given by Equation 4.21, each individual curve i belongs to only one group; and hence to uniquely identify the group effects $\beta_g(t), g = 1, \ldots, G$, the constraint $\sum_g \beta_g(t) = 0$ for all t is required. Let $\boldsymbol{\beta}(t) = (\beta_o(t), \beta_1(t), \ldots, \beta_G(t))$ be the vector of functional regression coefficients, then the model in Equation 4.21 can be re-expressed more compactly in vector-matrix notation as:

$$\mathbf{x}(t) = \mathbf{Z}\boldsymbol{\beta}(t) + \boldsymbol{\epsilon}(t),$$

where $\mathbf{x}$ is the vector of the response functions, $\boldsymbol{\epsilon}$ is the vector of the residual functions and $\mathbf{Z}$ is the design matrix of order $N \times (G+1)$. Each row of the matrix $\mathbf{Z}$ corresponds to a functional observation and each column corresponds to a coefficient function needed to be estimated. The first column of $\mathbf{Z}$ corresponds to the overall mean function and is a column of all ones, while the $(g+1)^{th}$ column corresponds to the g^{th} group effect and has 1 in the i^{th} row if the i^{th} observation belongs to the g^{th} group and zero otherwise.

With analogy to linear regression, the least squares method can be used to estimate the vector of functional regression coefficients by minimizing the following fitting criterion:

$$\int \big[\mathbf{x}(t) - \mathbf{Z}\boldsymbol{\beta}(t)\big]^\top \big[\mathbf{x}(t) - \mathbf{Z}\boldsymbol{\beta}(t)\big]\,dt, \tag{4.22}$$

subject to the constraint $\sum_{g=1}^{G} \beta_g(t) = 0$. Thus, each regression coefficient is a smooth function that can be expressed in terms of a set of basis functions. Assuming that the same set of basis functions, $\{\theta_1(t), \ldots, \theta_{p_\beta}(t)\}$, is used to smooth each regression coefficient function, the vector $\boldsymbol{\beta}(t)$ can be written in matrix notation as follows:

$$\beta(t) = \mathbf{B}\boldsymbol{\theta}(t),$$

where $\boldsymbol{\theta}(t)$ is the vector of basis functions $\{\theta_1(t), \dots, \theta_{p_\beta}(t)\}$ and $\mathbf{B}$ is the $(G+1) \times p_\beta$ matrix of basis coefficients. Let the vector of response curves, $\mathbf{x}(t)$, be also expressed in terms of its basis functions $\{\psi_1(t), \dots, \psi_{p_x}(t)\}$ as follows:

$$\mathbf{x}(t) = \mathbf{A}\boldsymbol{\psi}(t),$$

where $\boldsymbol{\psi}(t)$ is the vector of basis functions $\{\psi_1(t), \dots, \psi_{p_\beta}(t)\}$ and $\mathbf{A}$ is the $N \times p$ matrix of basis coefficients. Then, the least squares criterion given by Equation 4.22 can be equivalently written as:

$$\int \big[\mathbf{A}\boldsymbol{\psi}(t) - \mathbf{Z}\mathbf{B}\boldsymbol{\theta}(t)\big]^\top \big[\mathbf{A}\boldsymbol{\psi}(t) - \mathbf{Z}\mathbf{B}\boldsymbol{\theta}(t)\big] dt. \tag{4.23}$$

If the same set of basis functions is used to smooth both the functional response and the regression coefficient functions, then $p = p_\beta$ and $\boldsymbol{\theta} = \boldsymbol{\psi}$. Like the functional response, the smoothness of the regression coefficient functions can be controlled using a roughness penalty, such that the estimated regression coefficient functions are obtained by minimizing the penalized L^2 norm squared error given by:

$$\int \big[\mathbf{A}\boldsymbol{\psi}(t) - \mathbf{Z}\mathbf{B}\boldsymbol{\theta}(t)\big]^\top \big[\mathbf{A}\boldsymbol{\psi}(t) - \mathbf{Z}\mathbf{B}\boldsymbol{\theta}(t)\big] dt + \lambda \int [\mathbf{L}\boldsymbol{\beta}(t)]^2 dt, \tag{4.24}$$

subject to the constraint $\mathbf{L}\boldsymbol{\beta}(t) = 0$, where λ is the smoothing parameter and $\mathbf{L}$ is any differential operator, say $\boldsymbol{\beta}''(t)$. Usually, the smoothing parameter λ and the differential operator $\mathbf{L}$ is held the same for all curves to ensure a fair comparison of the estimates. The smoothing, in functional linear models, should be carried out within the estimation of the regression coefficient functions instead of the response functions which should be smoothed very little or even not at all (Ramsay and Silverman, 2005). This is mainly to guarantee that the potential variability within the individual response curves, playing a role in the estimation of the regression coefficient functions, is not lost by smoothing.

It is often desirable to construct confidence intervals for the estimated regression coefficients functions after fitting the model. Suppose that all the x_i's share the same measurements time t_j, $j = 1, \dots, T$, then the point-wise residuals are computed by:

$$\hat{\epsilon}_i(t_j) = x_i(t_j) - \mathbf{Z}_i\hat{\boldsymbol{\beta}}(t_j),$$

where $x_i(t_j)$ is the i^{th} curve evaluated at the time point t_j, $\mathbf{Z}_i$ is the i^{th} row of the design matrix $\mathbf{Z}$ and $\hat{\boldsymbol{\beta}}(t_j)$ is the estimated regression function evaluated at time t_j.

Subsequently, the error covariance matrix is estimated by:

$$\hat{\Sigma}_\epsilon = \frac{1}{N-G}\sum_{i=1}^{N}\hat{\epsilon}_i\hat{\epsilon}_i^\top,$$

which can then be used to obtain the standard error for each regression coefficient.

To evaluate the statistical difference between the mean EpCO$_2$ daily curves of the clusters, identified by the K-means clustering algorithm of the dynamic FPC scores, a one-way fANOVA model with 3 treatment groups is fitted. This fANOVA model is fitted with the EpCO$_2$ daily curves being the dependent (functional) variable and the clustering structure obtained from the K-means clustering of the dynamic FPC scores being the explanatory variable. For the fNAOVA model, the smoothed daily curves of the DD EpCO$_2$ are estimated by fitting 15 cubic B-splines, without adding a roughness penalty term, to the data within each day. The design matrix of the fANOVA model is constructed such that each row corresponds to a day and each column corresponds to a coefficient function needed to be estimated. According to the above clustering structure, 3 clusters are optimally identified and hence the number of coefficient functions to be estimated in this fANOVA model is 4. The first coefficient function reflects the overall mean function and the other 3 functions correspond to the deviations of each cluster from the mean curve of EpCO$_2$. The regression functions are estimated using penalized cubic B-splines with a smoothing parameter equal to 1, which is the same value chosen to fit the penalized cubic B-spline functions to the discrete data according to the sensitivity analysis shown in Figure 3.2. The roughness penalty is only introduced for the estimated regression coefficient functions and not the response functions, to ensure that the important variability within the daily curves of EpCO$_2$ is not missed by over-smoothing the data.

The estimated regression coefficient functions are displayed in Figure 4.12 along with their ±2 s.e. bands. The top left panel presents the grand mean effect and each of the other 3 panels reflects the effect of the corresponding cluster and how significant is the difference between the curves in that cluster and the grand mean. That is, the cluster is considered significantly different from the mean if the dashed zero line is outside the ± 2 s.e. bands. According to Figure 4.12, the average pattern of the turquoise cluster is significantly below the mean level indicating that the EpCO$_2$ in that cluster is lower than average and although the EpCO$_2$ is close to the overall average at midday, the EpCO$_2$ is much lower than average during night; the blue cluster is not significantly different from the overall mean level at all time points of the day; and the purple cluster is significantly above the mean level except at some points around midday, indicating

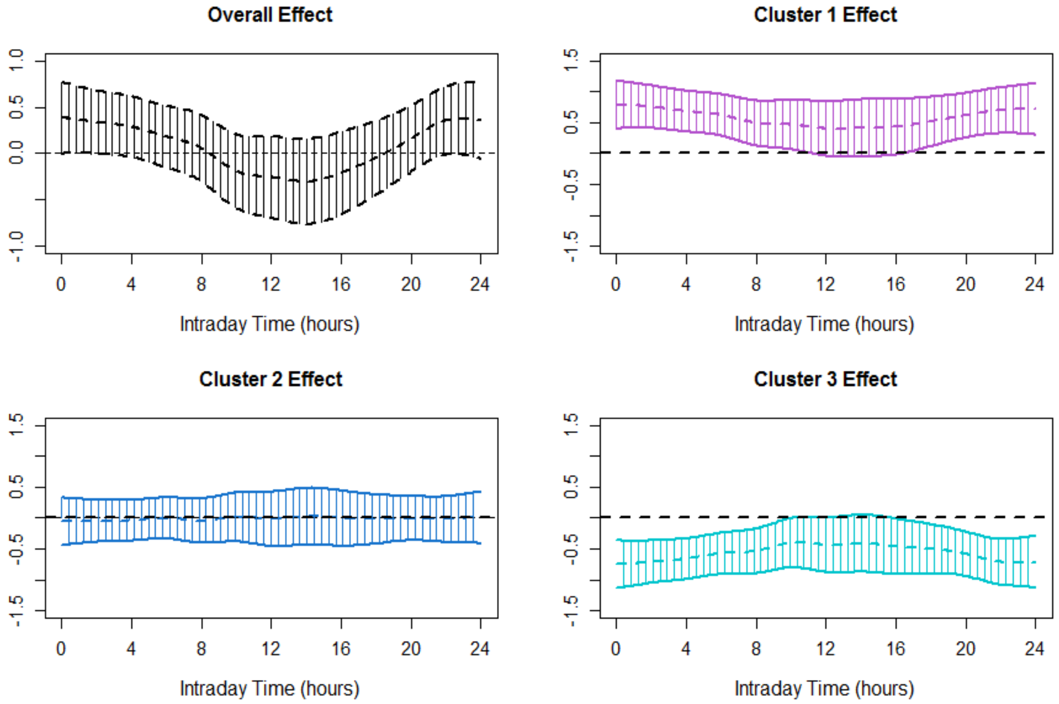


FIGURE 4.12: Estimated regression coefficient functions of the fANOVA model for the overall mean function of $EpCO_2$ and the cluster effects with ± 2 s.e. bands.

that although the $EpCO_2$ is higher than average in that cluster, it is much higher than average during night. Thus, the curves in the turquoise class correspond to days with a weak daily cycle and a relatively low $EpCO_2$ level; while the blue class consists mainly of the days with an average $EpCO_2$ level and a shallow daily cycle; and the curves in the purple class relate to days with a clear daily cycle and a relatively high $EpCO_2$ level with noticeable drop in this level during midday (see also Figure 4.13). Following from this, it can be concluded that the main difference between the 3 clusters is the functional mean level reflecting the magnitude of the contrast between day and night. It is also evident that the statistical significant difference between the 3 clusters occurs mainly at night, whereas the mean $EpCO_2$ during the light hours appear to be not significantly different in the 3 clusters. Note that the estimated standard errors functions of the regression coefficient function are not adjusted for the correlation between the curves and hence the obtained ± 2 s.e bands might be narrower than what it should be.

4.5 Permutation Functional F-Test

The primary objective of fANOVA is to assess if there is any statistically significant difference between the mean curves of the G groups anywhere at time t, whereas a permutation F-test is used to test the null hypothesis:

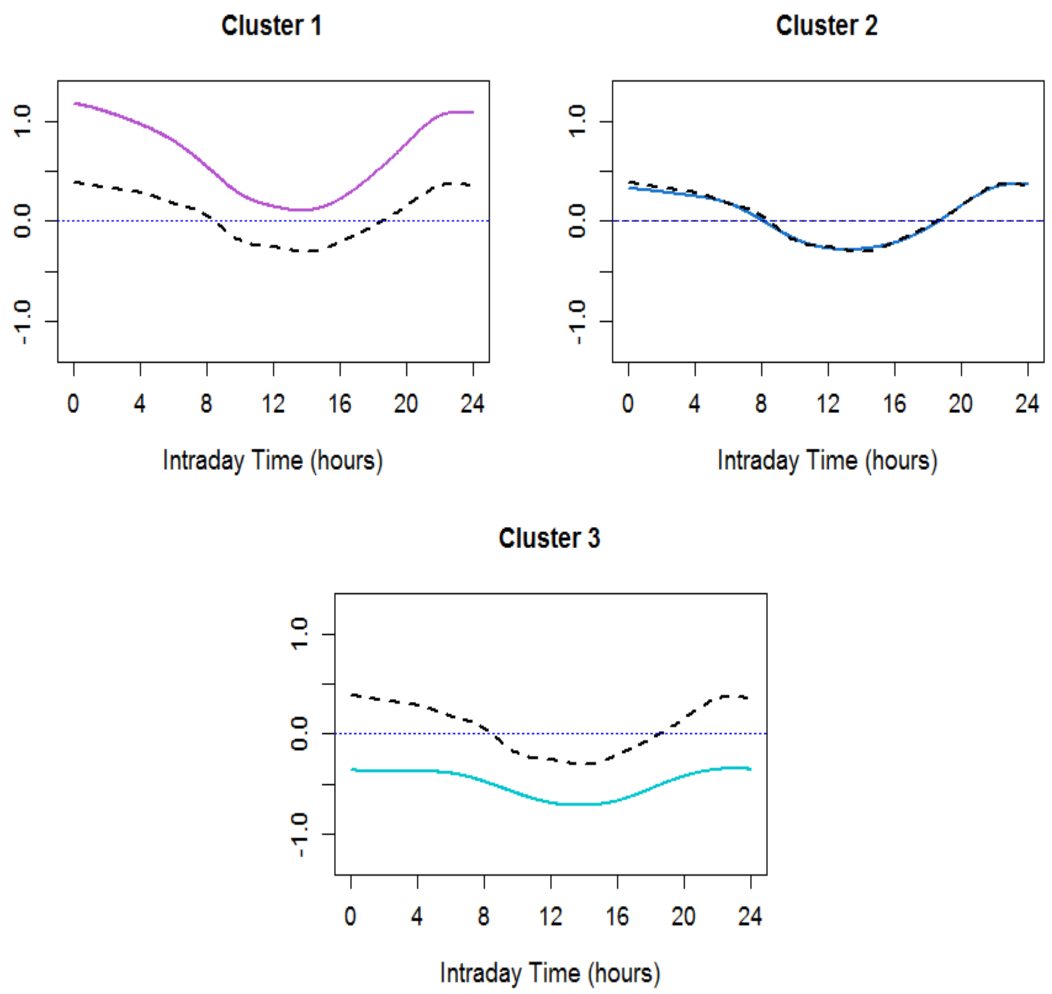


FIGURE 4.13: Estimated $EpCO_2$ average profiles of each identified cluster in the fANOVA model (solid curves). The dashed black curve is the overall mean function of $EpCO_2$.

H_o: No difference between the G groups;

versus the alternative hypothesis:

H_A: There is a difference between at least 2 of the G groups.

To evaluate the above test of hypothesis, Ramsay and Silverman (2005) suggested a point-wise permutation F-test, in which the functional responses are evaluated on a finite grid of points $\{t_1, \ldots, t_T\} \in \mathcal{T}$ then univariate F-test is performed at each time point $t_j = 1, \ldots, T$. Ramsay and Silverman (2005) considered the following point-wise F-statistic:

$$F(t_j) = \frac{\text{Var}[x(t_j)]}{\sum_{i=1}^{n}(x_i(t_j) - \hat{x}_i(t_j))^2}, \tag{4.25}$$

where $\hat{x}_i(t_j)$ is the predicted value of $x_i(t_j)$ obtained from the fitted fANOVA model. To perform the inference across the time t, the values of $F(t_j)$ and the corresponding permutation α-level critical values are plotted at each time point t_j. If the plotted $F(t_j)$ exceeded the permutation critical value over a certain time period, the test is declared to be statistically significant at that time period. However, this point-wise approach does not account for the multiple testing problem at each time point t_j. Therefore, Ramsay et al. (2009) suggested to perform an overall significance F-test using the maximum of the F-statistic given by Equation 4.25. This overall F-test detects differences anywhere in t instead of performing inference at each time point t_j across t. The distribution of the overall F-statistic, obtained as the maximum of the F-statistic function, under the null hypothesis is constructed by randomly permuting the curves and calculating this overall F-statistic across all permutations. Finally, the p-value of the test is obtained as the proportion of times where the maximum value of the permutation F-statistic function is greater than the maximum of the observed F-statistic function. Alternatively, Shen and Faraway (2004) proposed a region-wise test, in which the functional F-test statistic is based on the L^2 norm of the residual functions.

A block-wise permutation F-test is used here to test the null hypothesis of no difference between the means of the 3 clusters of EpCO2 daily curves versus the alternative hypothesis of at least 2 clusters are different, taking into account the temporal correlation between the curves. In block-wise permutation tests, the observations (curves) sequence of length N is first divided into B non-overlapping sequences (or blocks) each of length $\mathcal{N}$ (i.e. $N = B\mathcal{N}$) and instead of permuting the individual curves $x_i(t)$, the blocks $x_{\mathcal{N}(b-1)+1}(t), \ldots, x_{\mathcal{N}b}(t), b = 1, \ldots, B$, are permuted. Second, for each permutation of blocks, the test statistic function is computed using the data ordered according to this random block permutation such that the order of observations within blocks is preserved. The number of observations within each block should be large enough to encompass enough information about the dependency structure in the data, so that the unknown stochastic properties of the time series under the null hypothesis are preserved after re-sampling (Carlstein et al., 1998).

The estimated auto-correlation functions of the DD EpCO$_2$ curves' basis coefficients in Figure 3.22 and the lagged cross-correlations between the 15-minute DD EpCO$_2$ in Figure 4.14 indicate that the daily curves tend to be significantly correlated up to a maximum of 30 lags (a month). Accordingly, the months are chosen to be the splitting blocks and the data are divided into 36 blocks. The permutation test is performed by randomly permuting these 36 blocks 500 times, while maintaining the order of observations within each block, and calculating the test statistic function for each random permutation individually. Finally, the null distribution of the test statistic is constructed and the p-value of the test is obtained as the proportion of sampled permutations where the maximum value of the permutation test statistic function is greater than the maximum of the observed test statistic function.

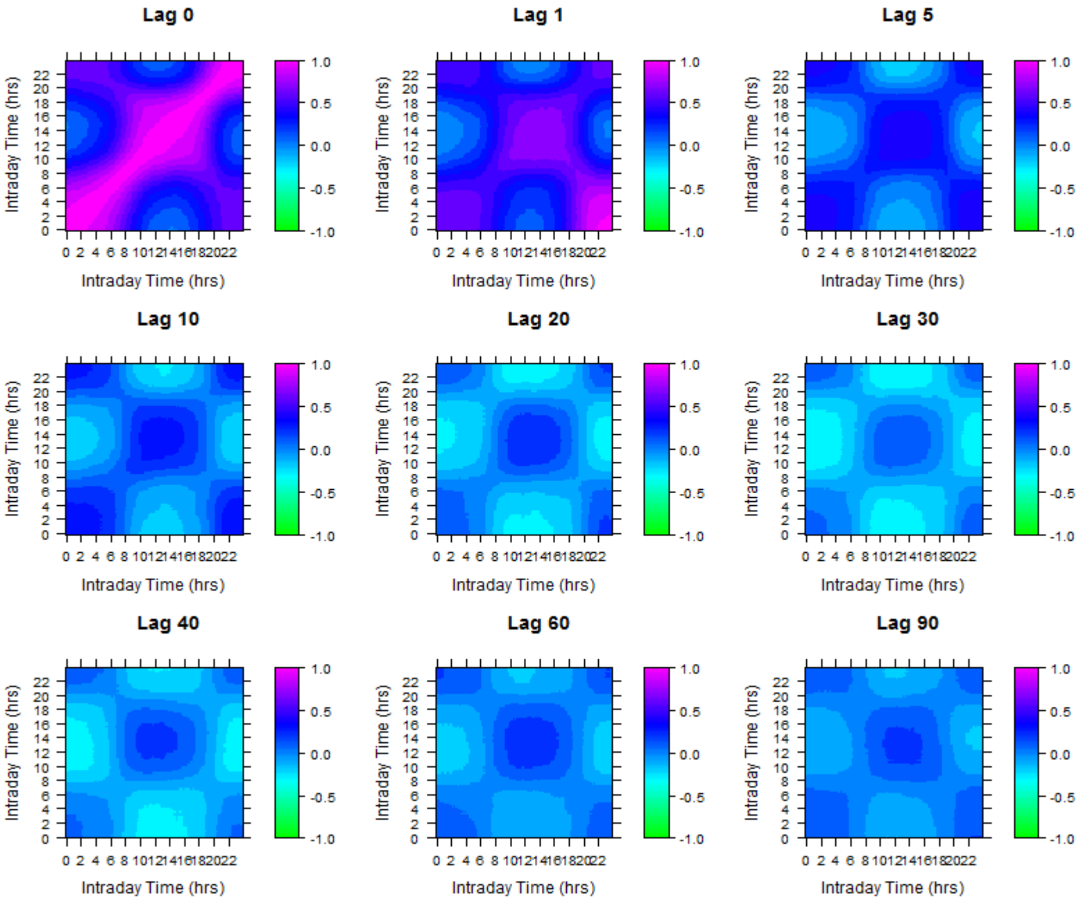


FIGURE 4.14: Estimated lagged cross-correlation matrix of the 15-minute DD EpCO$_2$ at the lags: 0, 1, 5, 10, 20, 30, 40, 60 and 90 days.

Figure 4.15 shows the results of the functional block-wise permutation F-test, where the solid black curve represents the observed F-statistic function; and the red dotted and blue dashed lines correspond to the point-wise critical value and the overall maximum critical value at 5% level of significance, respectively. It is evident that there is a statistical significant difference between at least 2 of the 3 clusters as the observed F-statistic falls above both the point-wise critical value and the overall maximum critical value at all time points. The p-value of the test is less than 0.001 indicating that, taking into account the dependence structure in the data, the alternative hypothesis of at least 2 groups are different is not rejected with a high probability. The largest difference is observed between the purple class, characterized by a relatively high EpCO$_2$ level and a clear daily cycle, and the turquoise class, characterized by a low EpCO$_2$ level and a shallower daily cycle.

Figure 4.16 compares the results of the single observation permutation F-test which ignores the time dependence structure between the curves and the block-wise permutation F-test which takes into consideration this temporal dependence structure. The

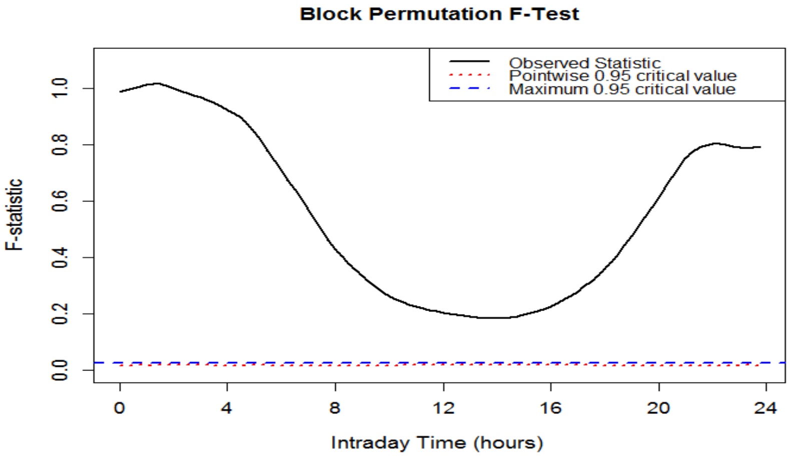


FIGURE 4.15: Functional block-wise permutation F-test for the daily curves of $EpCO_2$.

histogram of the calculated F-statistic under the null distribution of block-wise permutation test is more spread out than that of the single-observation permutation test and hence, the 95% quantile of the null distribution of the block-wise permutation test is relatively larger than that of the single-observation permutation test. However, the observed F-statistic ($F_{\text{obs}} = 1.02$) is far larger than the 95% quantile of the null distribution of both tests. This might suggest that the effect of correlation is relatively small compared to the size of the differences between the identified clusters.

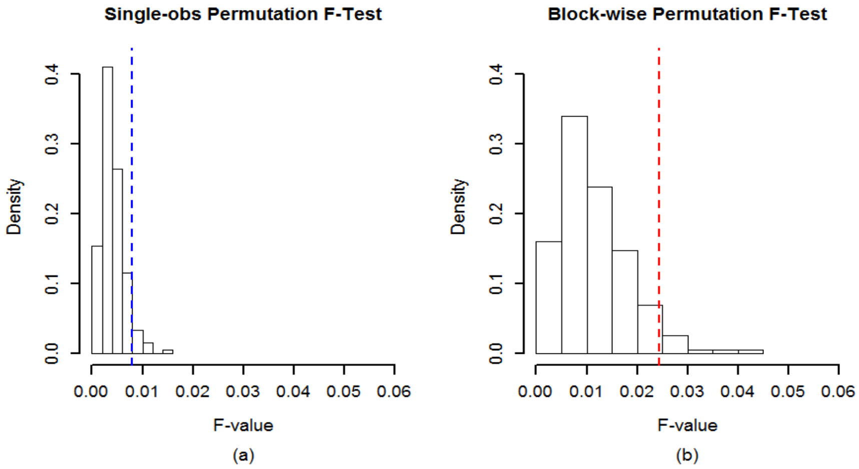


FIGURE 4.16: Histogram of the functional F-statistics calculated from 500 random permutations under the null distribution of the (a) single observation and (b) block-wise permutation functional F-tests.

4.6 Explaining $EpCO_2$ Clustering Structure using River Hydrology

In theory, the water hydrodynamics partly contribute to the $EpCO_2$ and its diurnal and seasonal variations (Li et al., 2013, Waldron et al., 2007, Yao et al., 2007). Therefore, it is of great concern to investigate and understand the nature of relationship between the different daily patterns of $EpCO_2$ and the river hydrology (reflected here by the measured specific conductivity). To study how the above clustering results based on the dynamic FPC scores reflect the status of the underlying hydrology, the distributions of specific conductivity (SC) in the 3 identified clusters are compared.

Firstly, the daily averaged SC in the 3 clusters of $EpCO_2$ daily curves are compared. The daily means of SC are calculated after trimming the lowest and highest 10% of the daily data to limit the effect of unusual SC values. The average SC is found missing for 27 days, where a natural cubic interpolating spline is used to interpolate them. Next, a smooth function is fitted to these daily averages, using penalized cubic B-splines with a knot every 3 days and a smoothing parameter equal to 1×10^{-7}. A large number of basis functions combined with a very small smoothing parameter are employed to closely follow the drops and rises in SC (see Figure 4.17). The fitted smooth function is then evaluated at the fine grid of daily time points and the distributions of the evaluated daily means in the 3 clusters are estimated and compared.

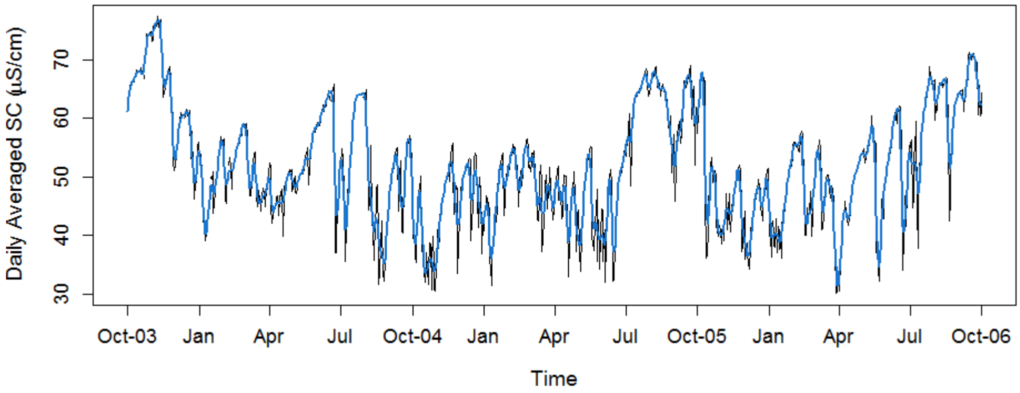


FIGURE 4.17: Time plot of the discrete daily average specific conductivity (black) and the corresponding fitted smooth curve (blue).

Figure 4.18(a-b) displays respectively the box-plot and the empirical cumulative distribution of the smoothed daily averaged SC for each cluster of curves. The Figure shows that the turquoise class, except few outliers, has mostly daily averaged SC centered around 50 reflecting stability in river hydrology, whereas the distributions of both the

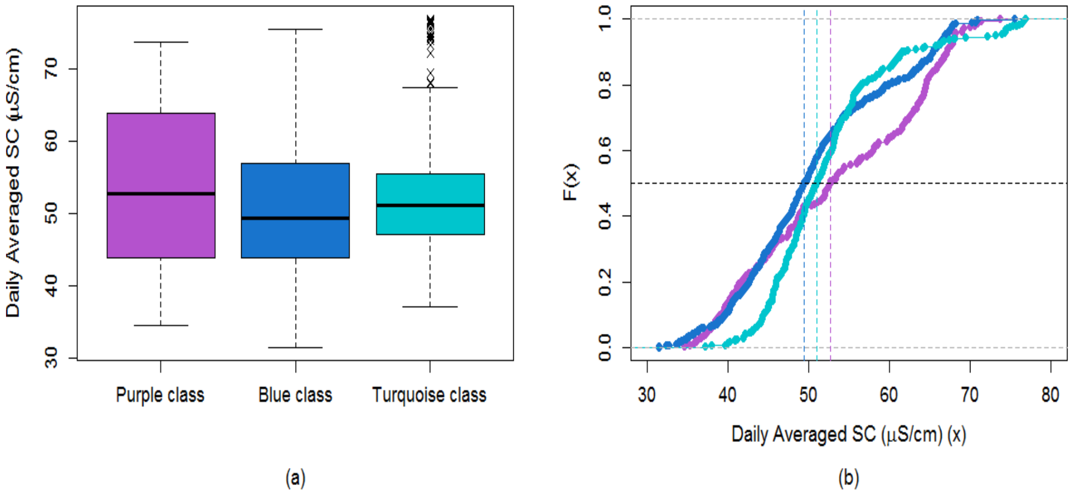


FIGURE 4.18: (a) Box-plots and (b) empirical cumulative distribution plots (probability plots) of the daily averaged SC in the 3 clusters of daily EpCO$_2$ curves identified based on the dynamic FPC scores. The horizontal dashed black line is the 50% quantile of the data and the vertical dashed purple, blue and turquoise lines correspond to the median of SC in each cluster.

purple and blue classes are relatively spread out over a wider range of SC. It is also evident that the probability function of the purple class tend to lie on the right of that of the blue class, implying that the blue class consists of days with relatively lower SC. This result suggests the persistent hydrological instability underlying the blue class, which makes the SC drop to very low levels that take a longer time to recover and rise back to normal levels.

To further assess the statistical significance of the difference between the average SC in the 3 clusters, a block permutation F-test is used. The test is employed to evaluate the null hypothesis of no difference between the clusters' means of daily averaged SC versus the alternative hypothesis of at least 2 clusters' means are different. To construct the null distribution of the test statistic, 500 random block permutations of the daily curves are generated under the null hypothesis of no difference and the test statistic is calculated for each permutation. It is clear from Figure 4.19(a) that the observed F-statistic is greater than the permutation test statistic (F_{obs} =7.9 and p-value=0.006), indicating the evidence of a significant difference between the daily averaged SC means of the 3 clusters of daily EpCO$_2$ curves. A one-way ANOVA model fitted with the daily average of SC as the response variable and the cluster membership as the explanatory variable shows that the blue class has a significantly lower average SC relative to the other two classes (see Figure 4.19(b)). This suggests that the days in that class are mostly characterized by strong hydrological instability, reflected by the very low values of SC. Moreover, although the purple and turquoise classes have very close average SC levels, more variability is observed in the purple class. This large variability indicates that the purple cluster involves a mixture of days with both low and high SC levels.

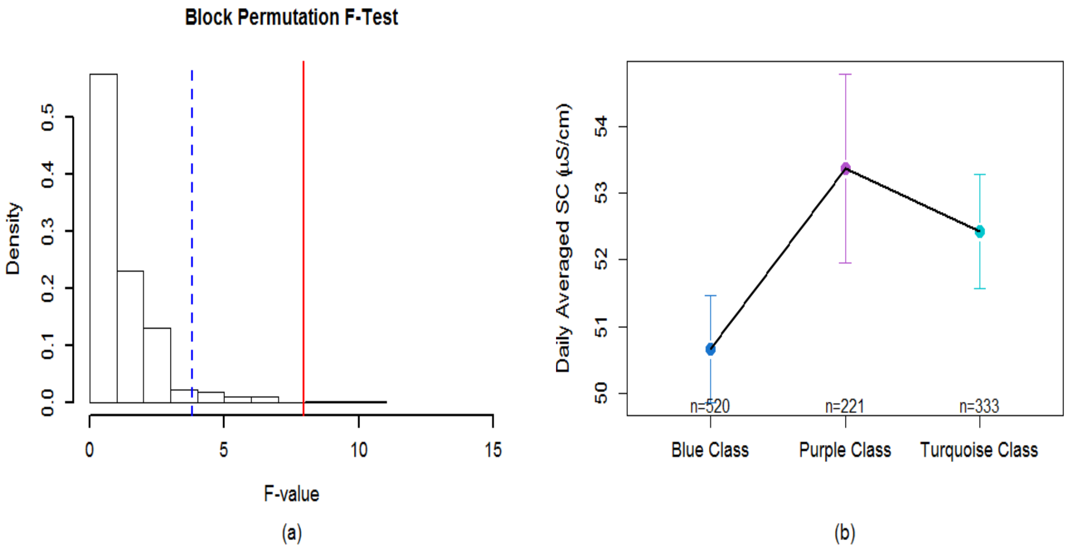


Figure 4.19: (a) Histogram of the F-statistics calculated from 500 random block permutations under the null hypothesis of no difference between the means of the daily averaged SC in the 3 clusters of daily EpCO$_2$ curves, identified based on the dynamic FPC scores. The blue dashed line is the 95% quantile of the null distribution and the red solid line is the observed statistic. (b) Plot of the mean daily averaged SC in each cluster and the corresponding ± 2 s.e bands.

Secondly, the distributions of the daily variances of SC in the 3 clusters of daily EpCO$_2$ curves are compared, where a relatively large variance reflects a considerable change in the SC within the day. Robust estimates for the daily variances of SC are obtained by calculating the variance of the SC data within each day after trimming the lowest and highest 10% of the data within each day. Because there exist long sequences of variances close to zero, natural cubic interpolating splines failed to interpolate the missing data with positive values. Instead, each missing value (of the 27) is interpolated with the average value of the last and first daily variances available before and after the period of missings plus a normally distributed random noise. Thereafter, a set of penalized cubic B-splines with a knot every 3 days combined with a roughness penalty is used to fit the smooth function for the daily variances, which is then evaluated at the fine grid of daily time points. Because the variance data are too noisy and many sequences of daily variances are very close to zero, some evaluated values turn to be negative when the fitted curve is less penalized. Ensuring that the estimated variances are all positive, the smoothing parameter is set equal to 1×10^{-5}. This smoothing parameter value provides an estimated smooth curve that captures the location of spikes but not their magnitude (see Figure 4.20).

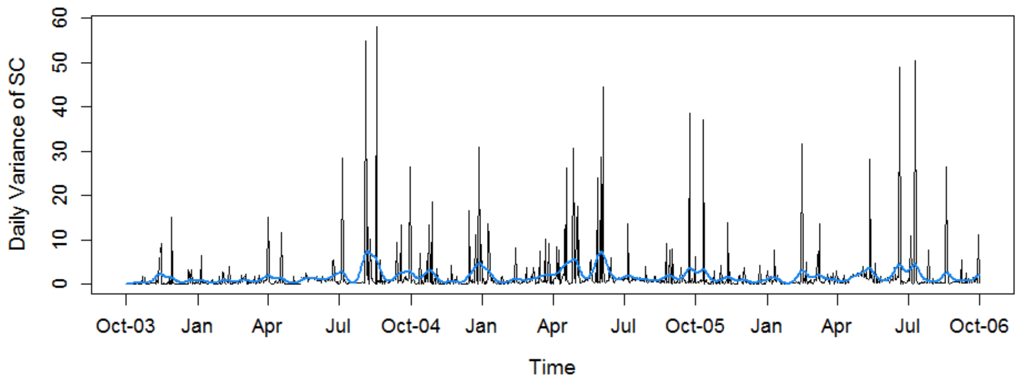


FIGURE 4.20: Time plot of the discrete daily variances of specific conductivity (black) and the corresponding fitted smooth curve (blue).

The empirical cumulative distribution function of the smoothed daily variances in each of the 3 clusters is estimated and displayed in Figure 4.21(a). It is noticed that the distribution function of the purple class lies to the right of that of the other 2 classes, implying that this class has relatively larger daily variances. This large within-day SC variability in addition to the between-day variability observed in Figures 4.18(a) and 4.19(b) indicate that the purple class encompasses the days with short-term SC drops that quickly recover back to its normal level. Whereas, the days belonging to the blue or turquoise classes, except few outliers, are characterized by more consistent hydrology within the day. This result suggests that the blue class underlies periods of long-term hydrological instability or extremely high flow events that take a longer time to recover. Thus, although the SC might be high within the days of that class, the variability between successive measurements should be relatively small reflecting the gradual recovery in hydrology. The extremely large variances in the blue and turquoise classes reflect the days in which the conductivity drops heavily as a result of a high flow event.

To further investigate if the clusters are formed such that they mirror long or short periods of hydrological instability; the first derivative of the estimated smooth function of the variances is obtained and evaluated at the same fine grid of daily time points, then the distribution functions of these derivatives in the 3 clusters are compared. Figure 4.21(b) shows that only a small proportion of the days in the purple class has a zero rate of change, revealing the relatively large changes in the SC variance from one day to another and consequently the short-term hydrological instability underlying the purple class. In contrast, a larger proportion of the days in the blue class has a zero or close to zero rate of change, indicating the persistent hydrological status of the days in that class.

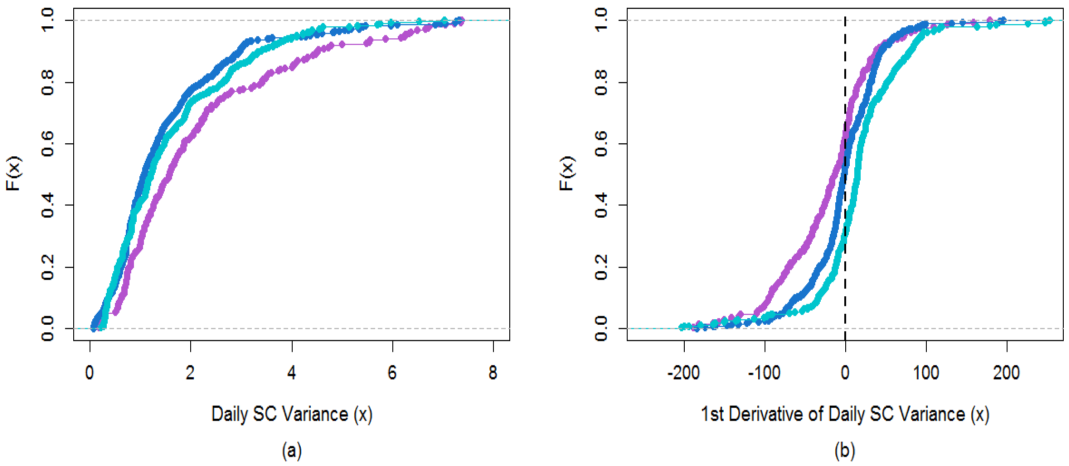


FIGURE 4.21: Empirical cumulative distribution plots (probability plots) of (a) the daily SC variances and (b) the first derivative of these daily variances in the 3 clusters of daily EpCO$_2$ curves identified based on the dynamic FPC scores.

4.7 Summary and Discussion

This chapter has investigated the effect of the temporal dependence structure in functional time series on the standard functional dimension reduction techniques. Traditional FPCA and FCA ignore the potential information, carried by the nearby observation, provided by the serial dependence structure of the functional data under study. For this reason, they fail to provide an adequate dimension reduction representation for the series. Therefore, Hormann et al. (2014) have recently proposed a dynamic version of FPCA that accounts for the temporal correlation between the curves in the frequency domain. In dynamic FPCA, the FPCs are analyzed componentwise, as the individual components are mutually independent at all lags and leads and account for most of the variations in the original process.

Here, the dynamic FPCA has been extended to approximate functional data estimated using any type of basis functions, not only orthogonal basis functions. Applying this version of dynamic FPCA to the EpCO$_2$ daily curves, approximated using penalized cubic B-splines, indicated that only 2 dynamic FPCs explain about 90% of the variability in the data. These 2 dynamic FPCs provide a better approximation to the original curves, compared to that using 2 static FPCs. The approximation has not only retrieved the overall pattern but also the local features of the curves and the relative differences between the curves. It has been also shown that the functional filters, which play the role of the eigenfunctions in traditional FPCA, fade down and converge to zero quite rapidly as the lag increases and approaches a month. In particular, the 3 central functional filters, i.e. the filters obtained at the lags $l = -1, 0, 1$, are the most influential filter elements. Accordingly, any curve $x_i(t)$ can be approximately expressed by $\sum_{m=1}^{2}\sum_{l=-1}^{1}\hat{y}_{m,i+l}\hat{\phi}_{ml}(t)$, where the scores $\hat{y}_{mi}$ are interpreted sequentially unlike the static FPCs. This expression has motivated our proposed methodology of clustering the daily curves based on the sequences of the current, previous and future scores of the first dynamic FPCs. Following from this, a K-means clustering with 3 centers is applied to the sequences $(\hat{y}_{1,i-1}, \hat{y}_{1,i}, \hat{y}_{1,i+1}, \hat{y}_{2,i-1}, \hat{y}_{2,i}, \hat{y}_{2,i+1})$.

It is evident that the classification of the daily curves based on the dynamic FPC scores does not only rely on the overall mean pattern of the curves but also on the time dependence structure between the curves. Therefore, interpreting the clustering results was not straightforward and required further statistical analysis and comparison of distributions. This subsequent statistical analysis indicated that one of the 3 classes, generally characterized by a high EpCO$_2$ level and a strong diel cycle, underlies the days with short-term SC drops that quickly recover back to its normal level (reflected by large within and between day SC variability). It has been also concluded that the class with a relatively lower EpCO$_2$ and a shallower diel cycle mostly highlights the periods with persistent and long-term hydrological events (revealed by large overall SC variance and small within-day SC variance), while the class characterized by the lowest EpCO$_2$ and absent diel cycle mainly underlies hydrologically stable periods (indicated by relatively smaller overall and within-day SC variances).

The clustering based on the dynamic FPC scores yielded a completely different structure from that based on traditional FPCs. This can be justified by the fact that both FPCA and dynamic FPCA are based on different methodology and are loading entirely different functions. From a statistical point of view, accounting for the time dependence structure by clustering based on the dynamic FPC scores is more preferred, as it provides a more appropriate dimension reduction for the data that takes into account the information carried by nearby observations.

One limitation of the dynamic FPCs is that they depend in their construction theory on the assumption that the functional time series is weakly stationary. However, such highly dynamic environmental systems are not necessarily stationary over time. Therefore, developing appropriate statistical tools for analyzing and clustering such functional time series, taking into consideration the potential changes in the system over time, is the main focus of the next chapter.

Chapter 5

Smooth Dynamic Functional Principal Component Analysis

Dynamic functional principal components rely on the assumption that the functional time series is weakly stationary. A functional process is weakly stationary if both the mean and auto-covariance operators, or its Fourier dual, are independent of time. This assumption implies that the spectral density operator, obtained based on the whole family of covariance operators at the different time lags, does not vary over time. However, in real life, numerous environmental systems are highly dynamic, which result in a covariance structure and hence a spectral density that evolves over time. In such case, the dynamic FPCs will not provide an adequate dimension reduction representation for the data, as they do not respond to the changes over time in the behaviour of the process underlying the data available for the study. Therefore, alternative techniques for FPCA that adapt to the changes in the covariance structure over time are needed.

Techniques for performing time-varying PCA for time series data are available in the literature in both time and frequency domains. In the time domain, Melnikov et al. (2016) used a dynamic principal component analysis to identify the relationship between multiple air pollutants and reveal the dynamics of the underlying data structure. In this dynamic PCA, a standard PCA is first applied on a sliding window of fixed width then the behaviour of the eigenvalues and PC loadings is studied over time. Alternatively, Miller and Bowman (2012) suggested a smooth principal component analysis approach to investigate whether and how the covariance structure changes over time, through weighting the covariance matrix obtained at each time point by a smoothing matrix then estimating the PCs at each time point individually. The authors have simultaneously developed inferential procedures based on the obtained time-varying eigenvalues and eigenvectors to detect changes in the variance and direction of the PCs. Since PCA can also be performed in the frequency domain (Brillinger, 1981), a time-localized frequency domain principal component analysis method for locally multichannel stationary signals is proposed by Ombao and Ho (2005). In this method, a time-varying spectral density is estimated by locally smoothing the spectral density estimate within locally stationary blocks, which enables the extraction of the time-varying features and the identification of channels responsible for the large variability over time.

Such time-varying principal component analysis methods have not been extended yet to functional data analysis. The recently developed dynamic FPCA, even though they have accounted for the time dependence structure in the data, have been developed for stationary functional time series. This is a potential limitation as most of real life processes, such as the air-water CO_2 fluxes, exhibit structural changes over time according to some internal and/or external factors. The main aim of this chapter is to extend the dynamic FPCA to account for the non-stationarity in functional time series.

The chapter here first examines whether and how the ordinary dynamic FPCs evolve over time using some heuristic approaches. A bootstrapping inference procedure is then proposed to detect changes over time in the spectral density of the functional process. Later on, time-varying dynamic FPCA is proposed in Section 5.3 as a means for better assessing whether or not the spectral density, and hence the covariance structure, varies significantly over time. These time-varying dynamic FPCs provide a more appropriate dimension reduction representation for locally stationary functional data. The performance of these time-varying dynamic FPCs is assessed using an extensive simulation study in Section 5.6. Finally, a clustering based on the time-varying dynamic FPCs is suggested and the results are discussed.

5.1 Local Dynamic Functional Principal Components

To initially explore whether and how the dynamic FPCs obtained for the $EpCO_2$ daily curves change over time, the functional time series is divided into monthly blocks and the dynamic FPCA, described in Chapter 4, is applied within each block separately. Subsequently, the absolute first and second eigenvalues and the corresponding proportion of variance explained by each of the first 2 dynamic FPCs across the full set of frequencies within each month is calculated by Equation 4.18 and plotted versus the month in Figures 5.1 and 5.2. It is evident that the contributions to explained variability by each of the dynamic FPCs differ from one month to another and so are the spectral characteristics. To assess the similarities and differences in the modes of variability between the different months, the 3 central functional filters of the first dynamic FPC are plotted for each month separately in Figures 5.3, 5.4 and 5.5. In the winter months,

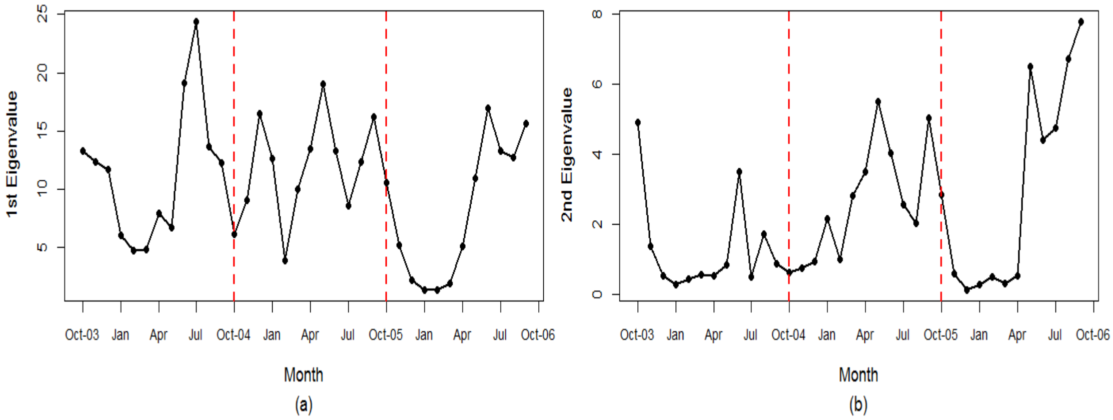


FIGURE 5.1: Plots of the (a) first and (b) second eigenvalues of the first 2 dynamic FPCs across the different months of the functional time series.

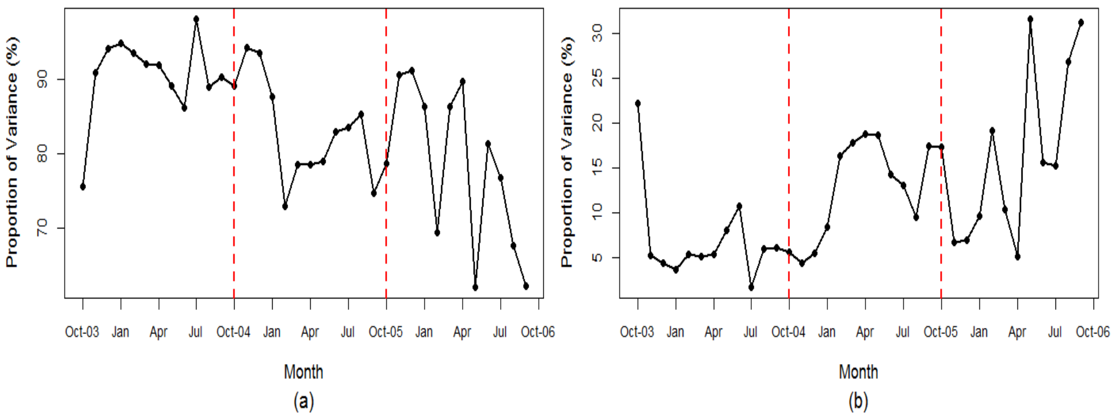


FIGURE 5.2: Plots of the proportion of variance explained by the (a) first and (b) second dynamic FPCs across the different months of the functional time series.

the functional filters tend to be flat across the 3 hydrological years. Whereas in the summer, more variability is present and the functional filters tend to exhibit different patterns across the different months and hydrological years.

To ensure a fair comparison of the spectral density across the different blocks, a more systematic method for extracting the time-varying features of the spectral density is proposed. Like in the Bartlett's method of estimating the spectral density, the entire signal is segmented into blocks of equal width and the dynamic FPCA is applied within each block individually. The eigenvalues of the first dynamic FPCs are then obtained within each block and the changes over time are studied. Note that the number of blocks increases but the frequency resolution decreases as the block width decreases. We, therefore, considered the case of overlapping blocks to increase the number of blocks without reducing the number of observations within each block. The tuning parameters of such algorithm are the fixed width $\mathcal{N}$ of the blocks and the amount of shift $\mathcal{H}$ between succeeding blocks. To estimate the spectral density within each block using the discrete

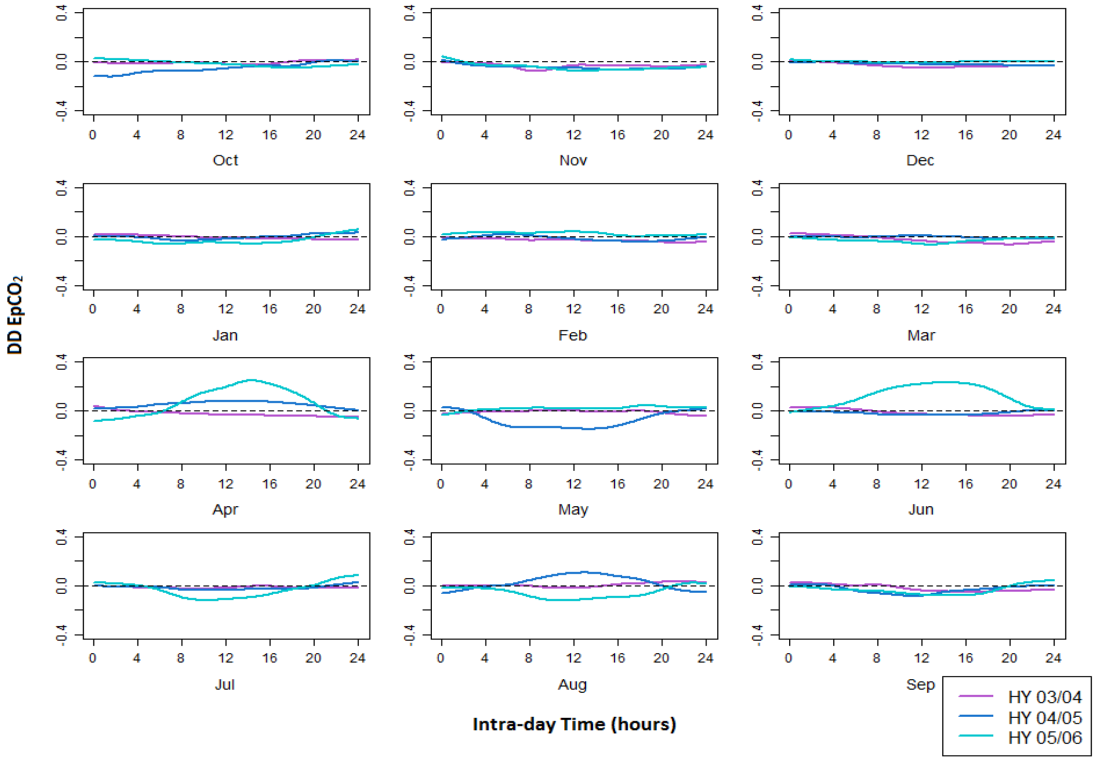


FIGURE 5.3: The functional filter $\phi_{1,-1}$ of the 1st dynamic FPC at $l = -1$ across the different months of the series. The purple, blue and turquoise curves belong respectively to the HYs 2003/2004, 2004/2005 and 2005/2006.

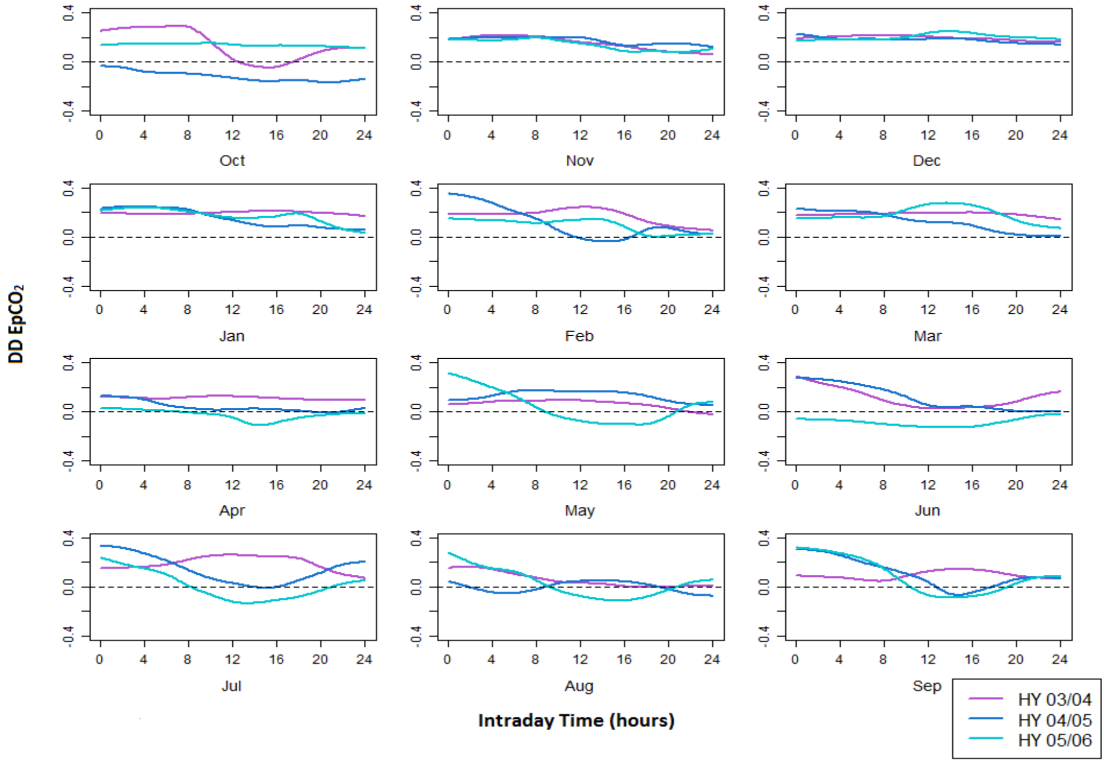


FIGURE 5.4: The functional filter $\phi_{1,0}$ of the 1^{st} dynamic FPC at $l = 0$ across the different months of the series. The purple, blue and turquoise curves belong respectively to the HYs 2003/2004, 2004/2005 and 2005/2006.

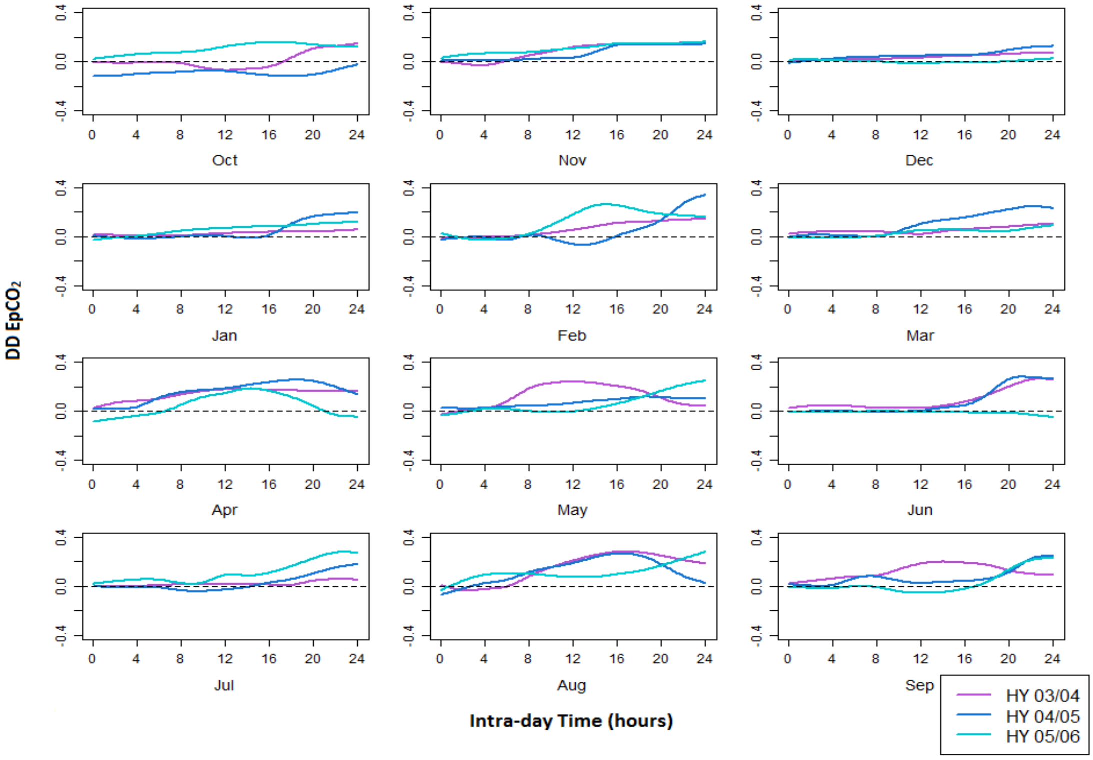


FIGURE 5.5: The functional filter $\phi_{1,+1}$ of the 1^{st} dynamic FPC at $l = +1$ across the different months of the series. The purple, blue and turquoise curves belong respectively to the HYs 2003/2004, 2004/2005 and 2005/2006.

Fourier transform, the block width is preferred to be dyadic, otherwise each block will be padded with zeros up to the next dyadic level.

Initially, the functional time series has been segmented into 34 non-overlapping succeeding blocks, each contains $2^5 = 32$ observations. A block size of 32 is chosen since earlier analysis has indicated that the correlation between successive curves is significant up to ≈ 30 days. Fixing the block size to 32 observations also ensures having a reasonable number observations for the estimation of spectral density, assuming local stationarity, within each block. Smaller block width values risk unreliable spectral density estimates due to the limited number of observations used to produce the estimates, while greater width values risk within-block non-stationarity. Figure 5.6, showing the first 2 eigenvalues of the non-overlapping blocks in purple, indicates that the eigenvalues do change considerably from one block to another over time. To evaluate changes more locally, a natural step forward is to estimate the dynamic FPCs at each time point by making the shift between succeeding blocks equal to 1 while fixing the block width to $2^5 = 32$ observations. This block width has been chosen on the basis of the reasons mentioned above, in addition to a sensitivity analysis that assesses the changes in the calculated eigenvalues to the changes in the block width. Figure 5.7 displays the sequences of the first 2 dynamic eigenvalues computed from the eigen-decomposition of the spectral

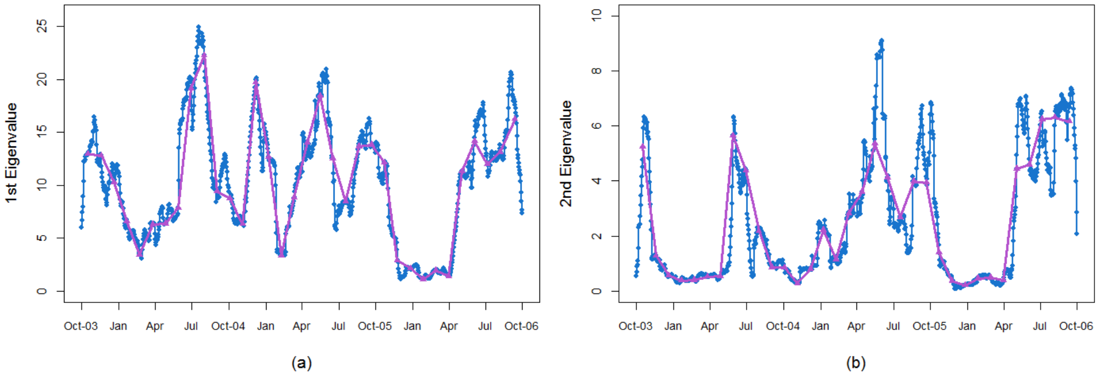


FIGURE 5.6: Eigenvalues of the (a) first and (b) second local dynamic FPCs obtained at each time point and averaged over all frequencies $\theta \in [-\pi, \pi]$ using overlapping (blue) and non-overlapping (purple) blocks of width equal to 32.

density estimated individually within blocks of size 16, 32 and 64, using the Bartlett window estimator with bandwidth $Q = \lfloor \sqrt{\mathcal{N}} \rfloor$. This sensitivity analysis shows that a smaller block width involves more variability and artefacts between successive eigenvalues, whereas a larger block width leads to smoother changes between blocks and misses some of the changes that could be of interest. In conclusion, the eigenvalues seem to vary over time and so are the spectral density and the covariance structure. However, a more formal test of stationarity is needed to assess how significantly the statistical properties of the process vary over time.

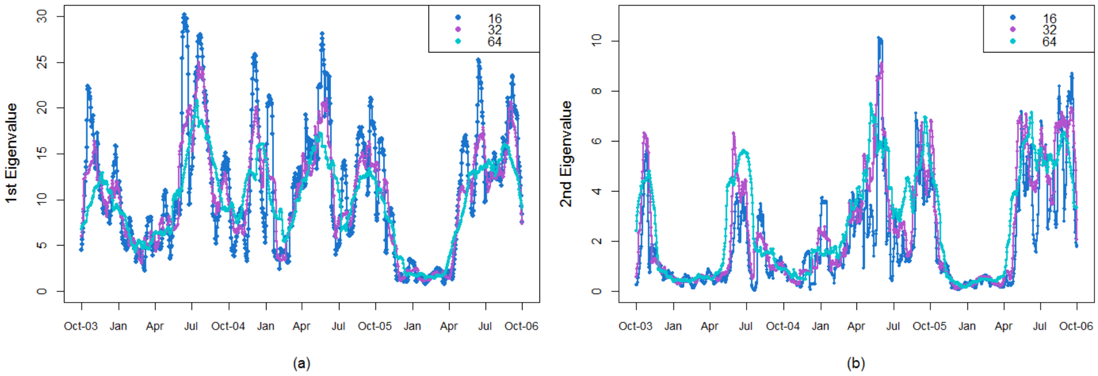


FIGURE 5.7: Eigenvalues of the (a) first and (b) second local dynamic FPCs obtained at each time point and averaged over all frequencies $\theta \in [-\pi, \pi]$ using different block widths: 16 (blue), 32 (purple) and 64 (turquoise).

Note that the time block used to estimate the spectral density at the time point i is defined by the interval $(i - \mathcal{N}/2 + 1, i + \mathcal{N}/2)$; and hence for $1 \leq i \leq \mathcal{N}/2$ and $N - \mathcal{N}/2 \leq i \leq N$, we set $X_{-\mathcal{N}/2+1} = \ldots, X_0 = X_{N+1} = X_{N+\mathcal{N}/2} = 0$. Padding the ends of the series with zeros creates some bias on the boundary of the observation period.

However, centering the time block around the i^{th} observation of interest allows for a more reliable estimate of the spectral density at this time point, for $i = \mathcal{N}/2 + 1, N - \mathcal{N}/2$.

5.2 Stationarity Test for Functional Time Series

In time series analysis, weak or second order stationarity is a fundamental assumption for estimation, forecasting and inference. However, non-stationary time series are very common in practice. Therefore, testing for stationarity has received a great attention over the years. A very simple and naive way to assess variance stationarity in a time series is to plot the partial sums of the signal $N^{-1} \sum_{t=1}^{N} x_t^2$ versus N and check if there is any trend (Mandlebrot, 1972). Throughout the years, more formal tests of stationarity have been developed in both time and frequency domains. The time domain approaches such as Kwiatkowski - Phillips - Schmidt - Shin (KPSS) (Kwiatkowski et al., 1992) and Dickey Fuller (Dickey and Fuller, 1979) tests often evaluate a null hypothesis of stationarity versus a non-stationary alternative of a unit root or random walk. Testing for a unit root is considered a very limited form of non-stationarity and more general tests are needed.

In the frequency domain, alternative approaches were developed to test for stationarity by assessing whether or not the spectral density of the signal varies over time. This approach is equivalent to testing whether or not the auto-covariance function changes over time (Berkes et al., 2009). Within this framework, Priestley and Subba Rao (1969) test if the evolutionary spectrum of the time series is constant over time using an ANOVA procedure that determines whether the time or/and frequency are responsible for the variability. Grasse et al. (2000) simplified this test by dividing the data into non-overlapping segments of the same length and estimating the spectrum for each segment separately, then testing if the spectrum is time invariant using an analysis of variance approach. Similar approaches, based on comparing spectral densities over various segments, include Paparoditis (2009), Giraitis and Leipus (1992) and Dwivedi and Subba Rao (2011). An alternative statistical test for stationarity, based on comparing the ratio of arithmetic and geometric means of spectra calculated from non-overlapping segments of the time series at a particular frequency, is proposed by Brcich and Iskander (2006). More recently, a simple test of stationarity in the frequency domain which also provides a graphical tool to represent the results has been introduced by Halliday et al. (2009). In this test, the time series is divided into B non-overlapping segments of the same width $\mathcal{N}$ and the periodogram is computed within each segment individually, such that the periodograms from non-overlapping segments are assumed uncorrelated. The ratio between the variance and mean of the periodograms obtained across the time segments at each frequency θ, known as the periodogram coefficient of variation and denoted by PCOV(θ), is computed and plotted versus the frequency θ. Under the null hypothesis, the estimated periodogram coefficient of variation is assumed to be normally distributed with mean 1 and variance $3/2B$. Thus, under the null hypothesis of stationarity, the plot of PCOV(θ) versus θ should not show any significant trend or departures from 1 by exceeding the 95% confidence limits given by $1 \pm 1.96\sqrt{3/2B}$. It is obvious that the confidence limits of this test statistic depends heavily on the number of segments such that they get wider as the segment width increases and the number of segments decreases.

Most of the above tests rely on the periodogram asymptotic distributional results and the independence assumption between the periodograms from non-overlapping segments. However, in reality, it is not reasonable to assume independence between segments of the same time series. In addition, all the former tests aim at checking the stationarity of a univariate time series and lack the ability to test for the stationarity of the covariance structure in a functional data setting. In general, there appears to be a relative paucity in stationarity tests for functional time series. At present, only change point tests using CUSUM methods are available for functional time series (Berkes et al., 2009, Horváth et al., 2010). Recently, Horváth et al. (2014) have proposed several procedures to test the null hypothesis of stationarity of functional time series in the time domain. These tests are extensions to the KPSS tests for functional time series where the alternative hypothesis includes change point and unit root, which are limited forms of non-stationarity, and hence more general tests are required. Thanks to the recent work of Hormann et al. (2014) which advances the spectral analysis of stationary functional time series, we propose a stationarity test to be based on the eigenvalues of the functional time series' spectral density operator.

5.2.1 Proposed Block Stationarity Test for Functional Time Series

The proposed test of stationarity primarily assesses whether or not the spectral density, and hence the covariance structure of a functional time series, changes over time. The null hypothesis H_0 of the test here is that the spectral density does not vary throughout time. This hypothesis will be investigated by assessing whether or not the spectral characteristics, summarized by the eigenvalues and/or eigenfunctions, of the functional spectral density change over time. The spectral density is uniquely identified by its eigenvalues and eigenvectors; and hence to assess the null hypothesis, that the spectral density does not vary throughout time, inference will be performed on the eigenvalues of the dynamic FPCs to investigate whether or not they change over time.

The proposed stationarity test rests on splitting the time series into non-overlapping blocks of equal width, then evaluating the spectral density of the functional time series within each block separately. According to the classical theory, the periodograms obtained at a certain frequency θ from non-overlapping segments are independent, such that each periodogram F_{θ}^{N} asymptotically follows a multiple of chi-square distribution with 2 degrees of freedom for $\theta \neq, -\pi, 0, \pi$ and 1 degree of freedom for $\theta = -\pi, 0, \pi$ (Shumway and Stoffer, 2011). However, these asymptotic distributional results may no longer be valid due to the persistent serial correlation between curves and the limited number of segments and observations within each segment. Therefore, the proposed test is alternatively based on obtaining the dynamic FPCs for each block separately then performing resampling-based inference to assess if these dynamic FPCs or their corresponding eigenvalues vary throughout time.

The reference information representing how a particular dynamic FPC and its corresponding spectrum (distribution of its corresponding eigenvalue across the set of all frequencies) would behave if the null hypothesis, that the spectral density does not vary over time, is true is constructed using time series bootstrapping. There exist several bootstrapping techniques for time series in the literature, where the reference distribution can be constructed either using block bootstrapping of time series (see Politis and Romano (1994)) or by simulating data from a fitted model with a known covariance structure (see, for example, Efron and Tibshirani (1986), Miller and Bowman (2012)). Due to the nature of functional time series and the serial correlation present between successive functions, the latter choice, known as the parametric bootstrap, is considered here to construct the reference distribution in the presence of correlation by simulating functional time series from a fitted stationary model with a known covariance structure that does not change over time.

In this parametric bootstrap, the serial correlation between curves in a functional time series is accommodated by simulating functional data according to a Functional Auto-Regressive of order 1, FAR(1), of the form:

$$X_{i+1} = \Upsilon(X_i) + \varepsilon_{i+1}, \tag{5.1}$$

where $\{X_i\}$ is a sequence of zero mean functions, Υ is an auto-regressive linear operator satisfying the condition $\int\int \Upsilon^2(t,s)dtds < 1$ and $\{\varepsilon_i\} \in L_H^2$ is a sequence of i.i.d zero mean functions such that $\|\varepsilon_i^2\| < \infty$. Practically, this simulation is performed in a finite dimension p. Let ψ_k, $k \in \mathbb{N}$, be the set of basis functions used to approximate the curves

$\{X_i\}$, then, owing to the linearity of Υ:

$$\begin{aligned}\langle X_{i+1}, \psi_{k'}\rangle &= \langle \Upsilon(X_i), \psi_{k'}\rangle + \langle \varepsilon_{i+1}, \psi_{k'}\rangle \\ &= \langle \Upsilon(\sum_{k=1}^{\infty}\langle X_i, \psi_k\rangle), \psi_{k'}\rangle + \langle \varepsilon_{i+1}, \psi_{k'}\rangle \\ &\approx \sum_{k=1}^{p}\langle X_i, \psi_k\rangle\langle \Upsilon(\psi_k), \psi_{k'}\rangle + \langle \varepsilon_{i+1}, \psi_{k'}\rangle.\end{aligned}$$

By letting $\mathbf{a_i} = (\langle X_i, \psi_1\rangle, \ldots, \langle X_i, \psi_p\rangle)^\top$ and $\boldsymbol{\varepsilon_i}^* = (\langle \varepsilon_i, \psi_1\rangle, \ldots, \langle \varepsilon_i, \psi_p\rangle)^\top$ be the set of basis coefficients of X_i and ε_i, respectively; the first p basis coefficients of X_i approximately satisfy the Vector Auto-Regressive of order 1, VAR(1), model given by:

$$\mathbf{a_{i+1}} = \mathfrak{R}\mathbf{a_i} + \boldsymbol{\varepsilon^*_{i+1}}, \tag{5.2}$$

where $\mathfrak{R} = (\langle \Upsilon(\psi_k), \psi_{k'}\rangle : k \geq 1, k' \leq p)$ is the matrix of auto-regressive parameters and $\boldsymbol{\varepsilon^*_i}$ are i.i.d N$(0, \Sigma)$. The covariance matrix Σ is assumed not to vary over time. The VAR(1) model is covariance stationary if the eigenvalues of $\mathfrak{R}$ lie inside the unit circle, i.e. if the eigenvalues of $\mathfrak{R}$ are less than 1 in modulus (Zivot and Wang, 2006).

According to the above procedure, curves are simulated under the null hypothesis that the spectral density, and hence the covariance structure, does not vary over time. By definition, the eigenvalues of the spectral density can be used to study the sampling properties of the power spectra of the sample dynamic FPCs. Accordingly, the null hypothesis that the spectral density does not change over time can be evaluated by performing inference on whether the eigenvalue spectra of a particular dynamic FPC vary over time. The reference distribution of such test is constructed by splitting each of the simulated time series into non-overlapping blocks of the same width $\mathcal{N}$ then, for each simulated data, evaluating the local spectral density and the corresponding eigenvalues over the frequencies $\theta \in [-\pi, \pi]$ for each time block individually. To perform a general test at all frequencies θ, the average of a particular eigenvalue m over all frequencies $\theta \in (-\pi, 0) \cup (0, \pi)$, referred to as general eigenvalue, is obtained for each time block and each simulated data. With analogy to the stationarity test by Brcich and Iskander (2006), the frequencies $\theta = -\pi, 0, \pi$ were excluded from the general test since the distributional results of the spectral density at the frequencies $\theta = -\pi, 0, \pi$ under H_0 are different from that at any other frequency $\theta \in (-\pi, 0) \cup (0, \pi)$. A $(1-\alpha)100\%$ reference interval for the m^{th} general eigenvalue is then obtained by computing the $\alpha/2$ and $1-\alpha/2$ quantiles of the simulated sequences of that general eigenvalue calculated across the sequence of time blocks, to highlight where the time sequence of that eigenvalue are expected to lie if the null hypothesis is true. Similarly, the observed signal is divided into the same number of time blocks and the same procedure of calculating the eigenvalues of the spectral densities for each time block is performed. The sequence of the observed m^{th} eigenvalue averaged over all frequencies $\theta \neq -\pi, 0, \pi$ is then plotted against time with the 95% reference bands. The spectral density is considered to vary significantly over time if the observed sequence of eigenvalues are exceeding the 95% reference bands of the null distribution. The test can be equivalently performed at a particular frequency of interest θ if desired, this is useful in studying the frequencies responsible for the largest variations and assessing the changes in the spectra at those particular frequencies.

It is also advantageous to have a test statistic and a corresponding p-value to evaluate the null hypothesis that the spectral density in general, or its eigenvalues in particular, do not vary over time. Following Miller and Bowman (2012), a proposed test statistic for assessing whether or not an eigenvalue of a particular dynamic FPC m changes at a certain frequency θ over time is obtained by calculating the absolute distance between the corresponding estimated eigenvalue of the local spectral density at each time block b of the observed time series at that particular frequency, denoted by $\hat{\gamma}_{mb}(\theta)$, and the corresponding estimated eigenvalue of the overall spectral density obtained for the full observed time series (ignoring blocks) at the same frequency, denoted by $\hat{\gamma}_m(\theta)$. The test statistic for the m^{th} eigenvalue at frequency θ is calculated by aggregating these distances across the whole grid of time blocks as follows:

$$\Lambda_m(\theta) = \sum_{b=1}^{B} |\hat{\gamma}_{bm}(\theta) - \hat{\gamma}_m(\theta)|, \qquad \theta \in [-\pi, \pi]. \tag{5.3}$$

To construct the distribution of the test statistic when the null hypothesis is true, the same process is repeated for all the bootstrap samples and the following quantity is calculated for each bootstrapped time series:

$$\Lambda^{\dagger}_m(\theta) = \sum_{b=1}^{B} |\gamma^{\dagger}_{bm}(\theta) - \gamma^{\dagger}_m(\theta)|, \qquad \theta \in [-\pi, \pi],$$

where $\gamma^{\dagger}_{bm}(\theta)$ is the m^{th} eigenvalue of the localized spectral density at block b of the bootstrapped sample of functions at frequency θ and $\gamma^{\dagger}_m(\theta)$ is the m^{th} eigenvalue of the overall spectral density of the bootstrapped functional data at the same frequency θ. The p-value is then computed as the proportion of times the value $\Lambda^{\dagger}_m(\theta)$ is greater than or equal to the value $\Lambda_m(\theta)$.

To obtain an overall p-value for assessing whether the eigenvalue at all frequencies collectively varies over time, the averages of $\hat{\gamma}_{bm}(\theta)$ and $\hat{\gamma}_m(\theta)$ over all $\theta \neq -\pi, 0, \pi$, denoted by $\bar{\bar{\gamma}}_{bm}$ and $\bar{\bar{\gamma}}_m$ respectively, are obtained and the following test statistic is computed:

$$\bar{\Lambda}_m = \sum_{b=1}^{B} |\bar{\bar{\gamma}}_{bm} - \bar{\bar{\gamma}}_m|. \tag{5.4}$$

Then for each simulated data, $\bar{\Lambda}^{\dagger}_m$ is calculated as $\sum_{b=1}^{B} |\bar{\gamma}^{\dagger}_{bm} - \bar{\gamma}^{\dagger}_m|$ where $\bar{\gamma}^{\dagger}_{bm}$ and $\bar{\gamma}^{\dagger}_m$ are the averages of $\gamma^{\dagger}_{bm}(\theta)$ and $\gamma^{\dagger}_m(\theta)$ over all $\theta \in (-\pi, 0) \cup (0, \pi)$, respectively. The p-value of the overall test statistic is calculated as the probability that the value $\bar{\Lambda}^{\dagger}_m$ is greater than or equal to the value $\bar{\Lambda}_m$.

5.2.2 Testing Stationarity of the Daily EpCO$_2$ Functional Time Series

Consider the functional time series of EpCO$_2$ daily curves described at length in Chapters 3 and 4; and suppose that we want to test the stationarity of this time series using the above proposed test. The reference distribution of the test is constructed, using the parametric bootstrap, by simulating 200 functional data sets according to the FAR(1) model given by Equation 5.1. By letting $\{X_i\}$ be the sequence of daily functions of DD EpCO$_2$, $\{\psi_k\}_{k=1}^{15}$ be the set of 15 B-splines basis functions used to approximate the curves $\{X_i\}$ and $\mathbf{a_i}$ be the vector of corresponding penalized basis coefficients; the following VAR(1) model:

$$\begin{bmatrix} a_{i,1} \\ a_{i,2} \\ \vdots \\ a_{i,15} \end{bmatrix} = \begin{bmatrix} \langle \Upsilon(\psi_1), \psi_1 \rangle & \langle \Upsilon(\psi_1), \psi_2 \rangle & \dots & \langle \Upsilon(\psi_1), \psi_{15} \rangle \\ \langle \Upsilon(\psi_2), \psi_1 \rangle & \langle \Upsilon(\psi_2), \psi_2 \rangle & \dots & \langle \Upsilon(\psi_2), \psi_{15} \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \Upsilon(\psi_{15}), \psi_1 \rangle & \langle \Upsilon(\psi_{15}), \psi_2 \rangle & \dots & \langle \Upsilon(\psi_{15}), \psi_{15} \rangle \end{bmatrix} \begin{bmatrix} a_{i-1,1} \\ a_{i-1,2} \\ \vdots \\ a_{i-1,15} \end{bmatrix} + \begin{bmatrix} \varepsilon^*_{i,1} \\ \varepsilon^*_{i,2} \\ \vdots \\ \varepsilon^*_{i,15} \end{bmatrix}$$

can be fitted to the basis coefficients $\mathbf{a_i}$ to estimate: the auto-regressive parameters matrix $\mathfrak{R} = (\langle \Upsilon(\psi_k), \psi'_k \rangle : k \geq 1, k' \leq 15)$, such that all the eigenvalues of $\mathfrak{R}$ are less than 1 in modulus to ensure stationarity, and the variance-covariance matrix Σ of the white noise $\boldsymbol{\varepsilon}^*$. Thus, to construct the reference information for the test, data (curves) are simulated by generating their corresponding ($p = 15$) B-splines basis coefficients according to the VAR(1) model given by Equation 5.2 based on the estimated $\mathfrak{R}$ matrix of auto-regressive parameters; where the randomness in the curves is introduced by simulating the errors $\boldsymbol{\varepsilon_i}^*$ independently from MVN$(0, \Sigma)$.

Next, as explained in Section 5.2.1, the observed and each of the simulated functional time series are divided into non-overlapping segments of fixed width. The width of each segment is set equal to $2^5 = 32$, which seems to be a plausible choice (see Section 5.1). The functional data is then assumed to be stationary within each segment and the spectral density of each time segment is estimated using the Bartlett kernel with bandwidth $Q = \lfloor \sqrt{\mathcal{N}} \rfloor = \lfloor \sqrt{32} \rfloor = 5$. To assess whether the spectral density at all frequencies changes over time, the test statistic $\bar{\Lambda}_m$ and the corresponding bootstrap p-value are computed for $m = 1, 2$, producing p-values less than 0.001. This result implies that the spectra of the first 2 dynamic FPCs and, roughly speaking, the spectral density are statistically significantly changing over time across all frequencies on average. Figure 5.8, showing the sequence of $\bar{\hat{\gamma}}_{bm}$ against the time block b along with the 95% bootstrap reference bands, emphasizes the non-stationarity in the spectra of the first 2 dynamic FPCs.

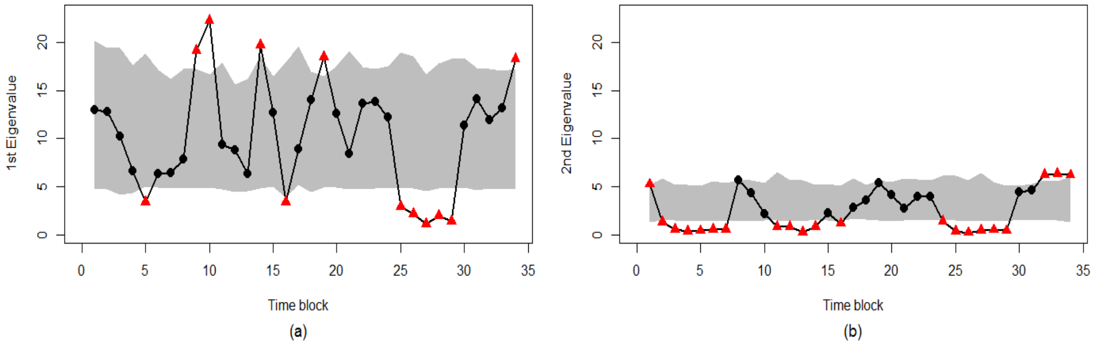


Figure 5.8: Plots of the (a) first and (b) second eigenvalues across the different time blocks based on the local spectral density estimated within each block individually with the 95% bootstrap reference bands displayed (▲, outside a 95% reference band; ●, within a 95% reference band).

It is also desired to investigate how the eigenvalues vary over time at the different time frequencies. Figure 5.9 displays the p-values of the test statistic $\Lambda_m(\theta)$ for $m = 1, 2$ at all frequencies $\theta \in [-\pi, \pi]$. It is evident that the dominant changes in the spectral characteristics are attributed to the high frequencies. Figures 5.10 and 5.11 illustrate the observed first 2 eigenvalues $\hat{\gamma}_{bm}(\theta), m = 1, 2$ estimated for each time block $b = 1, \ldots, 34$ at a particular set of frequencies $\theta = \frac{\pi}{8}, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}$ along with the corresponding 95% reference bands computed based on the simulated data. The figures show that the eigenvalues are significantly varying over time at all the selected frequencies, but the variations seem to be highly significant at the higher frequencies although they have less (spectral) power relative to the lower frequencies components.

In conclusion, it is evident that the spectral density of the EpCO$_2$ daily curves is changing over time and is not fixed over the whole study period. Consequently, the dynamic FPCs described earlier in Chapter 4 might not offer an adequate representation of the data as they do not adapt to the changes in the process underlying the data, relying on the weak stationarity assumption of the process. Therefore, a time-varying version of the dynamic FPCs is proposed in the following section.

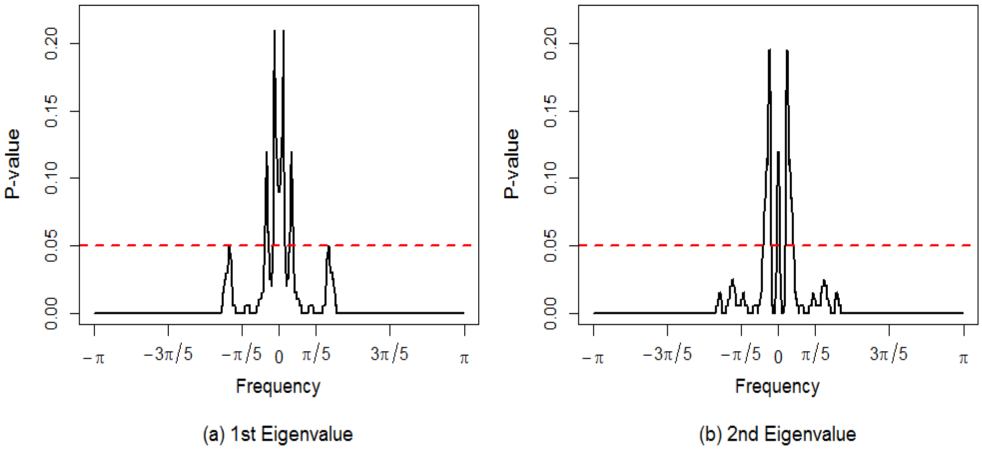


FIGURE 5.9: Plots of the p-values of the test statistic calculated by Equation 5.3 for the (a) first and (b) second eigenvalues. The red dashed line represents the 5% level of significance.

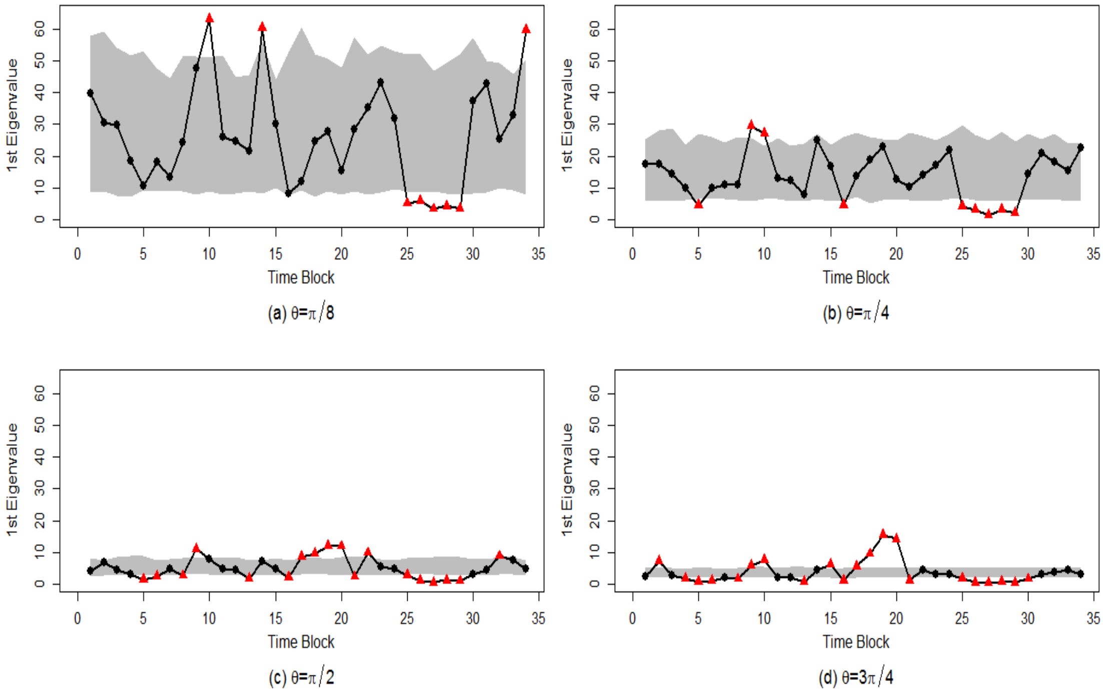


FIGURE 5.10: Plots of the first eigenvalue across the different time blocks based on the local spectral density estimated within each block individually and evaluated at the frequencies $\theta =$ (a) $\frac{\pi}{8}$, (b) $\frac{\pi}{4}$, (c) $\frac{\pi}{2}$ and (d) $\frac{3\pi}{4}$ along with the corresponding 95% bootstrap reference bands (▲, outside a 95% reference band; ●, within a 95% reference band).

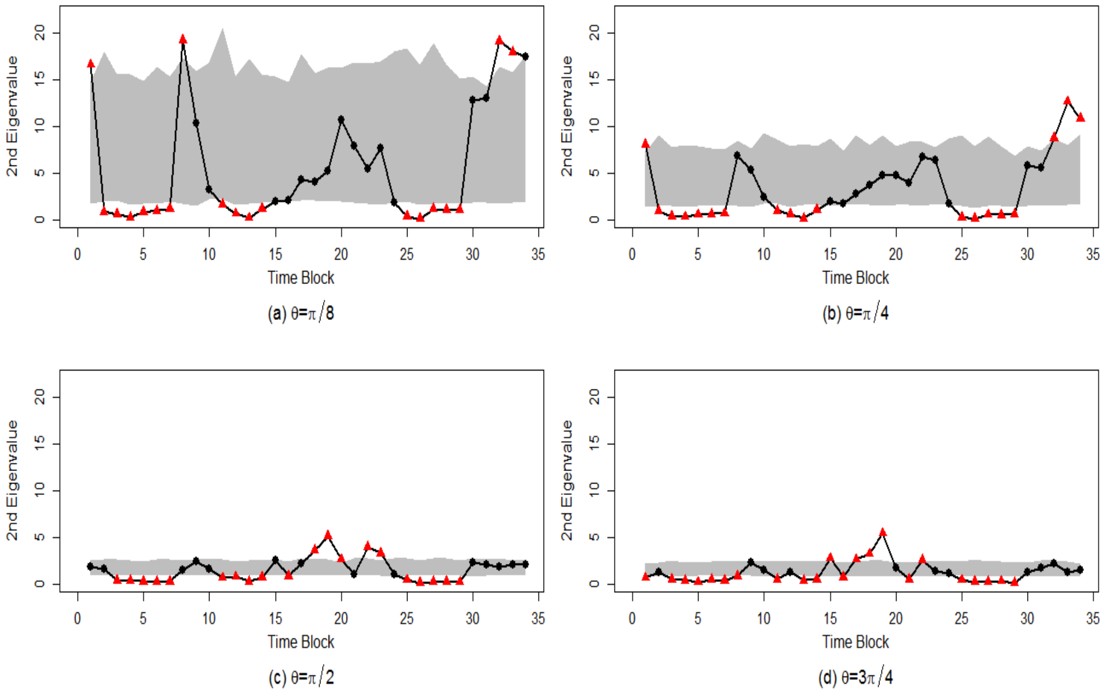


FIGURE 5.11: Plots of the second eigenvalue across the different time blocks based on the local spectral density estimated within each block individually and evaluated at the frequencies $\theta =$ (a) $\frac{\pi}{8}$, (b) $\frac{\pi}{4}$, (c) $\frac{\pi}{2}$ and (d) $\frac{3\pi}{4}$ along with the corresponding 95% bootstrap reference bands (▲, outside a 95% reference band; ●, within a 95% reference band).

5.3 Smooth Dynamic Functional Principal Components

For non-stationary functional time series, a time-varying version of the dynamic FPCA is appealing to provide a more adequate data reduction that simultaneously accounts for the autocorrelation between the functional data and the time-changing covariance structure. Here, we propose a smooth dynamic FPCA, in which the dynamic FPCs are obtained at each time point i by smoothing all the lag h covariance matrices. Because of the limited number of observations at each time point i (one functional observation per time point), a weighting function is proposed to estimate the lag h covariance matrix at each time point i, where the amount of neighbouring data contributing to the estimated covariance is controlled by the choice of the weighting kernel and the smoothing parameter.

To obtain the covariance structure at each time point i, a smooth covariance matrix for each lag h is estimated at this time point using a weighting kernel that assigns higher weights to the nearby observations and less weights to the further ones. Accordingly, the estimated lag h covariance matrix in terms of the functions' basis coefficients obtained at the time point i, assuming that the data is centered, can be defined by:

$$\begin{aligned} \hat{V}_h^{\mathbf{a}}(i) &= \frac{1}{\sum_{i'=1}^{N-h} \omega_{i,i'}} \sum_{i=1}^{N-h} \mathbf{a}_{i'} \omega_{i,i'} \mathbf{a}_{i'+h}^\top, && h \geq 0 \\ \hat{V}_h^{\mathbf{a}}(i) &= (\hat{V}_{-h}(i))^\top, && h < 0, \end{aligned} \tag{5.5}$$

where $\mathbf{a}_{i'}$ and $\mathbf{a}_{i'+h}$ are the vectors of basis coefficients that correspond to the functional observations i' and $i'+h$, respectively; and $\omega_{i,i'}$ is the value of the weighting kernel assigned to the pair of observations i' and $i'+h$, based on the distance between i' and i. The weighting kernel $\omega(.)$ is a monotonically decreasing function of the distance $|i-i'|$ regardless of h, i.e. $\omega_{i,i'} = \omega(|i-i'|)$ such that $i \leq N-h$, ensuring that the highest weights are given to the observations near the target point i. Here, the weighting kernel $\omega(.)$ is taken to be a Gaussian density with mean 0 and standard deviation s, that is:

$$\omega(i-i', s) = \exp\{-\frac{1}{2}\Big(\frac{i-i'}{s}\Big)^2\}. \tag{5.6}$$

The standard deviation s is the smoothing parameter controlling the width of the kernel and the degree of smoothing, by determining the amount of neighbouring data contributing to the estimated lag h covariance. Clearly, according to Equation 5.6, the weight $\omega(i-i', .) = 1$ for $i = i'$ and monotonically decreases towards zero as the distance $|i-i'|$ increases. Alternative weight functions can be used based on the nature of the variable of interest. In matrix notation, Equation 5.5 can be written equivalently as:

$$\hat{V}_h^{\mathbf{a}}(i) = \frac{1}{\sum_{i'} \omega_{i,i'}} A^\top \Omega_i A_h, \tag{5.7}$$

where A is an $(N-h) \times p$ matrix of basis functions' coefficients, such that the first row corresponds to the first functional observation and the last row corresponds to the $(N-h)^{th}$ observation; and A_h is also an $(N-h) \times p$ matrix of basis coefficients, where the first row corresponds to the $(1+h)^{th}$ observation and the last row corresponds to the N^{th} functional observation in the series. Ω_i is an $(N-h) \times (N-h)$ diagonal matrix containing the vector of weights $\omega_{i,i'}, i = 1, \dots, N-h$, obtained from Equation 5.6 for smoothing the lag h covariance at time i.

After estimating the covariance structure $\hat{V}_h^{\mathbf{a}}(i), |h| \leq Q$, for $i = 1, \dots, N-Q$, the spectral density is estimated at each of those time points using the Bartlett lag window estimator, with bandwidth $Q = \lfloor \sqrt{N} \rfloor$, as follows:

$$\hat{\mathfrak{F}}_\theta^{\mathbf{a}}(i) = \frac{1}{2\pi} \sum_{|h| \leq Q} w\Big(\frac{h}{Q}\Big) \hat{V}_h^{\mathbf{a}}(i) \exp(-\mathrm{i}h\theta), \tag{5.8}$$

where $w(\frac{h}{Q}) = 1 - |\frac{h}{Q}|$. Next, define the Hermitian matrix $\hat{\mathfrak{F}}_\theta^X(i) = \mathbf{W}^{1/2} \hat{\mathfrak{F}}_\theta^{\mathbf{a}}(i) \mathbf{W}^{1/2\top}$

($\mathbf{W} = \int \psi(t)\psi(t)^\top dt$ and $\mathbf{W}^{1/2}$ is the Cholesky decomposition of $\mathbf{W}$) and calculate the corresponding real eigenvalues $\hat{\lambda}_m(i, \theta)$ and Hermitian eigenvectors $\hat{\varphi}^*_m(i, \theta)$, $m \geq 1$. The eigenfunctions $\hat{\varphi}_m(i, \theta)$ of the spectral density operator corresponding to the kernel $\hat{\mathfrak{F}}^{\mathbf{a}}_\theta(i)$ are then obtained by $(\mathbf{W}^{1/2})^{-1}\hat{\varphi}^\star_m(i, \theta)$. The time domain functional filters for the lags $l = -L, \ldots, L$ can then be obtained at each time point through the following numerical integration:

$$\hat{\phi}_{ml}(i) = \frac{\boldsymbol{\psi}^\top}{2\pi(2N_\theta + 1)} \sum_{j=-N_\theta}^{N_\theta} \hat{\varphi}_m\Big(i, \frac{\pi j}{N_\theta}\Big)\exp\Big(-\mathrm{i}l\frac{\pi j}{N_\theta}\Big) = \boldsymbol{\psi}^\top \hat{\tilde{\boldsymbol{\phi}}}_{ml}(i) \qquad (N_\theta >> 1), \tag{5.9}$$

where $\hat{\tilde{\boldsymbol{\phi}}}_{ml}(i)$ are the corresponding basis coefficients of the functional filters $\hat{\phi}_{ml}(i)$.

Subsequently, the scores of the m^{th} smooth dynamic FPCs are estimated at $i = -L + 1, \ldots, N - L$ by:

$$\hat{\mathrm{Z}}_{mi} = \sum_{l=-L}^{L} a_{i-l}^\top \mathbf{W} \hat{\tilde{\boldsymbol{\phi}}}_{ml}(i). \tag{5.10}$$

For the ends of the series $1 \leq i \leq L$ or $N - L + 1 \leq i \leq N$, set $X_{-L+1} = \ldots = X_0 = X_{N+1} = \ldots = X_{N+L} = \mathbb{E}[X_1]$. With analogy to the dynamic FPCs, padding the ends of the series by the functional mean creates some bias on the boundary of the observation period.

Finally, the original curves can be approximated using the q-term time-varying (smooth) dynamic FPCs as follows:

$$\hat{X}_i = \sum_{m=1}^{q} \sum_{l=-L}^{L} \hat{\mathrm{Z}}_{m,i+l}\hat{\phi}_{ml}(i). \tag{5.11}$$

For $i \in \{-L + 1, \ldots, 0\} \cup \{N + 1, \ldots, N + L\}$, set $\hat{\mathrm{Z}}_{mi} = 0$ if the functional data are originally centered or set $\hat{\mathrm{Z}}_{mi} = \bar{\hat{\mathrm{Z}}}_{m.}$ if the data are not centered. Note that with $L < Q$, which is usually the case, we are only able to reconstruct the curves $i = 1, \ldots, N - Q$ for which the whole family of lag h covariances ($|h| \leq Q$) is available.

5.4 Modified Stationarity Test for Functional Time Series

According to the above methodology of time-varying dynamic FPCs, the eigenvalues of the smooth dynamic FPCs are available at each time point i. Following from this, the eigenvalues' sequences of the smooth dynamic FPCs can be used to assess whether or not the spectral density change throughout time. A test statistic, similar to the aforementioned block stationarity test statistic, is calculated at each time point rather than within blocks using the eigenvalues of the weighted spectral density estimate obtained at each time point. Under the null hypothesis of no change in the spectral density over time, the reference distribution of the test is constructed by: first simulating a number of functional data sets using the same parametric bootstrap approach explained before, then calculating the eigenvalues of the time-varying spectral density evaluated at each time point for each simulated data.

A general test at all frequencies θ, concerning the eigenvalue of a particular PC m, relies on computing the average of that particular eigenvalue across all frequencies $\theta \in (-\pi, 0) \cup (0, \pi)$ at each time point for each simulated data and estimating the corresponding point-wise 95% reference band. For the sequence of observed functional curves, the same procedure is employed and the sequence of the m^{th} eigenvalue averaged over the same set of frequencies is compared to the 95% bootstrap reference band. The null hypothesis of the test is rejected if the observed sequence of the m^{th} eigenvalue exceeded its corresponding 95% reference band. The corresponding general test statistic used to evaluate whether or not the eigenvalue of the m^{th} smooth dynamic FPC change over time on average is defined by:

$$\bar{\Lambda}_m^{(\mathrm{sm})} = \sum_{i=1}^{N} |\bar{\hat{\gamma}}_m^{(\mathrm{sm})}(i) - \bar{\hat{\gamma}}_m|, \tag{5.12}$$

where $\bar{\hat{\gamma}}_m^{(\mathrm{sm})}(i)$ is the average of the m^{th} eigenvalue of the time-varying spectral density evaluated at time i, $\hat{\gamma}_m^{(\mathrm{sm})}(i, \theta)$, across all frequencies $\theta \in (-\pi, 0) \cup (0, \pi)$ and $\bar{\hat{\gamma}}_m$ is the corresponding m^{th} eigenvalue of the overall spectral density obtained for the whole time series and averaged over the same set of frequencies. After calculating a similar quantity $\bar{\Lambda}_m^{\dagger(\mathrm{sm})}$ for each bootstrapped functional time series, the p-value of the test is estimated as the proportion of times $\bar{\Lambda}_m^{\dagger(\mathrm{sm})}$ is greater than or equal $\bar{\Lambda}_m^{(\mathrm{sm})}$. If desired, the test can also be performed at a particular frequency θ and a similar test statistic can be calculated individually for this frequency θ to asses if the power of a particular principal component m at a particular frequency of interest θ varies over time.

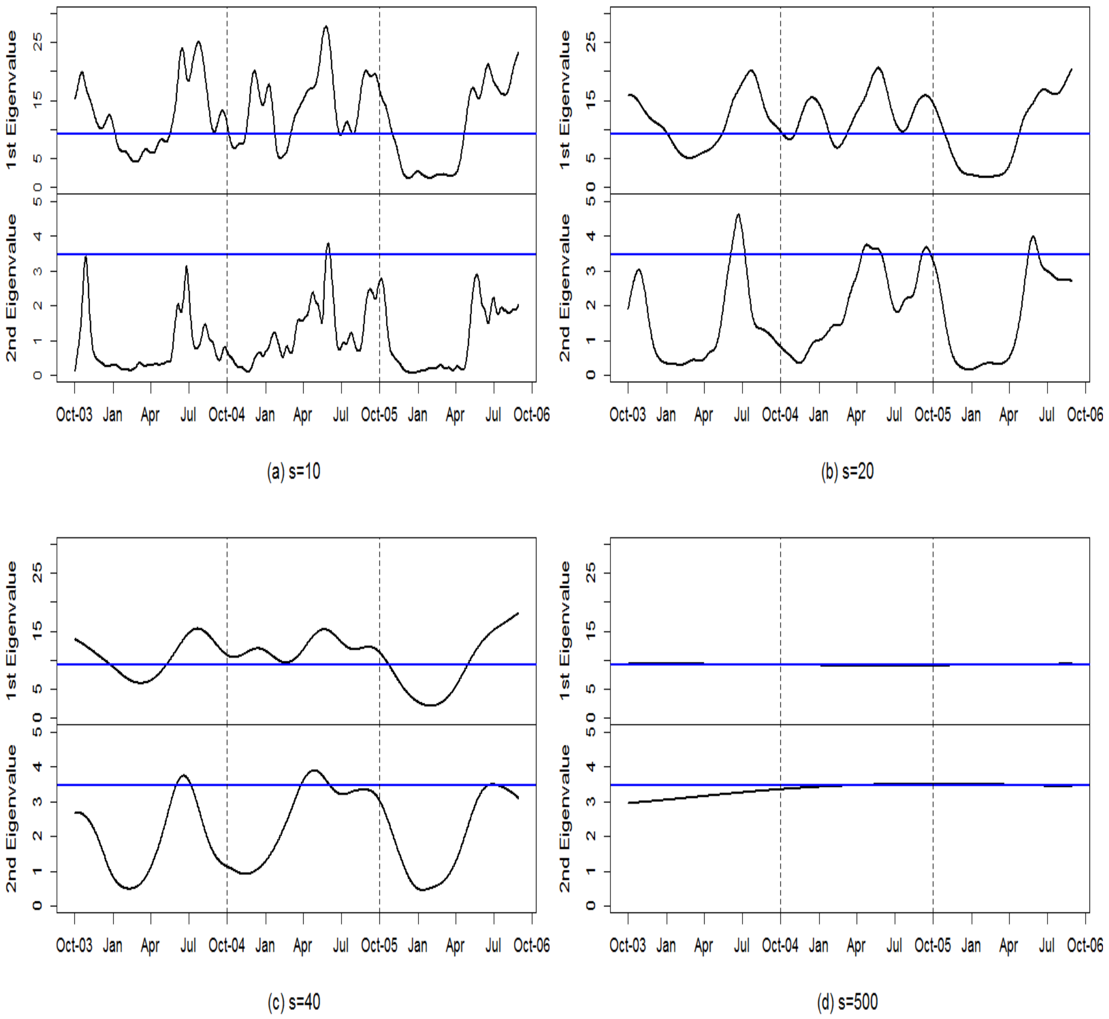
5.5 Application of Smooth Dynamic FPCA to the EpCO$_2$ Data

The preliminary block stationarity test proposed in Section 5.2.1 has shown that the spectral density of the EpCO$_2$ daily curves is varying over time (see Section 5.2.2). This result implies that the covariance structure and the modes of variability in the data are not the same for the whole time series. We, therefore, need to employ the smooth dynamic FPCA, presented in Section 5.3, to find the directions of maximum variability at each time point and assess whether they change over time using the test proposed in Section 5.4. These smooth dynamic FPCs can then be used to provide a better dimension reduction representation for the daily curves of EpCO$_2$ by accounting for both the serial correlation between the curves and the non-stationarity over time.

At each time point, the smooth covariance matrix for lag h, $|h| \leq Q$, is estimated using Equation 5.5 with the Gaussian weighting kernel given by Equation 5.6. The width of this Gaussian kernel is determined by a pre-specified smoothing parameter (standard deviation) which controls the amount of neighbouring data contributing to the estimation of lag h covariance matrix at this time point. The smooth spectral density is then estimated as the Fourier transform of the product of the smooth covariance structure and the Bartlett lag window function with bandwidth $Q = 32$ at each time point. The tuning parameter Q is set equal to 32 to allow for a fair comparison between the smooth and original dynamic FPCs. A more sophisticated calibration for the choice of Q can be used to obtain better results; however in such an environmental system correlation between observations is not expected to go beyond a month. From the spectral density estimate obtained at each time point, the frequency domain eigenvalues and the corresponding eigenfunctions are computed, and the functional filters $\phi_{ml}(i)$ for $l = -L, \ldots, L$ are estimated individually for each time point i. In dynamic FPCA, the maximum L is chosen such that such that $\sum_{-L \leq l \leq L} ||\hat{\phi}_{ml}||^2 \geq 1 - \epsilon$, for some small threshold ϵ, e.g. $\epsilon = 0.01$. Here, in smooth dynamic FPCA, the maximum lag L is set equal to 30 for which the average of $\sum_{l=-L}^{L} ||\hat{\phi}_{1l}(i)||^2$ across all time points i is approximately 0.99.

A sensitivity analysis approach is initially performed to determine the most appropriate value for the smoothing parameter of the Gaussian weighting kernel used to estimate the lag h covariance matrices. Figure 5.12 displays only the eigenvalues of the first 2 smooth dynamic FPCs, averaged over all frequencies $\theta \in (-\pi, 0) \cup (0, \pi)$, for different values of the smoothing parameter. As expected, the changes in a particular eigenvalue become smoother over time and approach the corresponding (overall) dynamic FPC's eigenvalue as the smoothing parameter increases. A small smoothing parameter (like s =10) seems to be sensitive to the very local and fine variations, while a considerably large smoothing parameter tends to miss the potential variability over time (see s ≥40). It is also noticed that by smoothing the covariance structure and allowing the dynamic FPCs to change over time, the first FPC tends to account for more variability relative to the subsequent FPCs (see the sequence of the second time-varying eigenvalues is below the eigenvalue of the second overall dynamic FPC).

Using different smoothing parameter values, the daily curves of the DD EpCO$_2$ are reconstructed based on the first 2 smooth dynamic FPCs. The smoothing parameter that minimizes the distance between the original curves and the reconstructed ones



(a) s=10

(b) s=20

(c) s=40

(d) s=500

FIGURE 5.12: Plot of the first 2 smooth dynamic FPCs' eigenvalues obtained at each time point and averaged over all frequencies $\theta \in (-\pi, 0) \cup (0, \pi)$ using different values of the smoothing parameter (a) $s = 10$, (b) $s = 20$, (c) $s = 40$ and (d) $s = 500$. The blue horizontal lines corresponds to the eigenvalue of the first 2 (overall) dynamic FPC averaged over the same set of frequencies θ.

is considered a good choice. More principal components can be considered for the reconstruction of the curves, but the first 2 PCs prove to account for the largest portion of variability (more than 90%) at any value of the smoothing parameter. Figure 5.13 displays a sample of 10 original smooth DD EpCO$_2$ daily curves (top left) and the corresponding reconstructed curves with the smooth dynamic FPCs using a wide range of values for the smoothing parameter, from which it is evident that using a very small standard deviation over-estimates the curves, while using a larger standard deviation overlooks the local variability in the curves. Figure 5.14, shows the NMSE, calculated using Equation 4.18, between the original smooth curves and the reconstructed ones at the different smoothing parameter values. Both figures indicate that a standard deviation of 20 provides a good approximation for the original curves with a NMSE of

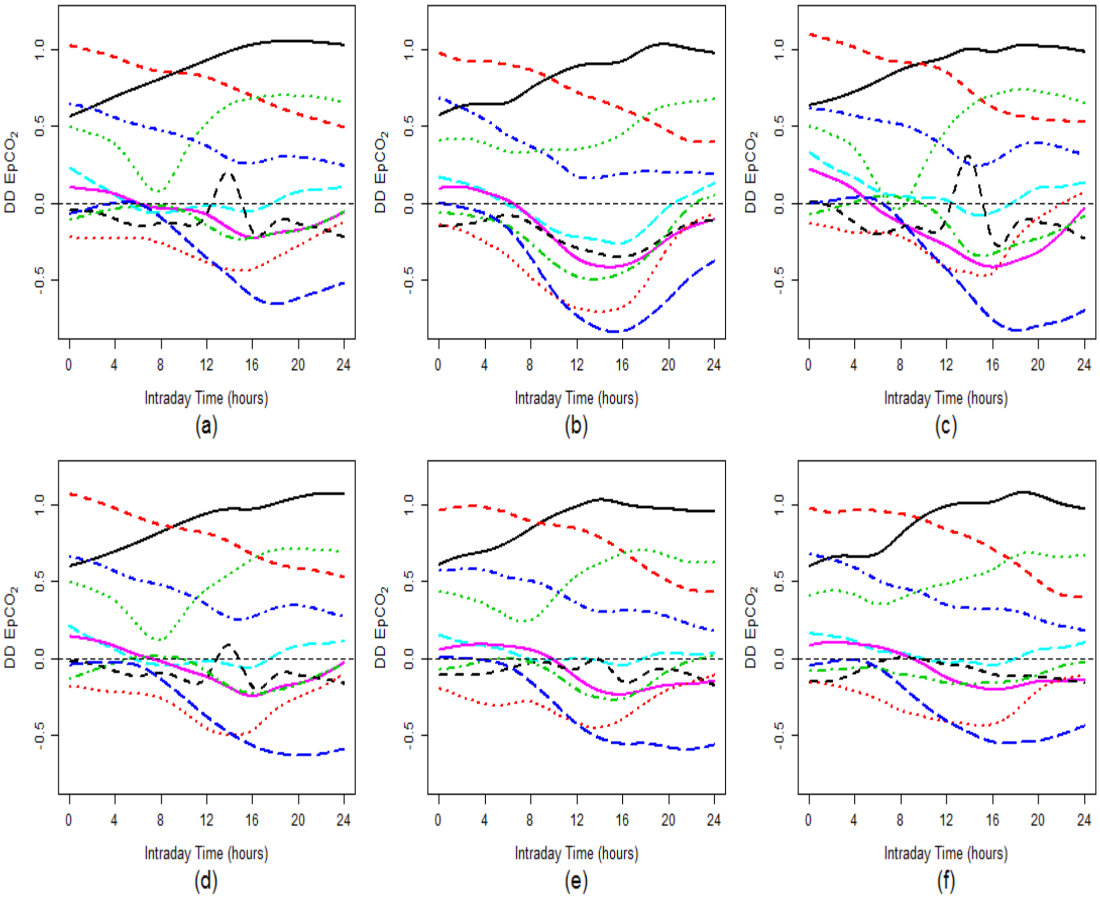


FIGURE 5.13: (a) 10 successive daily smooth curves and the corresponding reconstructions based on (b) the first 2 dynamic FPCs and (c-e) the first 2 smooth dynamic FPCs using the smoothing parameters $s = 10, 20, 40$ and 500, respectively.

3.4%.

Using a smoothing parameter equal to 20 for the Gaussian weighting kernel, the first 2 smooth dynamic FPCs collectively explain (1−NMSE) $*\, 100 \approx 96.5\%$, and the first component solely accounts for 94% of the variability. Moreover, Figure 5.15 shows that the first dynamic FPC scores and the corresponding smooth scores are close to one another, although the smooth dynamic FPC appears to account for more variability. In contrast, the second smooth dynamic FPC seems to account for a smaller amount of the total variations, contrary to its overall counterpart. Therefore, only the first smooth dynamic FPC, which seems to take over the role of explaining the major variations in the data, is retained.

Figure 5.16 displays the 3 central filter elements of the first smooth dynamic FPC, $\hat{\phi}_{1l}(i), l = -1, 0, +1, i = 1, \ldots, N-Q$, from which it is obvious that the effect of the filters changes throughout time. This implies that the consecutive effect of the corresponding dynamic scores varies over time. It also appears from Figure 5.17, displaying the sum of squared norm $\sum_{l=-1}^{1} ||\hat{\phi}_{1l}(i)||^2$ at $i = 1, \ldots, N-Q$, that the 3 central filters are not

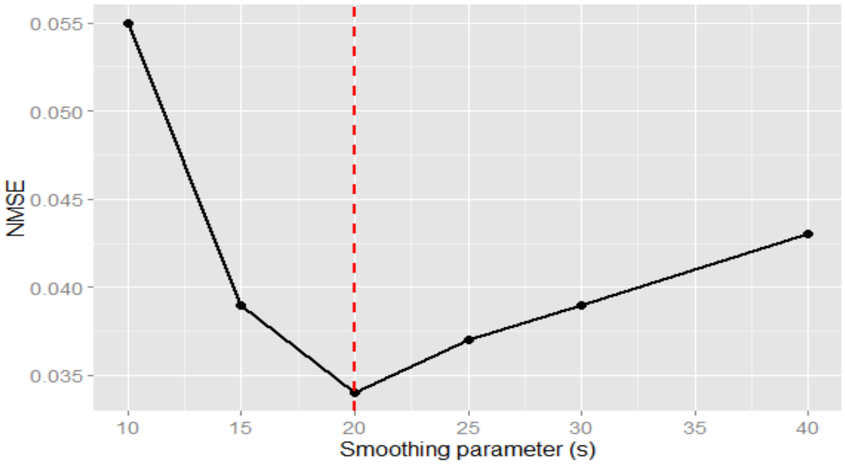


FIGURE 5.14: Normalized mean square error between the original smooth curves and the curves reconstructed based on the first 2 smooth dynamic FPCs calculated at different values of the standard deviation (smoothing parameter) of the Gaussian weighting kernel.

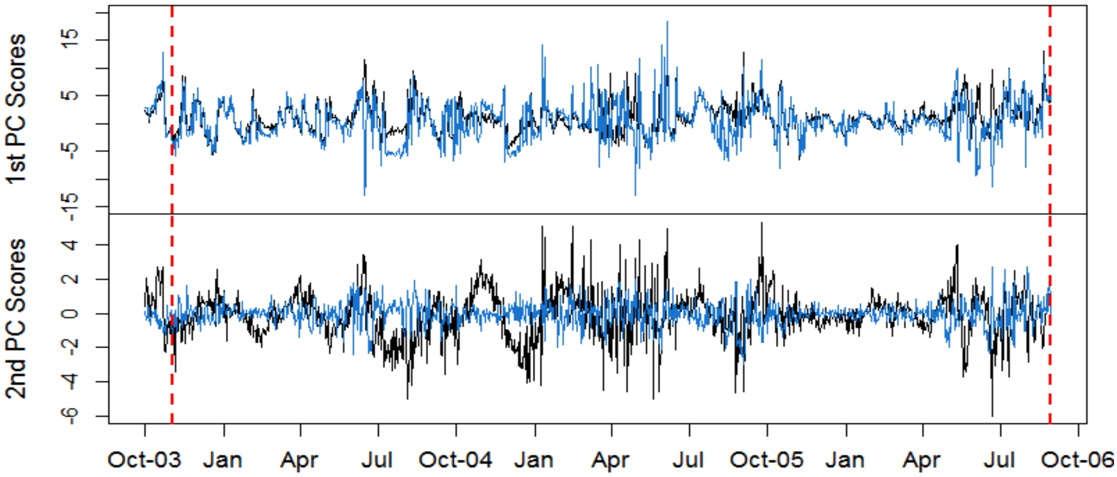


FIGURE 5.15: Time plots of the first (top) and second (bottom) dynamic FPCs' scores (black) and the corresponding smooth dynamic FPCs' scores (blue) obtained for $i = 1, \ldots, (1095 - 32)$. The red dashed vertical lines indicate the boundary areas of biased scores.

ultimately the most dominant filters for all time points $(\sum_{l=-1}^{1} ||\hat{\phi}_{1l}(i)||^2 << 0.9$ for some time points i). This implies that the effect of the first PC in some time periods is distributed over a wider range of lags.

Following the estimation of the smooth dynamic FPCs, the sequence of eigenvalues corresponding to the leading (first) smooth dynamic FPC can be used to evaluate whether or not the spectral density changes over time. The null distribution of the test, that the spectral density of a particular dynamic FPC at all frequencies $\theta \in (-\pi, 0) \cup (0, \pi)$ does

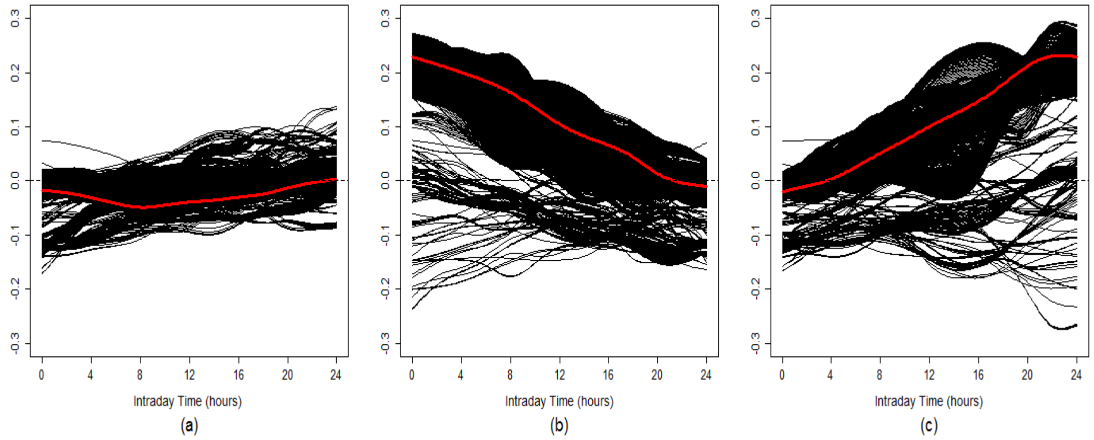


FIGURE 5.16: The 3 central functional filters of the first smooth dynamic FPC $\hat{\phi}_{1,-1}(i)$ (left), $\hat{\phi}_{1,0}(i)$ (middle), $\hat{\phi}_{1,+1}(i)$ (right), for $i = 1, \ldots, N - Q$. The red lines are the corresponding 3 central filters elements of the first (overall) dynamic FPC at the lags -1,0 and 1.

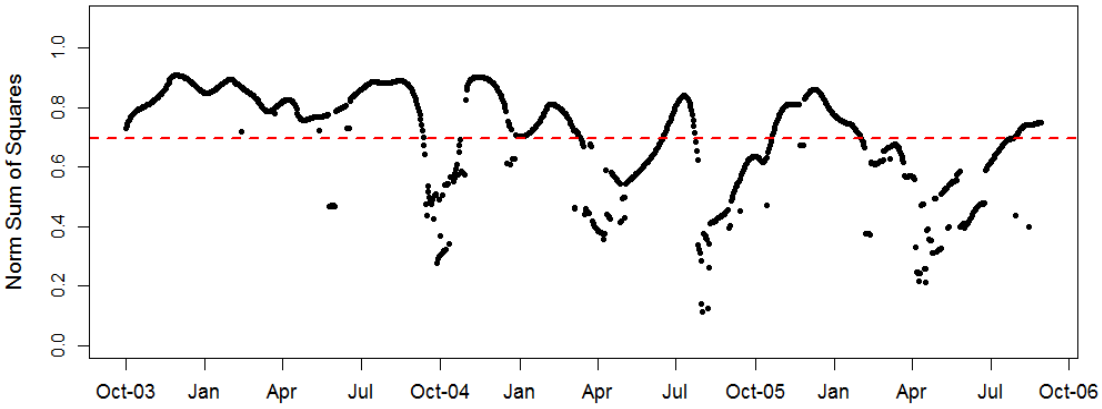


FIGURE 5.17: Plot of the norm sum of squares $\sum_{l=-1}^{1} \|\hat{\phi}_{1l}(i)\|^2$ at $i = 1, \ldots, N - Q$. The red dashed line is the mean norm sum of squares over all the time points i.

not change over time on average, is constructed using parametric bootstrap by simulating multiple data sets from the FAR(1) model explained at length in Section 5.2.1. Then for each simulated functional time series, having the same number of curves as the original data, the smooth covariance structure and spectral density are estimated at each time point using the Gaussian weighting kernel with the same smoothing parameter identified for the original time series. At each time point, the eigen-decomposition of the spectral density in the frequency domain is performed and the spectrum of the first eigenvalue over the range of frequencies $\theta \in (-\pi, \pi)$ is obtained. To perform the test at all frequencies θ, the average of the first eigenvalue over the whole set of frequencies $\theta \in (-\pi, 0) \cup (0, \pi)$, known as the general eigenvalue, is computed at each time point and the corresponding point-wise 95% reference band is estimated. Figure 5.18 shows the

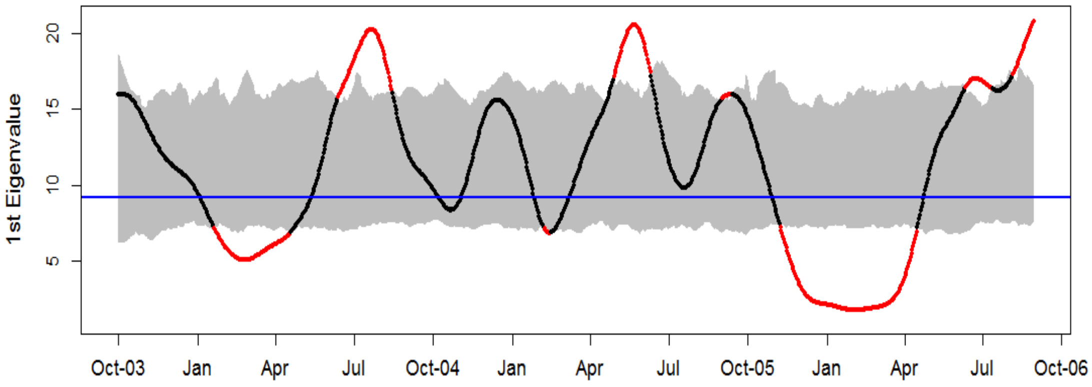


FIGURE 5.18: Plot of the first smooth dynamic FPC eigenvalues obtained at each time point and averaged over all frequencies $\theta \in (-\pi, 0) \cup (0, \pi)$ along with the corresponding 95% reference band. The red points highlight the periods exceeding the reference band. The blue horizontal line corresponds to the eigenvalue of the first (overall) dynamic FPC averaged over the same set of frequencies θ.

sequence of the first general eigenvalue across the whole time series and the corresponding 95% bootstrap reference band. It is clear that the spectrum of the first dynamic FPC significantly changes over time with higher values observed towards the end of the hydrological year (summer months) and considerably lower values during winter. Calculating the test statistic given by Equation 5.12 after performing the parametric bootstrap produces a p-value less than 0.001 emphasizing the significant changes in the first eigenvalue, on average, throughout time.

To investigate how the power of the first smooth dynamic FPC at a particular frequency θ changes over time, the same test procedure is applied for each frequency separately. Figure 5.19 illustrates the sequence of the first smooth dynamic FPC eigenvalues at the frequencies $\theta = \frac{\pi}{8}, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}$ and the corresponding point-wise 95% bootstrap reference bands. The Figure shows evidence of statistical significant changes in the power of the first smooth dynamic FPC at the different frequencies. Figure 5.20 maps the periods where the observed sequence of eigenvalues at each frequency θ in the range $[-\pi, \pi]$ exceeded the 95% reference band. The significant changes in the process are often observed in the period December – May, where the first smooth dynamic FPC has significantly lower spectral power (at all frequencies) than expected in all HYs except in the driest hydrological year 2004/2005 which has significantly larger power than expected in the period April – June.

The proposed smooth dynamic FPCA has proven to be a useful tool for assessing whether or not the spectral density, and hence the covariance structure, of a functional time series varies over time. By choosing an appropriate smoothing parameter for the weighting kernel used to smooth the covariance structure, the smooth dynamic FPCs appear to

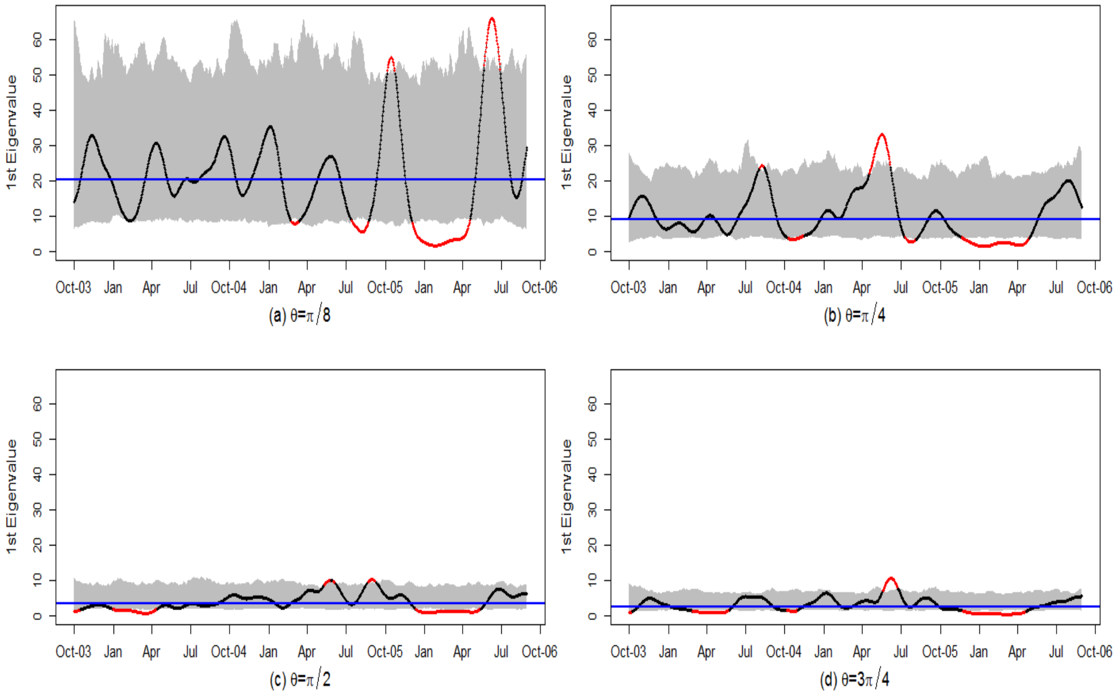


FIGURE 5.19: Plots of the the first smooth dynamic FPC eigenvalues obtained at each time point at the frequencies $\theta =$ (a) $\frac{\pi}{8}$, (b) $\frac{\pi}{4}$, (c) $\frac{\pi}{2}$ and (d) $\frac{3\pi}{4}$ along with the corresponding 95% reference bands. The red points highlight the periods exceeding the reference bands. The blue horizontal lines represent the corresponding eigenvalues of the first (overall) dynamic FPC at the same frequencies θ.

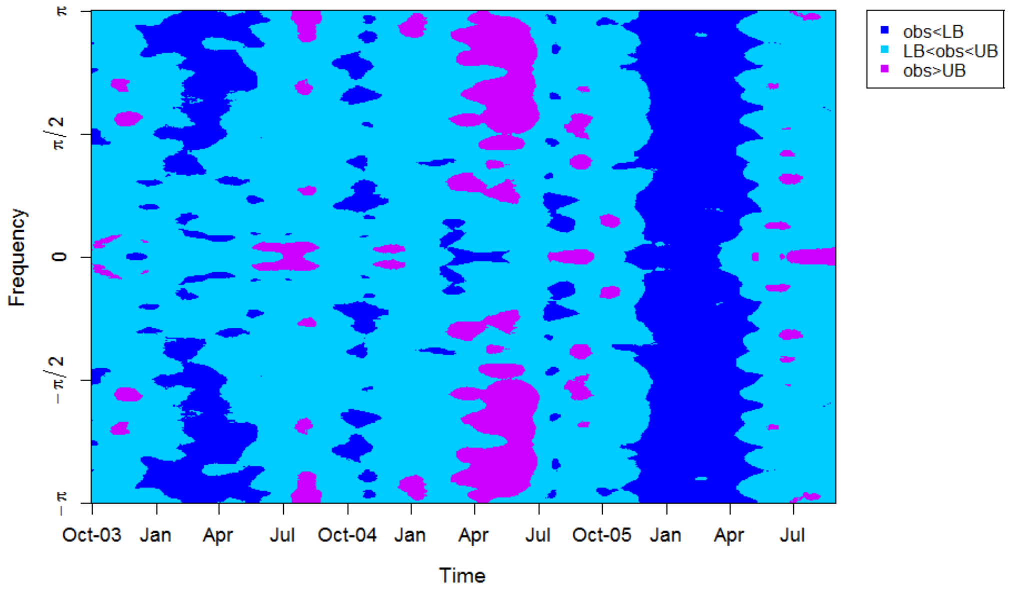


FIGURE 5.20: Plot of where the observed sequence of the first smooth dynamic FPC eigenvalues at each frequency $\theta \in [-\pi, \pi]$ exceeds the corresponding 95% reference band.

provide a good approximation to the original curves and a more adequate dimensionality reduction representation for the data. To statistically evaluate the performance of these smooth dynamic FPCs relative to the original dynamic FPCs proposed by Hormann et al. (2014), an extensive simulation study is performed. This simulation study is presented in full details in the following section.

5.6 Simulation Study

The main objective of this simulation study is to compare the performance of our proposed smooth dynamic FPCs with that of the original dynamic FPCs under a variety of data-generating processes. This simulation study aims at evaluating the performance of the smooth dynamic FPCA under both stationary and non-stationary conditions. In the simulation study, we assess, under some specific conditions, whether the smooth dynamic FPCs provide a better approximation for the original functional data than the stationary dynamic version. For each simulated functional time series, the dynamic FPCs and smooth dynamic FPCs are estimated and the original functional time series is recovered using both the first q dynamic FPCs and the first q smooth dynamic FPCs. The performances of these approximations are then measured and compared, in terms of their corresponding normalized mean squared errors (NMSE) computed by:

$$\begin{gathered}\sum_{i=1}^{N-Q} \|X_i - \hat{X}_i^{\text{dyn}}(q)\|^2 \Big/ \sum_{i=1}^{N-Q} \|X_i\|^2, \\ \sum_{i=1}^{N-Q} \|X_i - \hat{X}_i^{\text{sm}}(q)\|^2 \Big/ \sum_{i=1}^{N-Q} \|X_i\|^2, \end{gathered} \tag{5.13}$$

respectively. $\hat{X}_i^{\text{dyn}}(q)$ is the i^{th} functional observation approximated using the first q dynamic FPCs and $\hat{X}_i^{\text{sm}}(q)$ is the i^{th} functional observation approximated using the first q smooth dynamic FPCs. The smaller the NMSE, the superior is the approximation.

The scenario of the simulation study is designed to mimic the characteristics of the EpCO$_2$ functional data. Accordingly, an initial stationary functional time series of (N =400) observations is simulated according to the FAR(1) model, given by Equation 5.1, fitted to the EpCO$_2$ data. As previously mentioned in Section 5.2.1, this simulation must be performed in a finite dimension p. Following from this, the corresponding basis coefficients of these functional data are generated from the VAR(1) model, given by Equation 5.2, using the auto-regression matrix $\mathfrak{R}$ and the variance-covariance matrix Σ of the noise estimated for the EpCO$_2$ functional data; assuming that the simulated functional data, like the observed data, are approximated using ($p = 15$) B-splines basis functions.

For this initially simulated stationary functional time series, an ordinary time domain functional eigen-analysis is performed and the corresponding eigenvalues $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ and eigenfunctions $E_1(t), \ldots, E_p(t)$ are obtained. Then, to generate functional data with a covariance structure that varies smoothly over a grid of B time blocks, the eigenvalues $\gamma_2, \ldots, \gamma_p$ are multiplied by $0.1b$, where b is a segment of time within a grid of B time blocks, producing $\gamma^*_{2b}, \ldots, \gamma^*_{pb}$ and the first eigenvalue γ_1 is multiplied by $[1+(0.1b)] \times \gamma^*_{2b}$ to obtain γ^*_{1b} for $b = 1, \ldots, B$. Here, B is chosen to be 20. This setting provides a vector of eigenvalues at each time block b, say $\boldsymbol{\gamma}^*_b = (\gamma^*_{1b}, \ldots, \gamma^*_{pb})$, where both the (absolute) eigenvalues and the proportion of variance explained by the first FPC increase smoothly over time.

Mardia et al. (1979) has shown that an $N \times p$ matrix of data $\boldsymbol{\mathcal{Z}}$ with p variables and N data points can be constructed from a set of eigenvectors and PC scores using the following equation:

$$\boldsymbol{\mathcal{Z}} = \bar{\boldsymbol{\mathcal{Z}}} + \mathbf{S}\mathbf{E}^\top$$

where $\bar{\boldsymbol{\mathcal{Z}}}$ is an $N \times p$ matrix for the mean of each of the p variables, $\mathbf{S}$ is an $N \times p$ matrix of PC scores and $\mathbf{E}$ is a $p \times p$ matrix of eigenvectors. This expression can be generalized to a functional context as follows:

$$\mathcal{Z}_i(t) \approx \bar{\mathcal{Z}}(t) + \sum_{m=1}^{p} S_{mi} E_m(t), \quad t \in \mathcal{T} \tag{5.14}$$

where $\bar{\mathcal{Z}}(t)$ is the functional mean of the data $\mathcal{Z}_1(t), \ldots, \mathcal{Z}_N(t)$, S_{mi} is the score of the m^{th} PC for the i^{th} observation and $E_m(t)$ is the m^{th} eigenfunction of the data.

Following Equation 5.14, new locally stationary functional data with changing power spectrum/covariance structure over time are generated by simulating blocks of N/B functional observations at $b = 1, \ldots, B$ using the eigenvalues $\boldsymbol{\gamma}^*_b$. At each time block b, N/B scores, for each PC, are computed from a VAR(1) process with an auto-regression matrix $\mathfrak{R}^* = (\langle \Upsilon^*(E_m), E_{m'} \rangle : m \geq 1, m' \leq p = 15)$ that is obtained by first generating a $p \times p$ matrix $G = (G_{m,m'} : m \geq 1, m' \leq p)$, such that the elements $G_{m,m'}$ are mutually independent $N(0, \delta_{m,m'})$, then setting $\mathfrak{R}^*$ equal to $\kappa G/\|G\|$. The matrix G is normalized by dividing it by the square root of the largest eigenvalue of $GG^\top$ denoted by $\|G\|$, then multiplied by a pre-specified dependence coefficient κ that determines the size of $\|\Upsilon^*\|$. By defining $\delta_{m,m'}$=exp$\{-(m+m')\}$, the operator Υ^* is bounded and $\mathfrak{R}^*_{m,m'} \to 0$ as $m, m' \to \infty$. Other choices for $\delta_{m,m'}$, ensuring that the operator Υ^*, is bounded can be considered. The noise of the VAR(1) model used to generate the PC scores within each block b is chosen to have a multivariate normal distribution with a diagonal variance-covariance matrix $\boldsymbol{\Sigma}^*_b$, with the diagonal elements being the eigenvalues $\boldsymbol{\gamma}^*_b$ within this time block. Finally, the functional data $\mathcal{Z}(t)$ within each time block are constructed by multiplying the corresponding scores by the eigenfunctions $E_1(t), \ldots, E_p(t)$. This produces a time-varying functional times series of N observations, where N/B functional data share the same power spectrum and covariance structure. Stationary functional data sets can be generated using the same procedure by simulating the whole time series using the eigenvalues $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ and the eigenfunctions $E_1(t), \ldots, E_p(t)$.

To compare the performance of the smooth dynamic FPCA with that of the dynamic FPCA under different conditions, the above simulation procedure is repeated for a range of different settings (see Table 5.1). For both the non-stationary and stationary scenarios, we consider the values $0.1, 0.3, 0.6$ and 0.9 for κ, which reflect weak to strong dependence structure in the data, respectively. We also consider the values $10, 20, 40, 100$ for the standard deviation s used to smooth the covariance structure and estimate the smooth dynamic FPCs at each time point. For each choice of κ and s, the above simulation is repeated 200 times under the non-stationarity conditions and the mean and standard deviation of the NMSE between the simulated data and its reconstructed version using 1,2,3,6 dynamic FPCs and smooth dynamic FPCs are reported in Table 5.2 and the corresponding box-plots of the NMSE are displayed in Figure 5.21. The simulation procedure is also repeated 200 times assuming stationarity and the corresponding NMSE are reported in Table 5.3 and displayed in Figure 5.22.

κ s	0.1	0.3	0.6	0.9
	10	10	10	10
	20	20	20	20
	40	40	40	40
	100	100	100	100

TABLE 5.1: Summary of the different simulation study scenarios.

In almost all settings, the smooth dynamic FPCs outperform the dynamic FPCs in terms of the NMSE. As the dependence coefficient increases, the performance of both the ordinary and smooth dynamic FPCs improves (the mean of NMSE decreases); and the difference between them shrinks. This appears to be attributed to the enhanced performance of ordinary dynamic FPCs at high dependence, as the functional process becomes quite smooth and exhibits less variability with highly correlated observations. Thus, it can be concluded that although the smooth dynamic FPCs approximate the original curves more closely in almost all settings, they tend to highly outperform the ordinary dynamic FPCs in reconstructing the data with low to moderate dependence structure. In addition, as expected, the differences between the approximations by dynamic FPCA and smooth dynamic FPCA become negligible as the number of PCs used in the reconstruction increases. It is also evident that using a smaller standard deviation (less smooth spectral density) leads to superior results when fewer PCs are used

PCs	$\boldsymbol{\kappa}$	**SFPC**	**DFPC**	**smDFPC**	**smDFPC**	**smDFPC**	**smDFPC**
				s =**10**	s =**20**	s =**40**	s =**100**
1	0.1	0.445	0.39	0.055	0.137	0.251	0.35
		(0.025)	(0.022)	(0.003)	(0.007)	(0.013)	(0.019)
	0.3	0.438	0.368	0.053	0.133	0.241	0.331
		(0.029)	(0.023)	(0.003)	(0.007)	(0.013)	(0.019)
	0.6	0.405	0.3	0.047	0.117	0.202	0.27
		(0.057)	(0.027)	(0.004)	(0.01)	(0.018)	(0.024)
	0.9	0.303	0.191	0.036	0.082	0.133	0.174
		(0.111)	(0.042)	(0.006)	(0.018)	(0.030)	(0.039)
2	0.1	0.206	0.155	0.03	0.033	0.071	0.124
		(0.022)	(0.014)	(0.003)	(0.003)	(0.005)	(0.010)
	0.3	0.2	0.148	0.029	0.031	0.067	0.118
		(0.022)	(0.013)	(0.003)	(0.003)	(0.005)	(0.01)
	0.6	0.174	0.123	0.024	0.026	0.056	0.098
		(0.023)	(0.013)	(0.003)	(0.003)	(0.006)	(0.01)
	0.9	0.118	0.080	0.017	0.019	0.038	0.065
		(0.034)	(0.018)	(0.004)	(0.004)	(0.008)	(0.014)
3	0.1	0.059	0.047	0.019	0.012	0.023	0.038
		(0.008)	(0.006)	(0.002)	(0.001)	(0.003)	(0.005)
	0.3	0.057	0.045	0.018	0.012	0.022	0.037
		(0.008)	(0.006)	(0.002)	(0.001)	(0.003)	(0.005)
	0.6	0.049	0.039	0.015	0.010	0.019	0.032
		(0.008)	(0.006)	(0.002)	(0.001)	(0.003)	(0.005)
	0.9	0.033	0.027	0.011	0.008	0.014	0.022
		(0.01)	(0.007)	(0.003)	(0.003)	(0.003)	(0.006)
6	0.1	0.008	0.007	0.018	0.007	0.005	0.005
		(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
	0.3	0.008	0.007	0.017	0.007	0.005	0.005
		(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
	0.6	0.007	0.006	0.014	0.006	0.004	0.005
		(0.001)	(0.003)	(0.002)	(0.001)	(0.001)	(0.002)
	0.9	0.004	0.005	0.010	0.004	0.003	0.004
		(0.001)	(0.002)	(0.003)	(0.003)	(0.002)	(0.004)

Table 5.2: Results of the simulations in the non-stationary case in terms of the NMSE between the simulated curves and their recovered versions. The numbers represent the mean of NMSE for the static, dynamic and smooth dynamic FPCs from 200 simulation runs. The numbers in parentheses are the standard deviations of the NMSE.

in the reconstruction, but they lead to relatively higher NMSE using a larger number of PCs (see the rightmost box-plots of the NMSE at 6 PCs in Figures 5.21 and 5.22). Therefore, caution should be taken while choosing both the number of PCs used in the approximation and the amount of smoothing controlling the trade-off between smoothing the time-varying dynamic FPCs and over-fitting the data. Although the variance of NMSE increases as the dependence coefficient increases, it is observed that the variation in the NMSE is systematically lower for the smooth dynamic FPCs especially with

PCs	$\boldsymbol{\kappa}$	**SFPC**	**DFPC**	**smDFPC**	**smDFPC**	**smDFPC**	**smDFPC**
				s =**10**	s =**20**	s =**40**	s =**100**
1	0.1	0.349	0.319	0.05	0.126	0.215	0.283
		(0.034)	(0.03)	(0.004)	(0.007)	(0.016)	0.025
	0.3	0.346	0.299	0.047	0.12	0.202	0.266
		(0.036)	(0.029)	(0.004)	(0.008)	(0.016)	(0.024)
	0.6	0.324	0.238	0.041	0.099	0.162	0.211
		(0.064)	(0.028)	(0.004)	(0.01)	(0.018)	(0.024)
	0.9	0.24	0.138	0.03	0.063	0.097	0.124
		(0.12)	(0.038)	(0.005)	(0.016)	(0.027)	(0.034)
2	0.1	0.161	0.128	0.023	0.026	0.056	0.097
		(0.017)	(0.013)	(0.003)	(0.003)	(0.005)	(0.009)
	0.3	0.156	0.121	0.022	0.025	0.053	0.092
		(0.017)	(0.012)	(0.003)	(0.003)	(0.005)	(0.009)
	0.6	0.132	0.099	0.018	0.021	0.043	0.075
		(0.019)	(0.012)	(0.003)	(0.003)	(0.005)	(0.009)
	0.9	0.083	0.06	0.013	0.015	0.028	0.046
		(0.029)	(0.015)	(0.004)	(0.004)	(0.006)	(0.011)
3	0.1	0.046	0.038	0.014	0.01	0.018	0.03
		(0.005)	(0.005)	(0.002)	(0.002)	(0.002)	(0.003)
	0.3	0.044	0.037	0.013	0.009	0.017	0.028
		(0.005)	(0.004)	(0.002)	(0.002)	(0.002)	(0.003)
	0.6	0.037	0.031	0.011	0.008	0.015	0.024
		(0.005)	(0.005)	(0.002)	(0.002)	(0.002)	(0.003)
	0.9	0.023	0.020	0.009	0.007	0.011	0.016
		(0.008)	(0.005)	(0.004)	(0.003)	(0.003)	(0.004)
6	0.1	0.006	0.006	0.013	0.005	0.004	0.004
		(0.0007)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
	0.3	0.006	0.006	0.013	0.005	0.004	0.004
		(0.0007)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
	0.6	0.005	0.005	0.010	0.004	0.003	0.004
		(0.0007)	(0.003)	(0.002)	(0.001)	(0.001)	(0.002)
	0.9	0.003	0.005	0.008	0.004	0.004	0.004
		(0.001)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)

Table 5.3: Results of the simulations in the stationary case in terms of the NMSE between the simulated curves and their recovered versions. The numbers represent the mean of NMSE for the static, dynamic and smooth dynamic FPCs from 200 simulation runs. The numbers in parentheses are the standard deviations of the NMSE.

lower standard deviation. The same conclusion applies to both the stationary and non-stationary case but the improvement introduced by the smooth dynamic FPCs is considerably smaller in the stationary case compared to the non-stationary one.

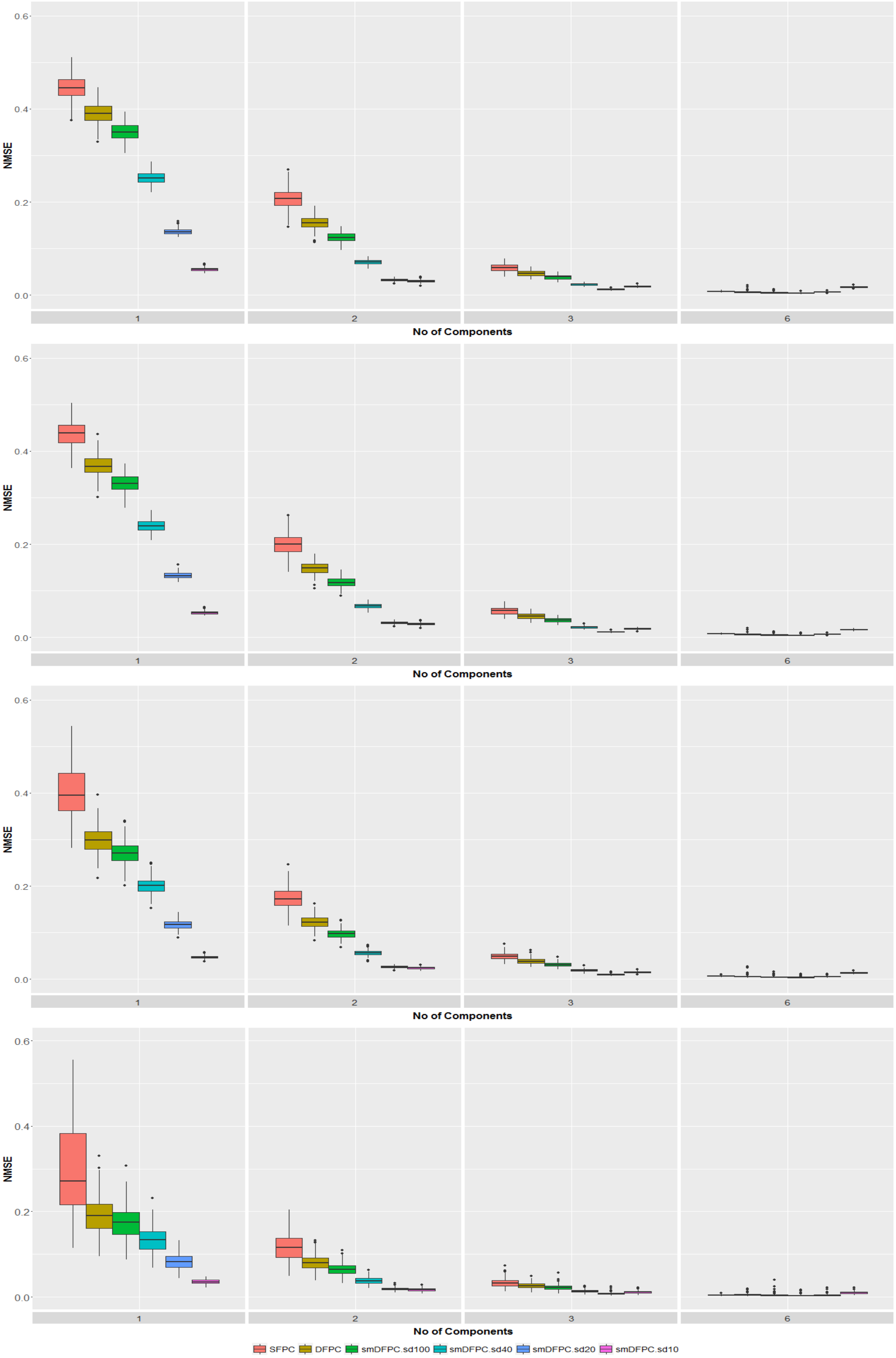


FIGURE 5.21: Box-plots of the NMSE between the simulated curves and their recovered versions using 1, 2, 3 and 6 (from left to right) static (red), dynamic (olive) and smooth dynamic FPCs with a standard deviation s =100 (green), 40 (turquoise), 20 (blue), 10 (purple) based on the results from 200 non-stationary simulation runs with $\kappa = 0.1, 0.3, 0.6, 0.9$ (from top to bottom).

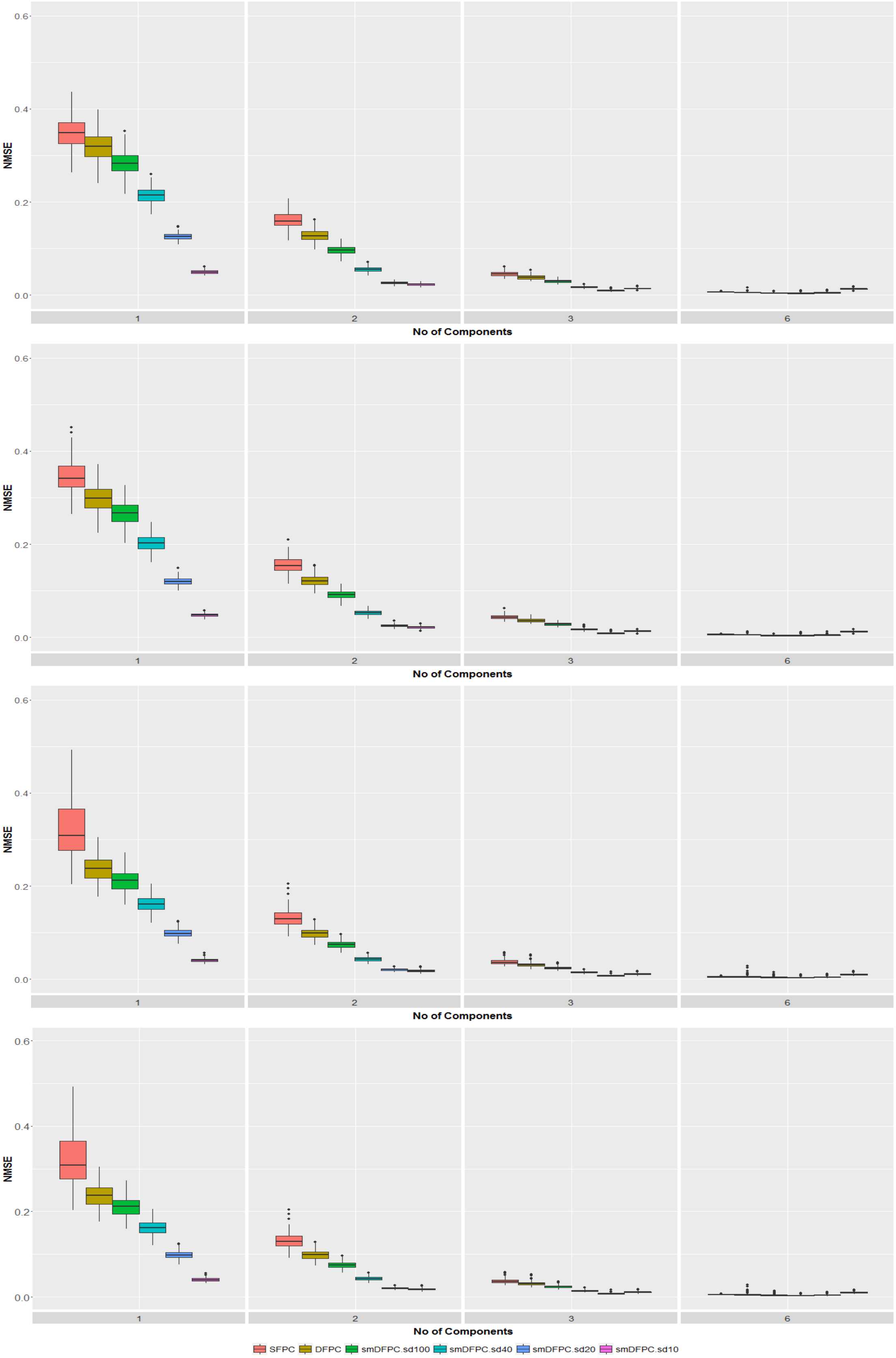


FIGURE 5.22: Box-plots of the NMSE between the simulated curves and their recovered versions using 1, 2, 3 and 6 (from left to right) static (red), dynamic (olive) and smooth dynamic FPCs with a standard deviation s =100 (green), 40 (turquoise), 20 (blue), 10 (purple) based on the results from 200 stationary simulation runs with $\kappa = 0.1, 0.3, 0.6, 0.9$ (from top to bottom).

Our proposed smooth dynamics, like the ordinary dynamic FPCs, are subject to approximation errors since they are based on approximate and not exact methods. These errors originate for instance from the numerical integration of the filters and the truncation of the filters at some finite lag L. These errors are negligible if a single component explains a large proportion of the variability. In contrary, these approximation errors are not recovered if the component explains an insignificant amount of the variance. This is mainly observed with the 6 PCs, when the NMSEs of the dynamic FPCs and the smooth version are larger than the static. There is another type of error stemming from the smoothing parameter which also become more noticeable when using a large number of PCs, see the case of 6 PCs, where the NMSE of the approximation based on the dynamic FPCs is smaller than that based on the smooth version. However, this effect disappears as the smoothing parameter gets larger.

To evaluate the bias of both the smooth dynamic FPCs and the dynamic FPCs in estimating the true eigenvalues underlying the original functional process in the case of stationary and locally stationary functional processes, the estimated eigenvalues of the time-varying spectral density and that of the overall spectral density are compared to the true frequency domain eigenvalues used to generate the data in both scenarios. For the non-stationary scenario, in each repetition, the scores within each time block b are generated from a VAR(1) with an auto-regression matrix $\mathfrak{R}^*$ defined as above and with noise $\boldsymbol{\epsilon}_i^*$ chosen as independent zero-mean Gaussian with a variance-covariance matrix $\boldsymbol{\Sigma}_b^* = \text{diag}(\gamma_{1b}^*, \ldots, \gamma_{pb}^*)$. Let $\mathbf{S}_i = (S_{1i}, \ldots, S_{pi})^\top$ be the vector of scores at time i, such that $\mathbf{S}_i = \sum_{j=0}^{\infty} \mathfrak{R}^{*^j} \boldsymbol{\epsilon}_{t-j}$, according to the infinite-order moving average representation of the VAR(1). Based on this expression, the lag h covariance matrix Γ_{hb} of $\mathbf{S}_i$ is defined by (Enders, 2004):

$$\Gamma_{hb} = \mathbb{E}\Big[\mathbf{S}_i \mathbf{S}_{i-h}^\top\Big] = \sum_{j=0}^{\infty} \mathfrak{R}^{*^j} \boldsymbol{\Sigma}_b^* \mathfrak{R}^{*^{j+h}\,\top}.$$

Accordingly, the variance of $\mathbf{S}_i$ within block b, denoted by Γ_{ob}, is obtained by substituting h with zero and is calculated as follows:

$$\text{vec}(\Gamma_{ob}) = (I_{15^2} - \mathfrak{R}^* \otimes \mathfrak{R}^*)^{-1} \text{vec}(\boldsymbol{\Sigma}_b^*),$$

where vec(.) is a linear operator converting a matrix into a column vector, I_{15^2} is a $15^2 \times 15^2$ identity matrix and $\otimes$ refers to the kronecker product between matrices. After computing Γ_{ob}, the covariance matrix at lag h within block b is computed recursively by $\Gamma_{hb} = \mathfrak{R}^{*^h} \Gamma_{ob}$. The corresponding spectral density of $\mathbf{S}_i$ within block b is then defined as the discrete Fourier transform of the covariance matrix Γ_{hb}:

$$\mathfrak{F}_b^{\mathbf{S}}(\theta) = \frac{1}{2\pi} \sum_{h \in \mathbb{Z}} \Gamma_{hb} e^{-\mathrm{i}h\theta}. \tag{5.15}$$

If all the eigenvalues of $\mathfrak{R}^*$ have modulus less than 1, then $\mathfrak{R}^{*^j}e^{-\mathrm{i}\theta}$ is an infinite geometric sequence such that:

$$\sum_{j=0}^{\infty}\mathfrak{R}^{*^j}e^{-\mathrm{i}\theta} = (I_{15} - \mathfrak{R}^*e^{-\mathrm{i}\theta})^{-1},$$

where I_{15} is a 15 × 15 identity matrix. Thus, the spectral density given by Equation 5.15 for a VAR(1) model can be equivalently computed in practice as follows:

$$\mathfrak{F}_b^{\mathbf{S}}(\theta) = \frac{1}{2\pi}(I_{15} - \mathfrak{R}^*e^{-\mathrm{i}\theta})^{-1}\mathbf{\Sigma}_b^*((I_{15} - \mathfrak{R}^*e^{-\mathrm{i}\theta})^{-1})^{\bullet} \tag{5.16}$$

such that $(.)^{\bullet}$ is the conjugate transpose of $(.)$.

In each repetition, a functional process $\mathcal{Z}(t)$ is generated by multiplying the simulated scores by the corresponding eigenfunctions $E_1(t), \ldots, E_p(t)$ following Equation 5.14 such that $E_m(t) = \boldsymbol{\psi}(t)^{\top}\mathbf{c}_m$, where $\boldsymbol{\psi}(t) = (\psi_1(t), \ldots, \psi_p(t))^{\top}$ is the same vector of B-splines basis functions used to simulate the original functional time series and $\mathbf{c}_m = (c_{m1}, \ldots, c_{mp})$ is the corresponding vector of coefficients. Following from this, the spectral density kernel of the functional data $\mathcal{Z}(t)$ is obtained by:

$$\mathfrak{F}_b^{\mathcal{Z}}(\theta) = \mathbf{C}\mathfrak{F}_b^{\mathbf{S}}(\theta)\mathbf{C}^{\top}\mathbf{W}^{\top}, \tag{5.17}$$

where $\mathbf{C}$ is a 15 × 15 matrix of eigenfunctions coefficients such that the m^{th} column is the vector $\mathbf{c}_m$ and $\mathbf{W} = \int \psi(t)\psi(t)^{\top}dt$. An eigen-analysis of $\mathfrak{F}_b^{\mathcal{Z}}(\theta)$ results in the true frequency domain eigenvalues of the process. In the case of non-orthogonal basis functions, $\mathfrak{F}_b^{\mathcal{Z}}(\theta)$ is not Hermitian and hence the corresponding eigenvalues might not be real. Using the re-parametrization proposed in Chapter 4, real eigenvalues are obtained by the decomposition of $\mathbf{W}^{1/2}\mathbf{C}\mathfrak{F}_b^{\mathbf{S}}(\theta)\mathbf{C}^{\top}\mathbf{W}^{1/2\top}$ instead, where $\mathbf{W}^{1/2}$ is the Cholesky decomposition matrix of $\mathbf{W}$. These eigenvalues are then compared to the corresponding ones obtained from the decomposition of the overall spectral density as well as those obtained from the decomposition of the time-varying spectral density using the mean squared difference as follows:

$$\begin{aligned}\frac{1}{N-Q}\sum_{i=1}^{N-Q}\|\bar{\gamma}_{im} - \bar{\hat{\gamma}}_{im}^{\mathrm{dyn}}\|^2 \quad \forall m \geq 1,\\ \frac{1}{N-Q}\sum_{i=1}^{N-Q}\|\bar{\gamma}_{im} - \bar{\hat{\gamma}}_{im}^{\mathrm{sm}}\|^2 \quad \forall m \geq 1,\end{aligned} \tag{5.18}$$

where $\bar{\gamma}_{im}$ is the m^{th} true eigenvalue averaged over all frequencies $\theta \in (-\pi, 0) \cup (0, \pi)$ at time point i, and $\bar{\hat{\gamma}}_{im}^{\mathrm{dyn}}$ and $\bar{\hat{\gamma}}_{im}^{\mathrm{sm}}$ are the m^{th} eigenvalue obtained from the eigen-analysis of the overall and time-varying spectral density averaged over the same set of frequencies at time point i, respectively.

The same procedure is repeated for the stationary case, but assuming that the scores of the functional data are wholly generated from the same VAR(1) model. Following from this, the true eigenvalues are obtained from the spectral decomposition of $\mathfrak{F}^{\mathcal{Z}}(\theta)$ expressed by $\mathbf{W}^{1/2}\mathbf{C}\mathfrak{F}^{\mathbf{S}}(\theta)\mathbf{C}^{\top}\mathbf{W}^{1/2\top}$, where $\mathfrak{F}^{\mathbf{S}}(\theta)$ is the spectral density of the scores generated using $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$.

The mean squared differences between the true eigenvalues and the corresponding estimated ones from the dynamic and the smooth dynamic FPCs, using different standard deviations, for both the non-stationary and stationary cases are reported in Tables 5.5 and 5.4 and displayed in Figures 5.24 and 5.23, respectively. For the stationary simulated data sets, the distance between the estimated dynamic eigenvalues and the true eigenvalues, in all the settings, is smaller than the distance between the time-varying eigenvalues and the true eigenvalues obtained under a VAR(1) model. This difference becomes more striking as the standard deviation decreases. In addition, as the dependence coefficient increases, the variance of the VAR(1) process and the lag-1 correlations increase since it involves more information about the past; and hence the system redundancy increases and the first few eigenvalues increase in absolute values. Therefore, the mean squared difference between the corresponding estimated and true eigenvalues get very large but the relative distance between the dynamic and the smooth dynamic eigenvalues get smaller. This result is more obvious for the first PC which accounts for the largest portion of variability in the data. Using the smooth dynamic FPCs with stationary data introduce some artefacts and bias that get magnified with a smaller standard deviation. Thus, we can conclude that the ordinary dynamic FPCs provide better approximations, than the smooth dynamic FPCs, for stationary functional time series. In contrast, the distance between the time-varying eigenvalues and the true eigenvalues is smaller than the distance between the dynamic eigenvalues and the true ones under the specified block-varying VAR(1) model. This result highlights the inappropriateness of the ordinary dynamic FPCs with non-stationary functional time series and the need for an equivalence that vary over time to adjust for changes in the covariance and power spectra. Again, the large scale distances at high correlation are just a result of the larger scale variance of the first principal components associated with large dependence coefficient. The large amount of variability in the results stems from the wide variability between the data sets simulated based on the dependence coefficient. There is an evidence as well that a small standard deviation like 10 tends to over-fit the data with a larger squared distance, whereas a standard deviation of 100 could under-fit the data, compared to the results of a standard deviation 20 or 40. The squared distance between the true and estimated eigenvalues are minimum at a standard deviation of 40 for the first 2 PCs and at 100 for the 3^{rd} and 6^{th} PCs. Thus, although a standard deviation of 10 leads to the minimum normalized mean squared error between the original and

No of PCs	$\boldsymbol{\kappa}$	**DFPC**	**smDFPC**	**smDFPC**	**smDFPC**	**smDFPC**
			s =**10**	s =**20**	s =**40**	s =**100**
1	0.1	15.029	6.073	2.176	0.889	1.827
		(3.711)	(3.105)	(1.589)	(0.758)	(0.807)
	0.3	17.554	6.125	2.256	0.976	2.192
		(4.383)	(3.270)	(1.654)	(0.813)	(1.035)
	0.6	31.425	8.347	3.611	1.955	4.552
		(8.498)	(6.246)	(3.184)	(1.773)	(2.429)
	0.9	160	80.602	49.113	33.897	43.886
		(107.662)	(111.629)	(66.533)	(45.412)	(50.765)
2	0.1	1.567	0.715	0.211	0.100	0.223
		(0.38)	(0.287)	(0.012)	(0.073)	(0.11)
	0.3	1.521	0.597	0.172	0.086	0.21
		(0.368)	(0.272)	(0.116)	(0.071)	(0.105)
	0.6	1.38	0.383	0.124	0.08	0.187
		(0.381)	(0.254)	(0.099)	(0.07)	(0.097)
	0.9	1.248	0.22	0.123	0.111	0.174
		(0.394)	(0.188)	(0.109)	(0.109)	(0.089)
3	0.1	0.662	1.385	0.704	0.366	0.274
		(0.106)	(0.232)	(0.155)	(0.115)	(0.096)
	0.3	0.654	1.353	0.688	0.359	0.27
		(0.107)	(0.231)	(0.154)	(0.115)	(0.096)
	0.6	0.618	1.216	0.61	0.316	0.242
		(0.115)	(0.245)	(0.163)	(0.119)	(0.099)
	0.9	0.537	0.916	0.423	0.203	0.168
		(0.128)	(0.275)	(0.172)	(0.115)	(0.095)
6	0.1	0.002	0.006	0.004	0.002	0.001
		(0.0004)	(0.0009)	(0.0007)	(0.0005)	(0.0004)
	0.3	0.002	0.006	0.004	0.002	0.001
		(0.0004)	(0.0009)	(0.0007)	(0.0005)	(0.0004)
	0.6	0.002	0.006	0.003	0.002	0.001
		(0.0004)	(0.0009)	(0.0007)	(0.0005)	(0.0004)
	0.9	0.002	0.006	0.003	0.002	0.001
		(0.0004)	(0.0009)	(0.0007)	(0.0005)	(0.0004)

Table 5.4: Results of the simulations in the non-stationary case in terms of the difference between the true and estimated eigenvalues. The numbers represent the mean of squared error between the true and estimated estimated eigenvalues for the dynamic and smooth dynamic FPCs from 200 simulation runs. The numbers in parentheses are the standard deviations of the squared error.

reconstructed functional series, the distance between the true and estimated eigenvalues is quite large. Whereas, a standard deviation of 40 generally leads to the minimum squared distance between the true and estimated eigenvalues; however the corresponding NMSE is not minimum. Therefore, the smoothing parameter should be chosen in a manner that fairly captures the variations over time without over/under fitting the original data.

No of PCs	κ	**DFPC**	**smDFPC**	**smDFPC**	**smDFPC**	**smDFPC**
			s =**10**	s =**20**	s =**40**	s =**100**
1	0.1	0.359	7.376	3.044	1.391	0.592
		(0.549)	(3.358)	(1.968)	(1.224)	(0.725)
	0.3	0.428	8.258	3.545	1.666	0.722
		(0.657)	(3.971)	(2.346)	(1.502)	(0.896)
	0.6	1.003	15.016	7.20	3.616	1.653
		(1.531)	(9.648)	(5.677)	(3.721)	(2.167)
	0.9	19.34	200.633	108.779	58.782	29.597
		(41.787)	(281.05)	(161.131)	(93.619)	(54.707)
2	0.1	0.024	0.346	0.097	0.057	0.034
		(0.039)	(0.197)	(0.061)	(0.0426)	(0.041)
	0.3	0.025	0.304	0.091	0.057	0.036
		(0.040)	(0.178)	(0.057)	(0.045)	(0.044)
	0.6	0.039	0.228	0.092	0.071	0.051
		(0.054)	(0.156)	(0.059)	(0.063)	(0.059)
	0.9	0.092	0.174	0.129	0.123	0.103
		(0.099)	(0.117)	(0.091)	(0.100)	(0.099)
3	0.1	0.037	0.987	0.463	0.205	0.072
		(0.03)	(0.169)	(0.104)	(0.067)	(0.041)
	0.3	0.036	0.965	0.453	0.201	0.071
		(0.029)	(0.168)	(0.103)	(0.067)	(0.04)
	0.6	0.031	0.874	0.407	0.179	0.062
		(0.027)	(0.179)	(0.109)	(0.069)	(0.04)
	0.9	0.019	0.669	0.289	0.117	0.037
		(0.023)	(0.209)	(0.122)	(0.070)	(0.035)
6	0.1	0.0002	0.0043	0.003	0.0012	0.0004
		(0.0001)	(0.0007)	(0.0005)	(0.0003)	(0.0002)
	0.3	0.00022	0.0043	0.003	0.0012	0.0004
		(0.0001)	(0.0007)	(0.0005)	(0.0003)	(0.0002)
	0.6	0.0002	0.0042	0.002	0.00118	0.0004
		(0.0001)	(0.0007)	(0.0005)	(0.0003)	(0.0002)
	0.9	0.0002	0.0041	0.002	0.0011	0.0004
		(0.0001)	(0.0007)	(0.0005)	(0.0003)	(0.0002)

Table 5.5: Results of the simulations in the stationary case in terms of the difference between the true and estimated eigenvalues. The numbers represent the mean of squared error between the true and estimated estimated eigenvalues for the dynamic and smooth dynamic FPCs from 200 simulation runs. The numbers in parentheses are the standard deviations of the squared error.

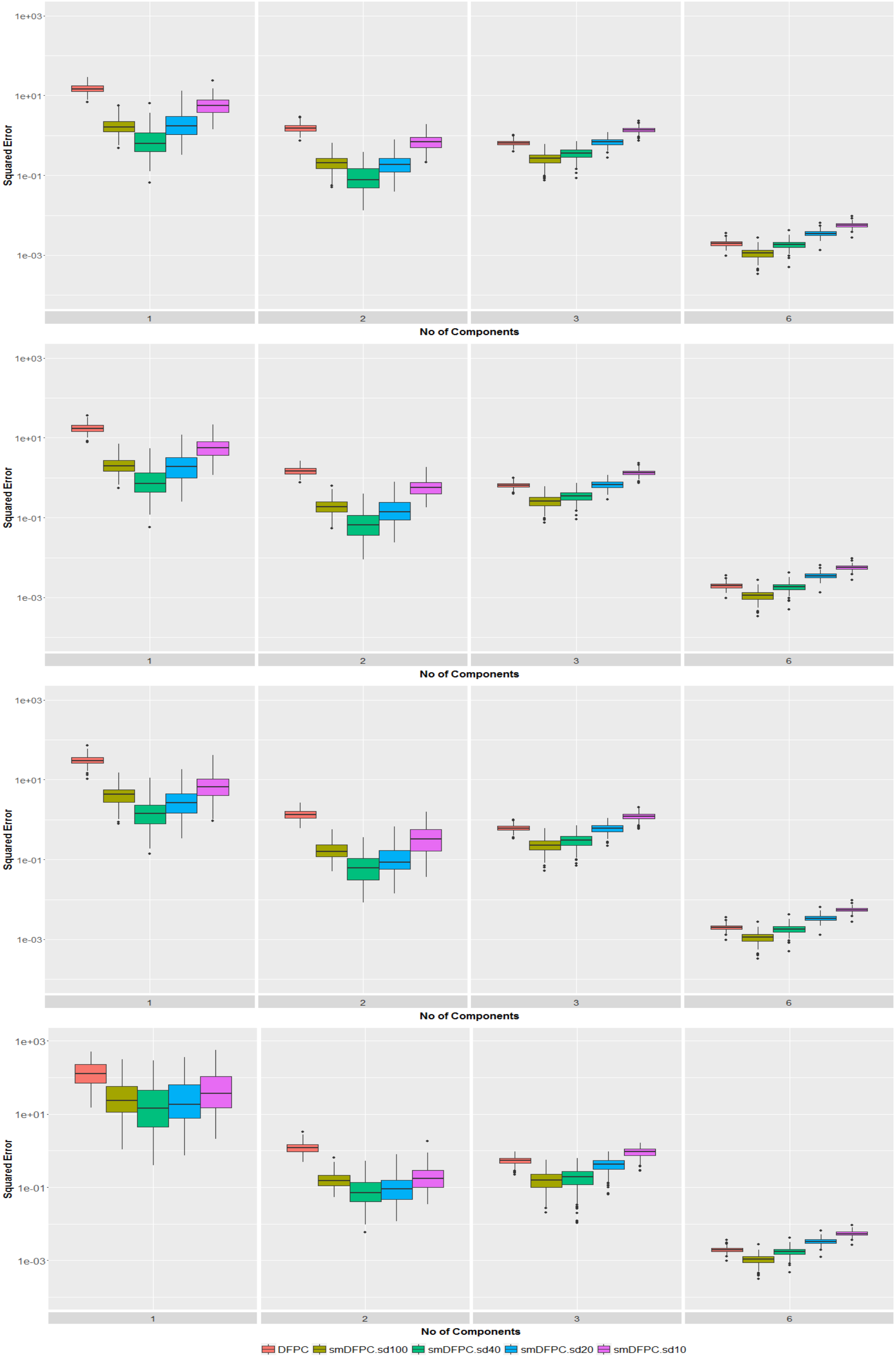


FIGURE 5.23: Box-plots of the squared errors between the true eigenvalues and the corresponding estimated eigenvalues of dynamic FPCs (red), and smooth dynamic FPCs with a standard deviation s =100 (olive), 40 (green), 20 (blue), 100 (purple) based on the results from 200 non-stationary simulation runs with $\kappa = 0.1, 0.3, 0.6, 0.9$ (from top to bottom). The results displayed (from left to right) concerns the 1^{st}, 2^{nd},3^{rd},6^{th} eigenvalues.

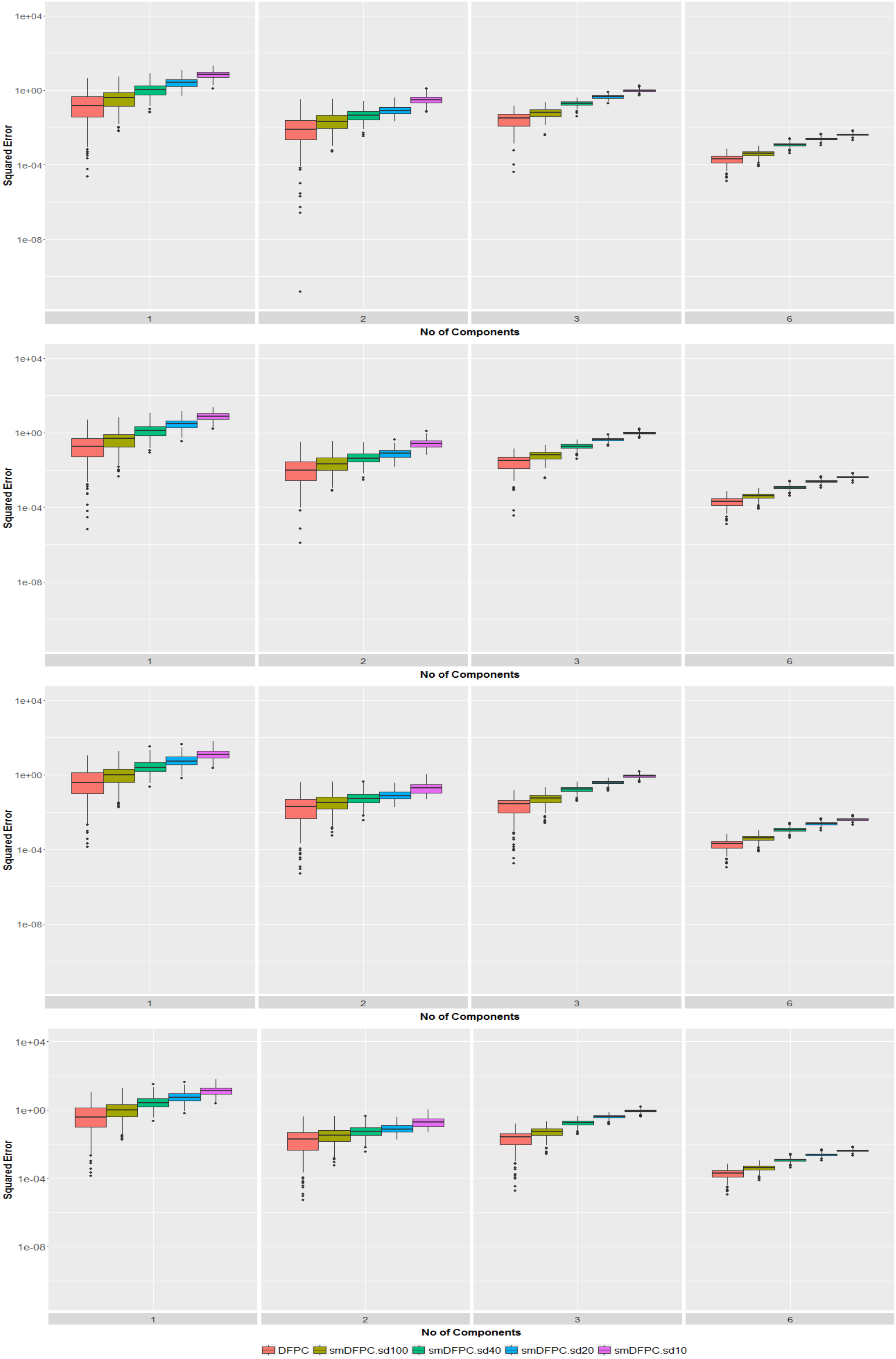


FIGURE 5.24: Box-plots of the squared errors between the true eigenvalues and the corresponding estimated eigenvalues of dynamic FPCs (red), and smooth dynamic FPCs with a standard deviation s =100 (olive), 40 (green), 20 (blue), 100 (purple) based on the results from 200 stationary simulation runs with $\kappa = 0.1, 0.3, 0.6, 0.9$ (from top to bottom). The results displayed (from left to right) concerns the 1^{st}, 2^{nd},3^{rd},6^{th} eigenvalues.

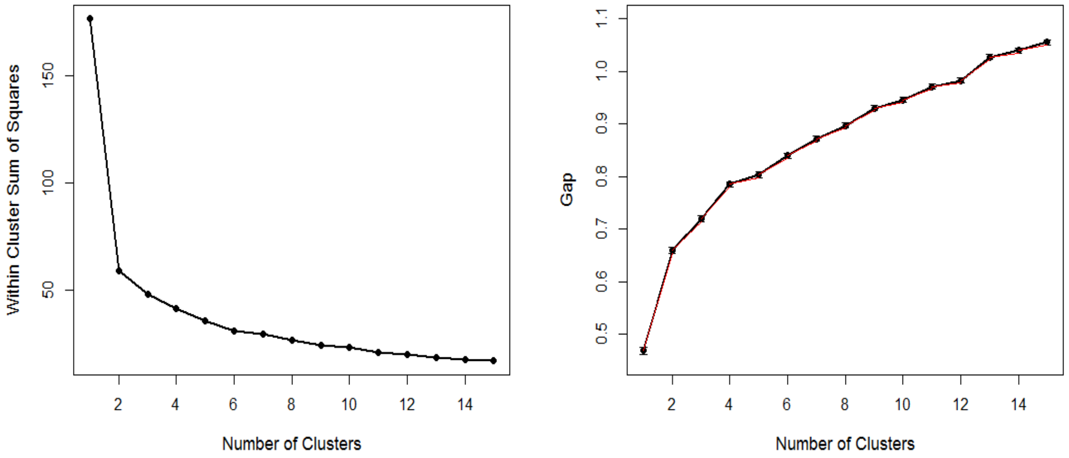
5.7 Clustering Functional Data using Smooth Dynamic FPCs

According to the previous simulation results, the time-varying version of the dynamic FPCs outperforms the original dynamic FPCs for non-stationary functional time series. Following from this result, it is more convenient to use the smooth dynamic FPCs to summarize non-stationary and highly dynamic data such as the EpCO$_2$ data. Thus, with analogy to the clustering procedure proposed in Section 4.3, it is worth investigating the results of classifying the daily functional curves of EpCO$_2$ based on their smooth dynamic FPC scores. Previous results have shown that the first smooth dynamic FPC explains solely about 94% of the variability in the data, see Section 5.5. Accordingly, we have initially decided to cluster the EpCO$_2$ daily curves, taking into account the correlation between the curves and the non-stationarity in the time series, based on the sequence of current, previous and future scores. We have chosen to cluster the daily curves based on the coefficients (scores) of the 3 central filters of the first smooth dynamic FPC, which are the most influential filters having on average $\sum_{l=-1}^{1} \|\hat{\phi}_{1l}(i)\|^2 = 0.7$. The previous, current and future scores of the first smooth dynamic FPC are denoted by $\hat{Z}_{1,i-1}, \hat{Z}_{1,i}, \hat{Z}_{1,i+1}$ and are computed for each i, $i = 1, \ldots, N-Q$, using the functional filter $\hat{\phi}_{1l}(i)$ estimated individually at that time point i as follows:

$$\begin{aligned}
\hat{Z}_{1,i-1} &= \sum_{l=-1}^{1} a_{i-1-l}^{\top} \mathbf{W} \hat{\hat{\boldsymbol{\phi}}}_{1l}(i), \\
\hat{Z}_{1,i} &= \sum_{l=-1}^{1} a_{i-l}^{\top} \mathbf{W} \hat{\hat{\boldsymbol{\phi}}}_{1l}(i), \\
\hat{Z}_{1,i+1} &= \sum_{l=-1}^{1} a_{i+1-l}^{\top} \mathbf{W} \hat{\hat{\boldsymbol{\phi}}}_{1l}(i).
\end{aligned}$$

Note that $\hat{\hat{\boldsymbol{\phi}}}_{1l}(i)$ are the basis coefficients of the functional filters $\hat{\phi}_{1l}(i)$.

Clustering based on the scores of the smooth dynamic FPCs could be misleading and meaningless. This is because the estimated functional filters are varying over time and hence each observation is loaded on different functional filters. For this reason, we argue that clustering is better preformed on the functional filters themselves and not the corresponding scores. According to the foundation theory of dynamic FPCs, the functional filters are constructed based on the same set of basis functions used to approximate the original curves. Following from this, we propose clustering the EpCO$_2$ functional curves using a K-means algorithm based on the corresponding estimated basis coefficients of the sequentially related functional filters of the first FPCs. Taking advantage of the fact that $\sum_{l=-1}^{1} \|\hat{\phi}_{1l}(i)\|^2 = 0.7$ on average for all i, we choose to cluster the curves based on the basis coefficients of the 3 central functional filters of the first smooth dynamic FPC obtained at the lags -1, 0 and 1 for each time point, denoted by $(\hat{\boldsymbol{\phi}}_{1,-1}(i), \hat{\boldsymbol{\phi}}_{1,0}(i), \hat{\boldsymbol{\phi}}_{1,+1}(i)), i = 1, \ldots, N - Q$. To determine the optimal number of clusters, the L-curve and gap statistic are employed. The within cluster sum of squares and the gap statistic are calculated for a number of clusters ranging between 1 and 15. The gap statistic fails to allocate a reasonable number of clusters for the data, however from the elbow plot a number between 2 and 6 seems plausible (Figure 5.25). For the sake of comparison with the classification results based on the dynamic FPC scores, a K-means clustering based on 3 centers is initially performed on the coefficients of the time-varying 3 central functional filters.



(a)

(b)

FIGURE 5.25: (a) L-curve and (b) gap statistic plot for the K-means clustering of daily DD EpCO$_2$ curves based on the coefficients of the 3 central time-varying functional filters of the first smooth dynamic FPC.

The classification results based on the central time-varying functional filters are displayed in Figures 5.26 and 5.27. Note the missing clustering at the end of the series, this is because the filters can only be obtained for the observations $i = 1, \ldots, N - Q$ (see Section 5.3). It is evident that the resulting clustering structure is different from that based on the dynamic FPC scores shown in Chapter 4 in Figure 4.11. The classification based on the time-varying functional filters of the first smooth dynamic FPC is smoother and no longer depends on the EpCO$_2$ mean level. This is because the clustering algorithm is based on the functional filters and not the corresponding scores, and the functional filters only determine how the current curve is sequentially related to the previous and future ones. It is interesting to observe the presence of two dominant and alternating patterns (that of the purple and blue classes) that persist over long time periods, in addition to some short transitional periods of a different pattern (turquoise class).

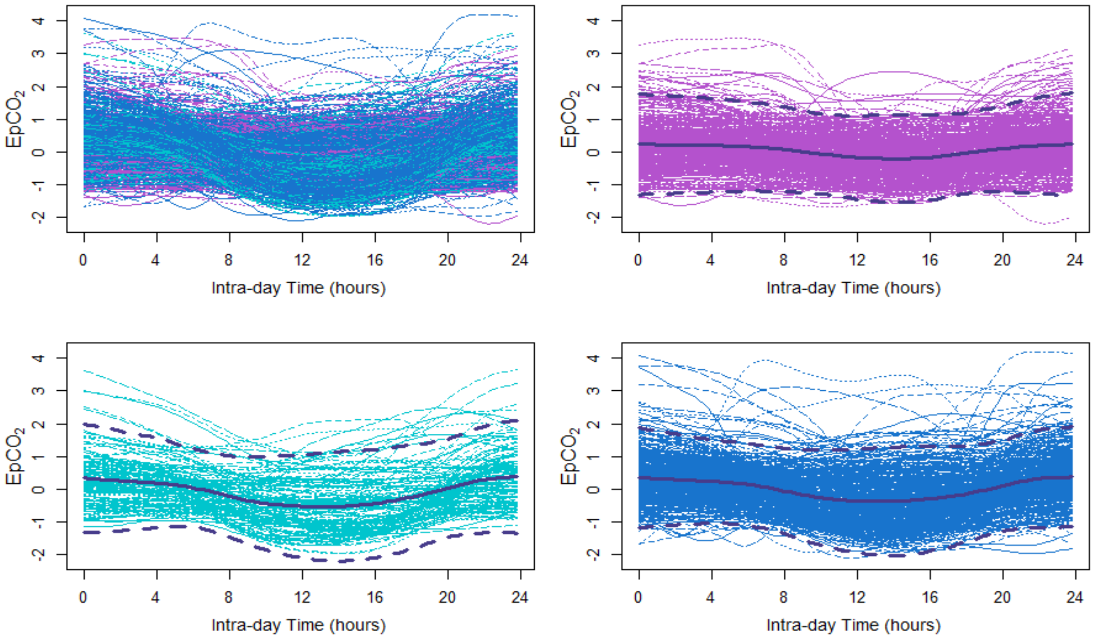


FIGURE 5.26: K-means clustering results of the daily DD EpCO$_2$ curves based on the coefficients of the 3 central time-varying functional filters of the first smooth dynamic FPC $(\hat{\boldsymbol{\phi}}_{1,-1}(i), \hat{\boldsymbol{\phi}}_{1,0}(i), \hat{\boldsymbol{\phi}}_{1,+1}(i), i = 1, \ldots, N - Q)$. The curves displayed here are recovered using only the first smooth dynamic FPC. The solid and dashed purple curves represent the mean curve and ± 2× standard deviation bands, respectively.

To illustrate the differences between the functional filters in the 3 clusters, the 3 central time-varying filters of the first smooth dynamic FPC are plotted for each cluster separately in Figure 5.28. Moreover, Figures 5.29, 5.30 and 5.31 display the overall mean curve $\hat{\mu}(t)$ plus the average effects of those 3 central filters $\sum_{l=-1}^{1} \delta_l \overline{\hat{\phi}_{1l}}$ within each cluster and show the differences between their average sequential effects on the mean level. Although the average sequential effect of the filters on the mean curve looks very similar in the purple and blue classes, the timing and length of the intra-daily trough, resulting from the sequential effect of 3 consecutive scores on the EpCO$_2$ level of the current day i, are different in both classes. Whereas, Figure 5.31 shows that the turquoise class is characterized by an opposite sequential effect when the 3 central scores are consecutively large (or small) and when two large (or small) scores are followed by a small (or large) score. Notice that 3 consecutive small FPC scores result in a negative shift in the blue and purple class and a positive shift that decreases as we approach the end of the day in the turquoise class. These differences in the effects, causing disturbances in the average daily cycle, can be attributed to the underlying climatological conditions including warm/cool temperature and short/long daytime as well as hydrological events such as storms and heavy rainfall. Such conditions influence the extent and strength of correlation between the days in the different clusters, which in turn affect the filters and hence the EpCO$_2$ pattern of day i.

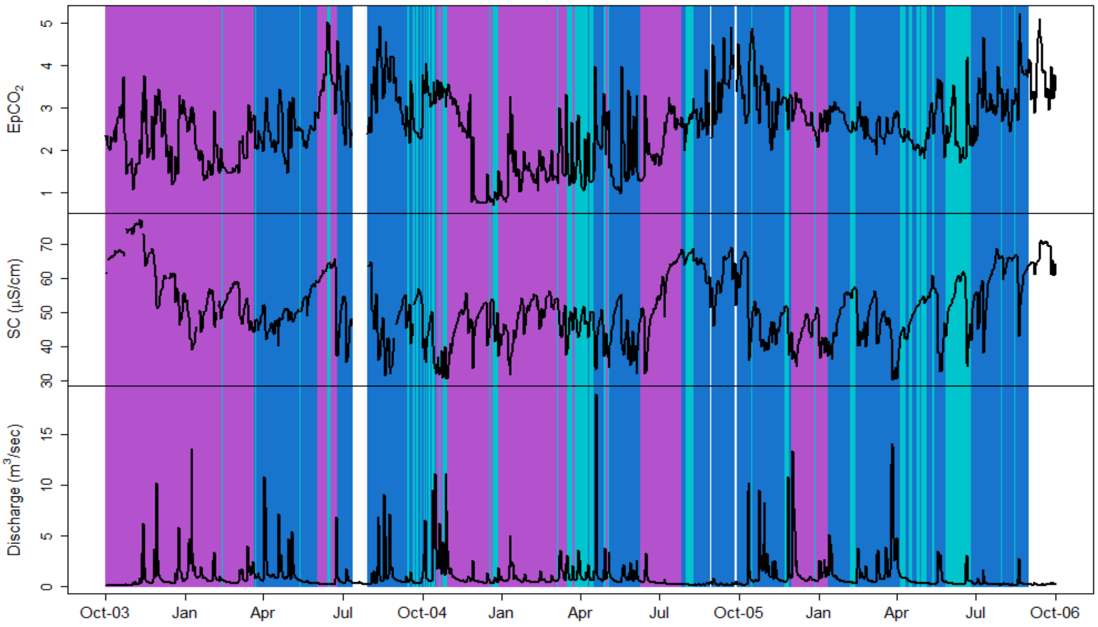


Figure 5.27: Class membership of days obtained using K-means clustering of the DD EpCO$_2$ curves based on the corresponding coefficients of the 3 central time-varying functional filters of the first smooth dynamic FPC. The solid curves are the corresponding daily averaged EpCO$_2$ (top), SC (middle) and water discharge (bottom).

The above resulting grouping structure poses many questions regarding the climatological and hydrological drivers of each class. One of these questions, for instance, is whether the blue class highlights warm periods associated with heavy but short-memory flow events. Such hypotheses need further investigations. To study how the grouping structure relates to the underlying hydrology, a similar analysis to that in Chapter 4 involving the comparison of the specific conductivity (SC) distributions in the 3 identified clusters is carried on. Using the same methodology in Chapter 4, the distributions of the smooth daily means and variances of SC are compared in the 3 clusters. Figure 5.32(a-b) displays respectively the box-plot and the empirical cumulative distribution of the smoothed daily averaged SC for each cluster of curves. The cumulative distribution function of the purple class tends to lie on the right of that of the other classes covering a wide range of SC.

By comparing the cumulative distribution functions of the daily variances in the 3 clusters (Figure 5.33), we found that the function of the purple class always lies on the left of that of the other two classes. This finding implies that the purple class is generally characterized by relatively smaller within-day variability. Thus, although the distribution function of the mean of the purple class covers a wide range of conductivity (Figure 5.32), the variability within these days is relatively small highlighting days with long

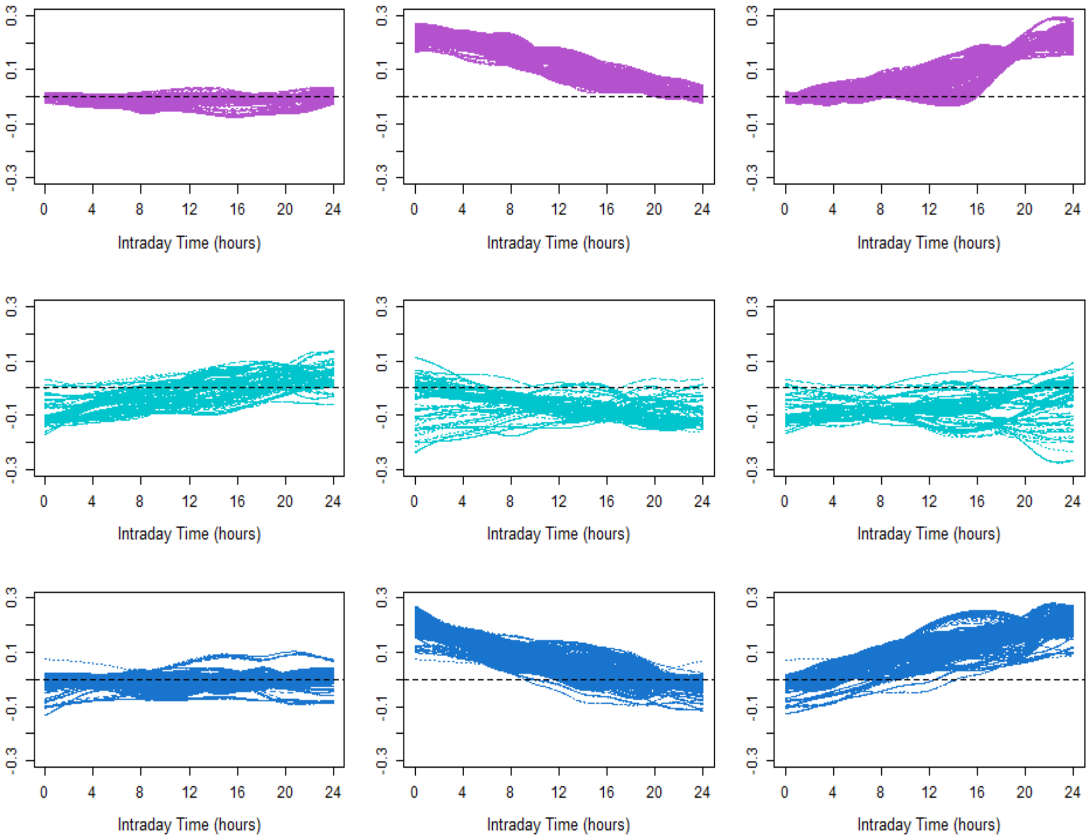


FIGURE 5.28: The 3 central functional filters of the first smooth dynamic FPC $\hat{\phi}_{1,-1}(i)$ (left),$\hat{\phi}_{1,0}(i)$ (middle),$\hat{\phi}_{1,+1}(i)$ (right), for $i = 1, \dots, N-Q$ colored by their corresponding class membership.

memory hydrological events or days within the recovery period of these events. In contrast, the blue class has a wider probability plot lying on the right hand side of that of the other two classes and is characterized by larger intra-daily variability.

According to the above classification, it seems that the pattern of the turquoise class is mostly driven by the biological activity in the catchment while those of the purple and blue classes are driven by the underlying hydrological conditions. The turquoise class seems to involve the periods with less correlated days. Whilst, the blue class is generally characterized by larger intra-day SC variability and is believed to highlight the periods with heavy and multiple rainfall events that bring more soil water, rich in CO_2, to the surface water. Alternatively, the purple class is characterized by less SC within-day variability, reflecting the gradual changes in river hydrology. This class mainly underlies periods with less carbon in the catchment possibly resulting from rainfall events that do not drive soil water to flow into the catchment or long-term events that dilute the CO_2 in the catchment. It is also noticed that producing more CO_2 within the catchment tend to accelerate the daily cycle of $EpCO_2$ by comparing Figure 5.31(c,d,e,f) with Figure 5.29(c,d,e,f).

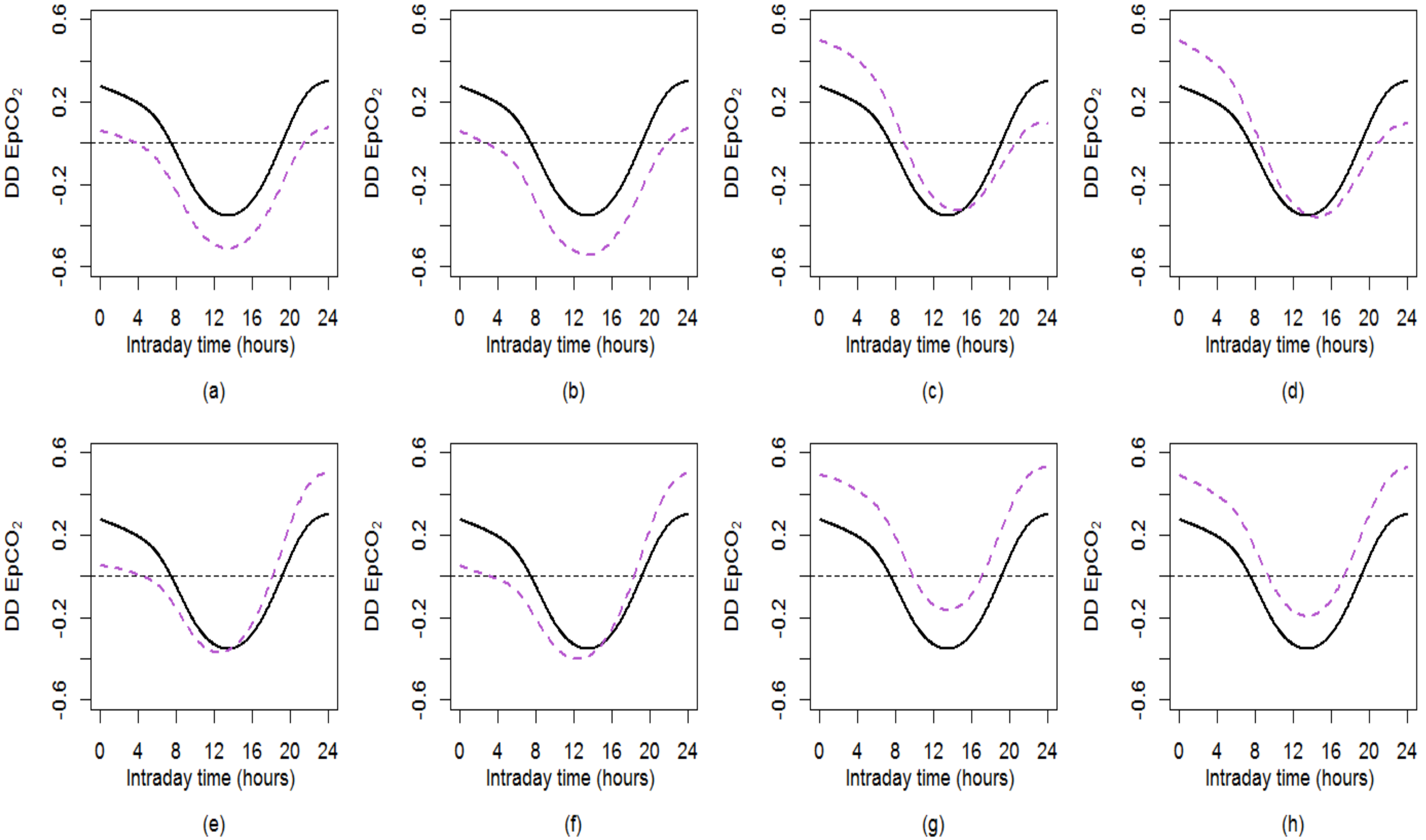


FIGURE 5.29: The functional mean $\hat{\mu}(t)$ of the DD EpCO$_2$ daily curves (solid line) and $\hat{\mu}(t) + \overline{\text{eff}}(\delta_{-1}, \delta_0, \delta_1)$ with $\delta_l = \pm 1$ of $\hat{\phi}_{1l}$(dashed line) for the purple class (a) $(\delta_{-1}, \delta_0, \delta_1) = (-1, -1, -1)$, (b) $(\delta_{-1}, \delta_0, \delta_1) = (-1, -1, 1)$, (c) $(\delta_{-1}, \delta_0, \delta_1) = (-1, 1, -1)$, (d) $(\delta_{-1}, \delta_0, \delta_1) = (-1, 1, 1)$, (e) $(\delta_{-1}, \delta_0, \delta_1) = (1, -1, -1)$, (f) $(\delta_{-1}, \delta_0, \delta_1) = (1, -1, 1)$ (g) $(\delta_{-1}, \delta_0, \delta_1) = (1, 1, -1)$, (h) $(\delta_{-1}, \delta_0, \delta_1) = (1, 1, 1)$

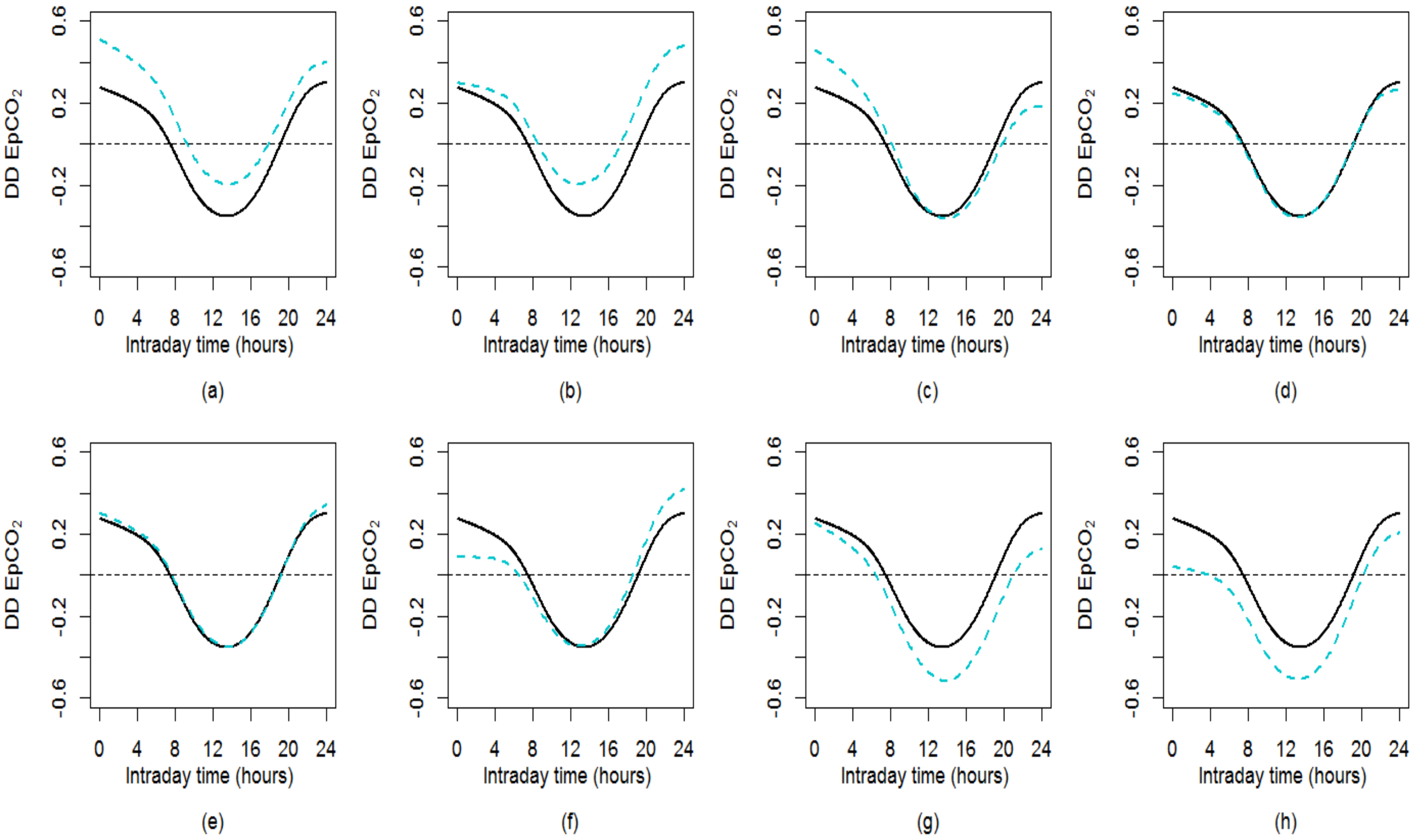


FIGURE 5.30: The functional mean $\hat{\mu}(t)$ of the DD EpCO$_2$ daily curves (solid line) and $\hat{\mu}(t) + \overline{\text{eff}}(\delta_{-1}, \delta_0, \delta_1)$ with $\delta_l = \pm 1$ of $\hat{\phi}_{1l}$(dashed line) for the turquoise class (a) $(\delta_{-1}, \delta_0, \delta_1) = (-1, -1, -1)$, (b) $(\delta_{-1}, \delta_0, \delta_1) = (-1, -1, 1)$, (c) $(\delta_{-1}, \delta_0, \delta_1) = (-1, 1, -1)$, (d) $(\delta_{-1}, \delta_0, \delta_1) = (-1, 1, 1)$, (e) $(\delta_{-1}, \delta_0, \delta_1) = (1, -1, -1)$, (f) $(\delta_{-1}, \delta_0, \delta_1) = (1, -1, 1)$ (g) $(\delta_{-1}, \delta_0, \delta_1) = (1, 1, -1)$, (h) $(\delta_{-1}, \delta_0, \delta_1) = (1, 1, 1)$

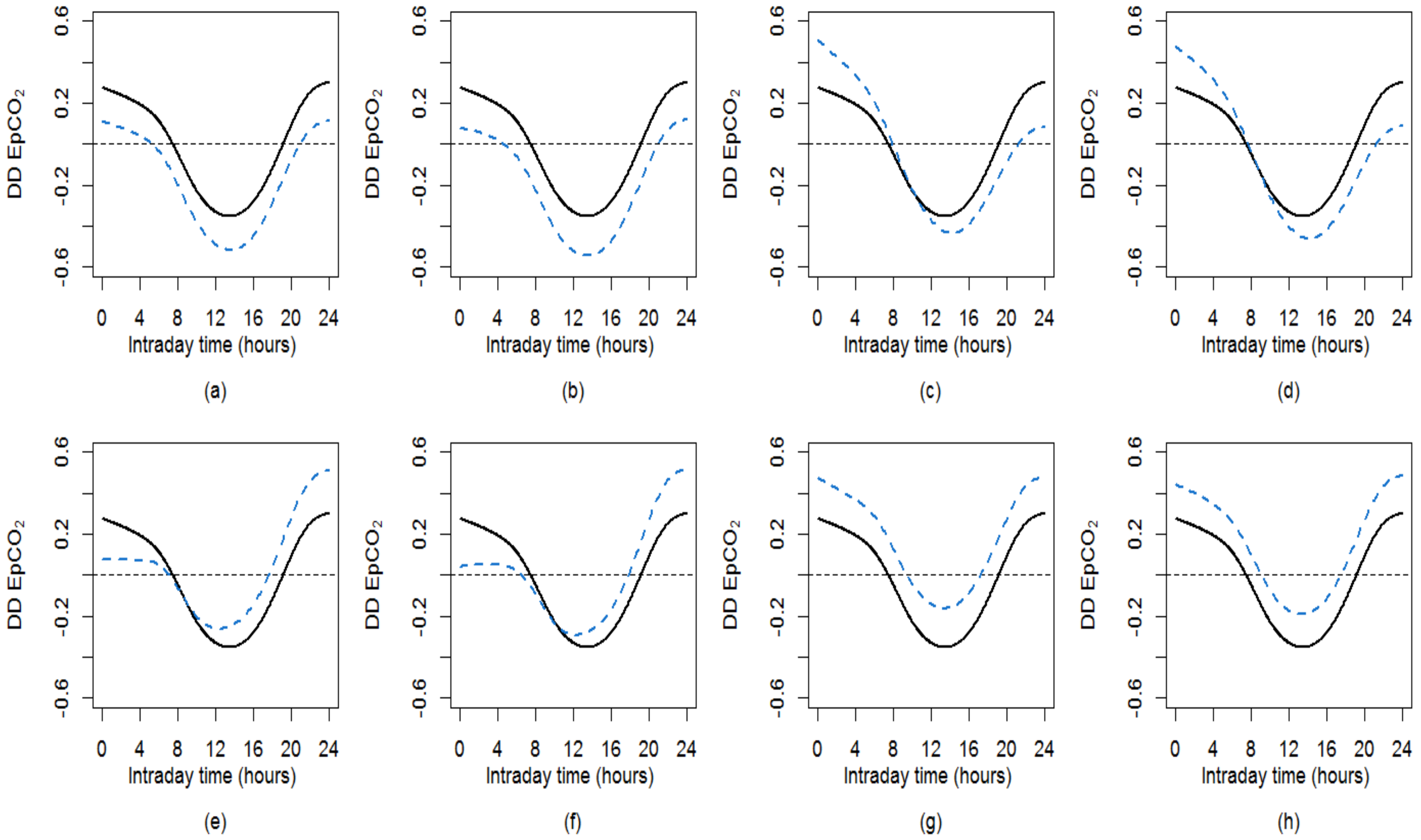


FIGURE 5.31: The functional mean $\hat{\mu}(t)$ of the DD EpCO$_2$ daily curves (solid line) and $\hat{\mu}(t) + \overline{\text{eff}}(\delta_{-1}, \delta_0, \delta_1)$ with $\delta_l = \pm 1$ of $\hat{\phi}_{1l}$(dashed line) for the blue class (a) $(\delta_{-1}, \delta_0, \delta_1) = (-1, -1, -1)$, (b) $(\delta_{-1}, \delta_0, \delta_1) = (-1, -1, 1)$, (c) $(\delta_{-1}, \delta_0, \delta_1) = (-1, 1, -1)$, (d) $(\delta_{-1}, \delta_0, \delta_1) = (-1, 1, 1)$, (e) $(\delta_{-1}, \delta_0, \delta_1) = (1, -1, -1)$, (f) $(\delta_{-1}, \delta_0, \delta_1) = (1, -1, 1)$ (g) $(\delta_{-1}, \delta_0, \delta_1) = (1, 1, -1)$, (h) $(\delta_{-1}, \delta_0, \delta_1) = (1, 1, 1)$

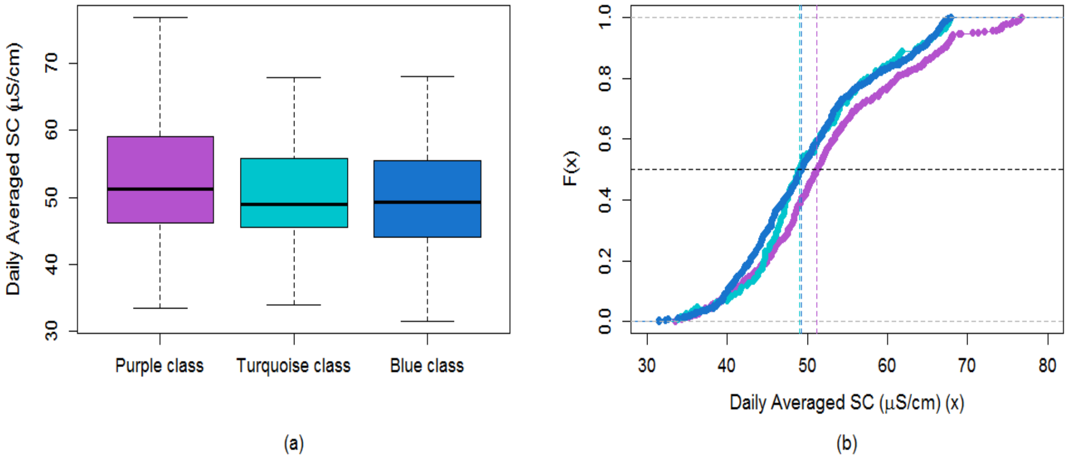


FIGURE 5.32: (a) Box-plots and (b) empirical cumulative distribution plots (probability plots) of the daily averaged SC in the 3 clusters of daily $EpCO_2$ curves identified based on the 3 central functional filters of the first smooth dynamic FPC. The horizontal dashed black line is the 50% quantile of the data and the vertical dashed purple, blue and turquoise lines correspond to the median of SC in each cluster.

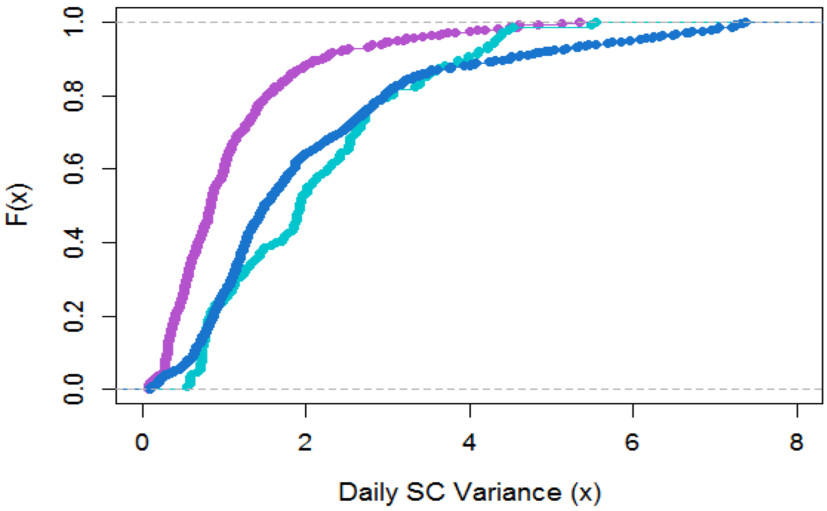


FIGURE 5.33: Empirical cumulative distribution plots (probability plots) of the daily SC variances in the 3 clusters of daily $EpCO_2$ curves identified based on the 3 central functional filters of the first smooth dynamic FPC.

5.8 Summary and Discussion

In real life, most of the environmental systems are highly dynamic and non-stationary over time. This may result in a covariance structure and a spectral density that evolve over time. In such case, the traditional FPCs and even the dynamic FPCs proposed by

Hormann et al. (2014) will not provide an optimal dimension reduction representation for the data as they do not adapt to the process changes over time. Therefore, in this chapter, a time-varying dynamic FPCA is proposed as a means of assessing whether and how the spectral density varies over time. To detect the statistically significant changes over time in the process's spectral density, a bootstrap inferential procedure based on the time-varying dynamic FPC is suggested.

The proposed methodology involves obtaining the orthogonal dynamic FPCs at each time point i by smoothing the lag h covariance matrices using a weighting kernel that assigns higher weights to the nearby observations and less weights to the further ones. The weighting kernel controls the amount of neighbouring data contributing to the estimated lag h covariances through a smoothing parameter. Although a Gaussian weighting kernel is used for our application and simulation study, alternative weight functions can be used to adapt the method of smoothing to the nature of the studied time series.

The proposed time-varying dynamic FPCs, compared to the dynamic FPCs, have provided a more appropriate approximation to the complex structure in the EpCO$_2$ daily curves, where the smoothing parameter was chosen on the basis of a sensitivity analysis. A simulation study has also proven that the smooth dynamic FPCA outperforms the ordinary dynamic FPCA in almost all settings but quite significantly when serial dependence is not very strong. This is mainly because the functional process becomes quite smooth and exhibits less variability at high correlation. In terms of the error between the true eigenvalues and the estimated ones from both the smooth dynamic and ordinary dynamic FPCs, the smooth dynamic FPCs have shown to provide superior estimation (smaller squared error) in the case of non-stationary functional time series. This result implies that the ordinary dynamic FPCs might not provide the optimal dimension reduction for non-stationary functional time series; and therefore an equivalence that varies over time is needed to adjust for changes in the covariance and power spectra. Oppositely, in the stationary case, the ordinary dynamic FPCA provides better estimation for the true eigenvalues; whereas the smooth dynamic FPCA tends to introduce more artefacts and bias.

The time-varying dynamic FPCs were subsequently used to determine whether or not the spectral density of the EpCO$_2$ daily curves changes throughout time. Using a bootstrap inferential procedure, the test indicated the significant changes in the spectral characteristics of the process over time. Based on that result, the ordinary dynamic FPCs will not provide an optimal dimension reduction to the data and hence a more appropriate time-varying version is needed.

After summarizing the functional data by their corresponding smooth dynamic FPCs, we proposed clustering the curves based on the corresponding functional filters that vary over time. Here, we suggest classifying the daily curves of $\mathrm{EpCO_2}$ based on the coefficients of the 3 central filters of the first smooth dynamic FPC which appears to be the most influential filters, on average. Clustering based on the functional filters accounts for the overall average effect of the 3 central functional filters, at the lags -1, 0 and 1, on the mean functional curve not just the filtered effect at some specific lags. A number of clusters between 2 and 6 seems plausible. However, a K-means algorithm based on 3 centers is performed for the matter of comparison with the clustering structure based on the dynamic FPC scores. The classification of the curves based on the time-varying functional filters looks quite different from that based on the 3 central consecutive scores of the dynamic FPCs. The resulting clustering structure is smoother than that based on the dynamic FPC scores, showing the presence of two dominant and alternating patterns that persist over long time periods in addition to some short transitional periods of a different pattern.

The functional filters differ from one class to another, implying a different dependence structure between the observations in the 3 clusters. These differences in the covariance structure can be attributed to the underlying climatological and hydrological conditions including short/long daytime, warm/cool temperature, clear/turbulent water, heavy/shallow storms, and continuous/intermittent rainfall events. Each mixture of those conditions affects the extent and degree of dependence between the curves differently, which in turn influences the functional filters and their contribution to the construction of the current time curve. Following from this, we found that one of the identified clusters is driven by the biological activity while the others are driven by different hydrological conditions.

The cluster driven by the biological activity involves the periods with less correlated days, where the effects of neighbouring days on the $\mathrm{EpCO_2}$ of the current day become negligible after a very small number of lags ($\sum_{l=-1}^{1} \|\overline{\hat{\phi}_{1l}}\|^2 \approx 0.8$). One of the other two classes includes the periods characterized by a correlation structure that persists over a large number of lags, where the estimated filters' effects fade down quite slowly ($\sum_{l=-1}^{1} \|\overline{\hat{\phi}_{1l}}\|^2 \approx 0.4$). This class appears to underlie the periods with less within-day SC variability and less carbon in the catchment resulting from rainfall events that do not drive soil water to flow into the catchment or long-term effect events that dilute the $\mathrm{CO_2}$ in the catchment. The last class comprises the periods characterized by a correlation structure that fades down more rapidly ($\sum_{l=-1}^{1} \|\overline{\hat{\phi}_{1l}}\|^2 \approx 0.7$). This class generally describes the periods with heavy and multiple short-term effect rainfall events that bring more soil water, rich in $\mathrm{CO_2}$, to the surface water.

In conclusion, determining the drivers and characteristics of the different $EpCO_2$ daily patterns is quite challenging and subject to different factors. This is mostly attributed to the complex dynamics of river systems, serial dependence between measurements, multi-drivers of environmental changes, in addition to the events causing disturbances to the system. Advanced supervised and unsupervised learning methods, that account for the inter-relationships between the $EpCO_2$ and the different hydrological and climatological conditions, are appealing to gain a better understanding of the $EpCO_2$ daily patterns and their drivers. For instance, Spezia et al. (2011) have developed a multivariate Hidden Markov Model (HMM) to identify the hidden states and the number of states underlying the water quality of some Scottish rivers. In this model, a Bayesian inference procedure is employed to estimate the unknown number of states. According to this methodology, the authors were able to identify a number of states reflecting different hydrological regimes; some of these states were persistent for long periods and others were intermittent. The method proposed has provided information on the covariance between series of water quality variables trying to understand the drivers of environmental changes, though interpreting the states is again complex and not straightforward. This is because the river systems are typically highly dynamic and responsive to both internal and external factors. Other attempts for identifying states and clusters in a river system include those proposed by Kannan and Ghosh (2011) and Whiting (2006). Kannan and Ghosh (2011) have introduced a down-scaling methodology coupled with multivariate techniques, including cluster analysis and principal component analysis, to predict the state of rainfall on a river basin from large scale climatological data. Three rainfall states "almost dry", "medium" and "high" were identified and the state-to-state transitional probabilities were calculated, indicating the persistence of states. The authors have highlighted that different transitions can lead to different hydrological conditions in the future. Alternatively, Whiting (2006) has suggested the use of auto-regressive HMMs and hidden semi-Markov models in modeling monthly rainfall data and identifying wet and dry states. This work has demonstrated the presence of spatially-consistent patterns of persistent wet and dry seasons. All these methods confirm the importance of the covariance structure between the different determinants across space and time in identifying the states underlying a certain hydrological system.

Chapter 6

Conclusions, Discussions and Future Work

The tremendous advances in sensor technology allow environmental monitoring programmes to collect and store measurements more efficiently at any arbitrary high-temporal resolution. These high-frequency data promote an increasingly comprehensive picture of many environmental processes, which are in reality continuous in time. However, benefiting from this increasing data resource poses various challenges in terms of statistical modeling using standard methods and software tools. Therefore, there is a demand for more advanced statistical methods that extract previously inaccessible information while accounting for the persistent and long memory serial correlation between observations, the complex and varying dynamics over the different timescales and the inter-relationships between the different drivers of the process.

In this thesis, a variety of statistical methods in both time and frequency domains have been used and developed to effectively explore and analyze hydrological high-frequency time series as well as to optimally reduce their dimensionality. Firstly, the statistical challenges of exploring, modeling and analyzing such large volumes of high-frequency time series have been investigated. Thereafter, more advanced statistical techniques have been applied and developed to: (i) better visualize and identify the different modes of variability in the data; (ii) classify and explain the common patterns in the data; and (iii) provide a more adequate dimension reduction representation, which takes into account the persistent serial dependence between observations. The methodology considered in this thesis has been extended to non-stationary time series whose covariance structure varies over time. The proposed method has been further employed to statistically evaluate, in the frequency domain, whether or not the covariance structure changes. Although a 15-minute resolution time series of excess partial pressure of carbon dioxide obtained for a small catchment in the River Dee in Scotland has been used as an illustrative data set throughout the thesis, the techniques used and developed are of general applications to any high-frequency time series which opens doors for a variety of future research.

This chapter is organized as follows. Section 6.1 summarizes the main findings of the wavelet analysis used as an exploratory tool for visualizing the variability in a high-frequency time series. Section 6.2 briefly explains the results and challenges of fitting additive models with multivariate interaction terms for the high-frequency and long memory $\mathrm{EpCO_2}$ time series. Section 6.3 discusses the results of statistically analyzing high-frequency time series using a functional data analysis approach. In this section, the results of analyzing the daily curves of $\mathrm{EpCO_2}$ ignoring and taking into account the time dependence structure between the curves using a frequency domain approach are compared. Next, the methodology proposed to statistically evaluate and account for the temporal changes in the covariance structure/spectral density of the process and its results are discussed. Finally, Section 6.4 suggests a variety of possible extensions to the developed methodology.

6.1 Wavelet Analysis

It is evident from the primary exploratory analysis that the $\mathrm{EpCO_2}$ exhibits variations over different timescales. To gain a better understanding of how the $\mathrm{EpCO_2}$ varies over time at the various timescales, wavelet analysis has been employed. Wavelets have proven to be an effective and efficient tool for deconstructing a high-frequency time series into several components then exploring the variability at each of those components. The maximal overlap discrete wavelet transform was preferred over the discrete wavelet transform to avoid the sample size limitations discussed in Chapter 1. To line up features in the filtered series with the original series, the least asymmetric filter was used here to circularly filter the original signal. The multiresolution analysis (MRA) based on the wavelet transform has helped visualize and analyze the $\mathrm{EpCO_2}$ high-frequency time series in the time and frequency domains simultaneously; and has been used, in particular, to identify the dominant seasonal and cyclical variations in the time series.

The wavelet analysis has supported the evidence of intra-daily, seasonal and inter-annual variations in the $\mathrm{EpCO_2}$ series. These variations can be attributed to changes in the relative strength of external (e.g., climatological) and internal (biological processing) drivers of resultant $\mathrm{EpCO_2}$. The MRA of the $\mathrm{EpCO_2}$ series within each hydrological year has indicated that an 8-hour signal is the major contributor to the variability of the

$EpCO_2$ series. This feature is considered representative for the intra-daily variability, reflecting the dark-light-dark cycle within the day. The amplitude of this diel cycle was found varying throughout the year with larger variability observed in summer when a pronounced diurnal cycle is present. It has also been noted that the variability in $EpCO_2$ resulting from the daylight cycle changes from one year to another, reflecting different balances of external and internal drivers of $EpCO_2$. Based on the smooth component of the wavelet transform, the $EpCO_2$ series exhibited a repeated seasonal pattern over the 3 hydrological years. Although this seasonal pattern is mainly characterized by high levels of $EpCO_2$ in winter relative to summer; its magnitude varies from one year to another. These changes in the variability from one year to another reflects the inter-annual variations and the non-stationarity of the series over the whole study period.

By comparing the MRA of the $EpCO_2$ series to that of water discharge, temperature, pH and specific conductivity, we found that the highest $EpCO_2$ variability is usually associated with little changes in discharge, consistent with internal fluvial carbon cycling, while the hydrological events are oppositely associated with compressed $EpCO_2$ variability. It has also been found that the variability in $EpCO_2$ evolves coherently with the variability in temperature and that the $EpCO_2$ is more variable during summer when there are larger fluctuations between day and night temperatures. These results suggest that river hydrodynamics contribute to the $EpCO_2$ variability but whether the temporal patterns in $EpCO_2$ can be described entirely or partially by hydrology was not clear yet. Therefore, additive models have been employed to investigate and describe the nature of these relationships.

6.2 Additive Models

Additive models have proven here to be very powerful and useful at (i) explaining the temporal variability in a high-frequency time series at the different timescales, and (ii) describing the multivariate non-parametric relationships between the response variable and the explanatory variables. The fitted additive models have indicated that high-frequency time series are subject to long memory and persistent correlation structure. A suitable description of this temporal correlation structure has been incorporated in the analysis of additive models, to ensure that the standard errors reflect the form of variation exhibited by the data. These additive models and large volumes of high-frequency data have allowed both the temporal variations and the mechanisms controlling the system changes to be accommodated through the use of bivariate smooth terms, expressing the interaction between two covariates.

The hierarchical set of additive models fitted over a day, a month and a year has revealed that the variability in $EpCO_2$ and its relationship with water hydrology are time and scale dependent. It has been concluded from the fitted models that the temporal variations in $EpCO_2$ as well as the multivariate relationships with river hydrology change across the different timescales and become more composite as the model is extended to cover a longer time period within the hydrological year. The fitted models have shown, specifically, that the $EpCO_2$ exhibits a 24-hour dark-light-dark cycle reaching the minimum value at noon and that the magnitude of this intra-daily cycle changes along the year and becomes more apparent during summer when the $EpCO_2$ reaches its maximum levels.

In addition to the temporal variations in $EpCO_2$, the fitted additive models have demonstrated that the river hydrology co-varies with the $EpCO_2$ and its intra-daily cycle. In particular, the intra-daily cycle dominates the $EpCO_2$ variations in summer during the absence of hydrological events and during low flows (evidenced by higher specific conductivity), while high flow events tend to dilute the DIC pool which dampens the diel cycle in winter. The relative contribution of these temporal and hydrological variations to the changes in $EpCO_2$ varies over time and from one timescale to another. Consequently, the additive models have encountered problems in uniquely identifying the sources of variability and the contribution of each variable in the interaction terms to the variability in $EpCO_2$. For this reason, a sensitivity analysis for choosing the basis dimension of each variable in the smooth bivariate terms has been performed first before automatically choosing the smoothing parameter using the REML criterion.

The residuals of the fitted additive models have shown a periodic auto-correlation structure that persists over a large number of lags. This dependence structure has not been accounted for by the different covariates in the model, especially at the higher timescales. Consequently, modeling these high-frequency data assuming independence is no longer valid. A two-stage fitting procedure has been rather employed, where some lagged dependent variables are added to the model; then an AR process is fitted to the adjusted model residuals and the standard errors of the estimates are updated. An alternative method that accounts for the correlation structure through fitting a GAMM has been also considered. However, incorporating the auto-correlation structure as a mixed effect component while fitting the additive model resulted in automatically selecting larger smoothing parameters which leaves the remaining correlation structure very complex to be accommodated using simple time series models. Whereas in the two-stage procedure of fitting, the first stage has resulted in optimally selecting smaller smoothing parameters assuming independent errors, using the REML criterion, which reduces the complexity of modeling required for the residuals in the second stage. Although autocorrelation can influence the smoothing parameter selection from automatic approaches, GAMMs appear to be computationally inefficient with large time series and numerically unstable because of the confounding between correlation and non-linearity.

6.3 Functional Time Series Analysis

Both the wavelet analysis and the additive models indicated that the EpCO$_2$ exhibits a dominant intra-daily cycle that varies over time depending on the underlying climatological (seasonal) and hydrological conditions. Following from this, a functional data analysis (FDA) approach has been developed to analyze these intra-daily patterns; then identify the main sources of variability between those daily curves as well as the most common daily patterns in the data and the drivers of each. FDA has shown to be an appropriate tool for extracting the most important characteristics of some possibly high dimensional data and reducing the data dimensionality by *stringing* the high-dimensional discrete data into functional data. This view of the data allows the patterns in a high-frequency time series to be studied without being concerned about the high-correlations between the (high-frequency) measurements within the same functional unit. Following this approach, functional principal component analysis and functional clustering analysis are considered very useful techniques for identifying the sources of variability and the common patterns in a functional data, respectively.

In this thesis, the 96 (15-minute) observations within each day have been considered as discrete observations of a continuous smooth daily function, which represents a great reduction in the dimension of the data under study. The initial classification of the daily EpCO$_2$ curves based on either their basis functions' coefficients or their first few FPCs has mainly relied on the mean level rather than any local features. This clustering structure was believed to be driven by different combinations of seasonal and hydrological conditions. Each combination has a different effect on the biological and hydrological activity of the catchment, inducing a particular daily pattern that reflects the processing of EpCO$_2$ along the day. For instance, in warm and dry periods the biological activity becomes more dominant and more CO$_2$ tends to be consumed during the day; whereas in wet periods the heavy rainfall dilutes the DIC pool in the catchment and dampens the intra-daily cycle of EpCO$_2$.

To avoid the effects of trend and seasonality on the mean level from dominating the clustering structure of the curves, the global trend as well as the seasonal effects have been estimated then subtracted from the discrete data before any further analysis. The FPCA of the detrended and deseasonalized EpCO$_2$ daily curves has suggested that the primary modes of variability are the deviations from the average daily pattern of EpCO$_2$ and the contrast between day and night levels of EpCO$_2$. By limiting the effects of trend and seasonality on the clustering structure of the curves, the water hydrology, to a great extent, has potentially become the major responsible for the resulting classification.

The major shortcoming of FPCA and FCA is that they have ignored the persistent time dependence between successive daily curves. Consequently, all the information carried by nearby observations have been discounted. This in turn could lead to inappropriate dimensionality reduction and misleading clustering structure. This limitation has motivated the use of the recently developed dynamic FPCA (Hormann et al., 2014) that accounts for the time dependence structure in the frequency domain.

6.3.1 Accounting for Temporal Correlation

In principle, dynamic FPCA relies on the spectral decomposition of the spectral density operator that contains all the information on the whole family of covariance operators - not just the covariance operator at time lag 0. In this thesis, dynamic FPCA has been extended to appropriately reduce the dimensionality of temporally dependent functional data estimated using any type of basis functions, not only orthogonal basis functions. The proposed general version of dynamic FPCA has been applied to the daily curves of EpCO_2, approximated using penalized cubic B-splines. The dynamic FPCs have proven to provide a better approximation to the original curves in smaller dimensions, compared to that using traditional FPCs, retrieving both the overall pattern and the local features of the curves taking into account the correlation structure in the data. The dynamic FPCs, unlike the traditional FPCs, are interpreted sequentially by studying the effect of a sequence of consecutive scores on the current day level. The filter elements of the dynamic FPCs were found to fade down quite rapidly as the lag increases demonstrating the dominant effect of the most central filters of the first dynamic FPCs on the current day level. Following from this result, we have proposed classifying the functional data based on the most central lags of the first dynamic FPC scores using the multivariate K-means algorithm.

Clustering the daily curves of EpCO_2 based on the dynamic scores of the first 2 FPCs at the central lags -1,0 and 1 has appeared to be mainly driven by the discrepancies in the mean level as well as the dependence structure between the curves. To study the hydrological drivers of each daily pattern, further statistical analysis of the specific conductivity in each cluster has been employed. This analysis has shown that one of the identified clusters mainly underlies the hydrologically stable periods and the others describe either the periods with short-term conductivity drops that quickly recover back to its normal level or the periods with persistent long-term hydrological events. This clustering structure is clearly different from that based on the traditional FPC scores. This is because the FPCA and the dynamic FPCA are based on entirely different methodology and are also loading different functions. Accounting for the time dependence structure using the dynamic FPCs has taken advantage of the potential information carried by nearby observations, and therefore a more appropriate dimension reduction for the data was provided.

6.3.2 Testing and Accounting for Non-stationarity

By definition, the dynamic FPCA depends in their construction on the assumption of second-order stationarity. This is considered a weakness, especially when the time series under study is believed to have a complex covariance structure that varies over time. A preliminary novel bootstrap stationarity test, based on comparing the dynamic FPCs' eigenvalues across a sequence of time blocks covering the whole signal under study, has shown significant changes in the spectral density over time. This result implies that the obtained dynamic FPCs might not offer an optimal dimension reduction for the data as they are static and do not adapt to the changes in the process over time. Therefore, smooth dynamic FPCA has been proposed, in Chapter 5, as a means of obtaining time-varying dynamic FPCs. These time-varying dynamic FPCs can be subsequently used to investigate whether and how the spectral density and covariance structure change over time.

According to the proposed methodology of smooth dynamic FPCA, the orthogonal dynamic FPCs are obtained at each time point i by smoothing the lag h covariance matrices using a weighting kernel that assigns higher weights to the nearby observations and less weights to the further ones. The weighting kernel controls the amount of neighbouring data that contributes to the the estimated covariance through a smoothing parameter. Here, a Gaussian weighting kernel has been employed and the smoothing parameter has been chosen based on a sensitivity analysis study. Note that alternative weighting kernels can be used depending on the nature of the studied time series. The proposed smooth dynamic FPCs are shown to provide a better approximation for the original curves using a smaller number of principal components that changes over time.

A novel simulation study, inspired by the one described in Hormann et al. (2014), has been presented in Chapter 5 to assess the performance of the smooth dynamic FPCs relative to the dynamic FPCs. Based on that simulation study, the smooth dynamic FPCs have proven to outperform ordinary dynamic FPCs in almost all settings but quite significantly when serial dependence is not very strong. This result can be justified by the fact that the process exhibits less variability at high correlations. In terms of estimating the true eigenvalues of the spectral density, the smooth dynamic FPCs have shown a considerable improvement (by having smaller squared errors between the true and estimated eigenvalues) when compared to ordinary dynamic FPCs in the case of non-stationary functional time series. This implies the inappropriateness of using ordinary dynamic FPCA with non-stationary functional data and supports the need for an alternative data dimensionality reduction that varies over time with the changes in the covariance structure and spectral density. On the contrary, the ordinary dynamic FPCs have provided better estimation for the true eigenvalues in the case of stationary functional time series, whilst the smooth dynamic FPCs have introduced more artefacts and bias by overestimating the true eigenvalues.

The simulation study has also shown that caution should be taken while choosing the smoothing parameter controlling the width of the weighting kernel and hence the amount of neighbours contributing to the estimation of the covariance structure. Employing a very small smoothing parameter tends to over-estimate the true eigenvalues, while using a relatively large smoothing parameter tends to overlook the changes in the process. As the smoothing parameter increases, the estimated time-varying dynamic eigenvalues approach the (constant) dynamic ones. Thus, the smoothing parameter controls the trade-off between over-fitting the data and smoothing the time-varying dynamic FPCs to account for the process changes.

The time-varying dynamic FPCs have been used subsequently in determining whether and when the spectral density of the EpCO$_2$ daily curves changes significantly throughout time, using a novel bootstrap inferential procedure. The proposed test has shown significant changes in the spectral characteristics of the EpCO$_2$ process over time. The first eigenvalue of the spectral density has been significantly lower, on average, in winter relative to summer. This result has supported the non-stationarity of the functional time series overall and at particular frequencies; and hence each curve is better approximated using the corresponding time-varying dynamic FPCs.

The functional filters obtained at each time point, using the smooth dynamic FPCA methodology, present the effect of the neighbouring curves on that time point. Following from this, a novel clustering approach based on the coefficients of the functional filters that vary over time has been proposed. The EpCO$_2$ daily curves have then been classified based on the coefficients of the most 3 central filters of the first smooth dynamic FPC which appeared to be the most influential filters, on average. A K-means clustering algorithm has been performed and the results have been compared to that based on the dynamic FPC scores. The clustering structure appeared to be smoother than that based on the dynamic FPC scores, showing the presence of some dominant patterns that persist over long periods in addition to some short transitional periods of a different pattern. The discrepancy between the classes here relies on the differences in the functional filters reflecting different dependence structures between observations, which can be attributed to the underlying climatological and hydrological conditions, in the sense that each mixture of conditions like short/long daytime, warm/cool temperature, clear/turbulent water, heavy/shallow storms, and continuous/intermittent rainfall events would affect differently the range and strength of the serial correlation structure in the data.

The clusters identified based on the time-varying functional filters are possibly driven by either biological activity or river hydrology. Biological activity is believed to be the main driver of the cluster involving the periods where the effect of neighbouring days on the $\mathrm{EpCO_2}$ of the current day becomes negligible after a smaller number of lags. The other two classes are driven by different hydrological conditions. One class mostly underlies the periods characterized by a persistent correlation structure, where the estimated filters' effects fade down quite slowly, describing the periods with less within-day SC variability and less carbon in the catchment. The other class includes the periods characterized by a rapidly decaying auto-correlation structure; ergo underlies the periods with heavy and multiple short-term effect rainfall events that bring more soil water, rich in $\mathrm{CO_2}$, to the surface water.

Note that, the 3 hydrological years under study have experienced different hydrological conditions. The HY2003/2004 is relatively wet in both summer and winter with an average total water discharge of 9,000 and 10,000 $\mathrm{m^3/s}$, respectively. In contrast, the HY2004/2005 has drier summer and winter with average total discharge of 4,000 and 6,500 $\mathrm{m^3/s}$ respectively; Whereas the HY2005/2006 is characterized by a relatively drier summer and wetter winter with 4,000 and 12,400 $\mathrm{m^3/s}$ average total discharge, respectively. Therefore, although the different daily patterns of $\mathrm{EpCO_2}$ have proven to be driven by different hydrological and climatological conditions, validating the conclusions drawn about the drivers of each pattern from any of the above clustering results still needs further investigation. A longer time period, involving more hydrological years, would be desirable so that one could repeat the same analysis to check whether the clustering structure repeats or not under the same hydrological and climatological conditions.

At last, it is concluded that determining and understanding the drivers and characteristics of the different $\mathrm{EpCO_2}$ daily patterns is quite challenging and subject to different factors. This is attributed to the complex dynamics of river systems, multi-drivers of environmental changes, different climatological conditions, temporal dependence between measurements, and system disturbances caused by irregular events. Some classifications have been based on the mean level while others have been based on the dependence structure in the data. Therefore, the objective of clustering has to be set in advance so that the most appropriate method of clustering is used. Advanced classification methods, that account for the inter-relationships between the EpCO_2 and the different hydrological and climatological conditions, are also required to gain a better understanding of the EpCO_2 daily patterns and their drivers.

6.4 Future Work

There are several potential extensions to the statistical methodology applied in this thesis to analyze high-frequency data in general and environmental high-frequency data in particular. The nature of these high-frequency data yet poses additional statistical challenges on the possible future developments.

In some situations, modeling the remaining dependence structure in the residuals of the fitted additive models using traditional time series models such as ARMA(p,q) might not be satisfactory. Such time series models do not necessarily reflect the complex correlation structure of the data including the long range dependence structure and the changes in variability and correlation throughout time. To account for such complex temporal dependence structure, it might be worth investigating in the future the use and developments of long memory heteroscedastic time series models (e.g. GARCH models) for large volumes of high-frequency time series.

In this thesis, the EpCO_2 daily curves have been always clustered using the K-means algorithm by first filtering the functional data using their basis functions' coefficients or their first few (time or frequency domain) FPC scores followed by a conventional multivariate K-means clustering. Alternative methods of unsupervised classification including model-based clustering techniques can be considered depending on the objective and nature of study. In model-based clustering, each observation is accompanied with a probability of cluster membership and hence is considered appropriate if further formal inference is of interest. One limitation of the model-based clustering techniques is that they could be computationally inefficient especially with large volumes of data. Another limitation is the increased difficulty of maximum likelihood estimations arising from the complex and long memory covariance structures characterizing high-frequency time series.

To account for the temporal dependence structure in a functional time series, it may be advantageous to study the lagged effect of covariates on the response variable. Distributed-lag models are typically used to model the temporal dependence structure of the dependent variable on the covariate when the effect of a covariate is delayed and distributed through time. Like functional regression, distributed-lag models can be extended to situations where the dependent variable or the covariate or even both are functions of time. Such models can be used to relate the value of the functional response at a ceratin time point to the value of the functional covariates at the same time point and the previous time points. These modeled functions can then be used within a functional cluster analysis.

Other potential future work directions are motivated by the recent developments involving the extension of the K-means algorithm to cluster functional data in the function domain (Garcia et al., 2015). This method can be extended to cluster functional time-series data using some smoothing techniques to account for the temporal correlations between successive curves. Moreover, this functional K-means clustering algorithm can be extended to a multivariate setting and so the different regimes in a hydrological system can be identified based on multiple functions of different attributes.

In smooth dynamic FPCA, the weighted spectral density of the whole functional time series is decomposed at each time point. Accordingly, this procedure consumes high computational time and large memory. One possible way to minimize the computational cost is to constrain the weighting matrix and hence the spectral density matrix at each time point to be sparse, which can subsequently speed up the eigen-decomposition of the spectral densities.

The smooth dynamic FPCs are obtained such that the smoothing parameter used to compute the covariance operators at all lags h is the same across the whole signal. Nevertheless, the amount of observations contributing to the estimation of the covariance structure at each time point might vary over time if the series is covariance non-stationary. This situation requires the use of an adaptive smoothing parameter that varies over time to provide a more reliable estimate for the covariance structure at each time point. Estimating this adaptive smoothing parameter at each time point introduces an additional computational cost and complexity to the methodology applied. Therefore, ways of efficiently selecting smoothing parameters need to be further investigated.

The dynamic FPCs and proposed smooth dynamic FPCs have successfully accounted for the temporal dependence in functional time series. Nonetheless, other applications involve functional data that are rather spatially dependent. Examples include annual temperatures measured on the earth's surface, water quality trends for a river network, etc. Following from this, future work could involve extending these dynamic FPCs and smooth dynamic FPCs to account for the spatial dependence between curves in a large spatial network. Further developments can be devoted to the study of dominant spatial and temporal patterns in a spatio-temporal framework by extending the T-mode and S-mode PCA to a functional dynamic setting. Examples of applications include daily patterns of EpCO_2 or any geophysical and environmental data across different locations.

Bibliography

Abraham, C., P. Cornillon, E. Matzner-Lober, and N. Molinari (2003). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics 50*(3), 581–595.

Adelfio, G., M. Chiodi, A. D'Alessandro, and D. Luzio (2010). Functional principal components direction to cluster eathquake waveforms. In *Geophysical Research Abstracts*, Volume 12 of *EGU General Assembly 2010*.

Akaike, H. (1973). Information theory and an extension of maximum likelihood principle. *Second International Symposium on Information Theory, Akademia Kiado*, 267–281.

Banfield, J. and A. Raftery (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics 49*(3), 803–821.

Bartlett, M. (1950). Periodogram analysis and continuous spectra. *Biometrika 37*(1-2), 1–16.

Berkes, I., E. Gombay, and L. Horváth (2009). Testing for changes in the covariance structure of linear processes. *Journal of Statistical Planning and Inference 139*(6), 2044–2063.

Blackman, R. and J. Tukey (1958). *The Measurement of Power Spectra from the Point of View of Communications Engineering.* Dover, New York.

Bouveyron, C. and J. Jacques (2011). Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification 5*(4), 281–300.

Bowman, A. and A. Azzalini (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-plus Illustrations* (1st ed.). Number 18 in Oxford Statistical Science Series. Oxford University Press.

Bowman, A., M. Giannitrapani, and M. Scott (2009). Spatiotemporal smoothing and sulphur dioxide trends over Europe. *Journal of the Royal Statistical Society 58*(5), 737–752.

Box, G. and G. Jenkins (1970). *Time Series Analysis, Forecasting, and Control.* Oakland, CA: Holden-Day.

Brcich, R. and D. Iskander (2006). Testing for stationarity in the frequency domain using a sphericity statistic. *IEEE International Xonference for Acoustics, Speech and Signal Processing 3*, 464–467.

Brillinger, D. (1981). *Time series Data Analysis and Theory.* San Francisco: Holden-Day.

Butman, D. and P. A. Raymond (2011). Significant efflux of carbon dioxide from streams and rivers in the united states. *Nature Geoscience 4*(12), 839–842.

Calinski, R. and J. Harabasz (1974). A dendrite method for cluster analysis. *Communications in Statistics 3*(1), 1–27.

Carlstein, E., K.-A. Do, P. Hall, T. Hesterberg, and H. Kunsch (1998). Matched-block bootstrap for dependent data. *Bernouilli 4*(3), 305–328.

Carvalho, L., C. Miller, E. Scott, G. Codd, P. Davies, and A. Tyler (2011). Cyanobacterial blooms: Statistical models describing risk factors for national-scale lake assessment and lake management. *Science of the Total Environment 409*(24), 5353–5358.

Chen, B., M. Sinn, J. Ploennigs, and A. Schumann (2014). Statistical anomaly detection in mean and variation of energy consumption. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pp. 3570–3575. IEEE.

Cleveland, W. S. (1979, December). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association 74*(368), 829–836.

Cole, J. J., N. Caraco, G. Kling, and T. Kratz (1994). Carbon dioxide supersaturation in the surface water of lakes. *Science 265*(5178), 1568–1570.

Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions. estimating the correct degree of smoothing by the method of generalized crossvalidation. *Numerische Mathematik 31*(4), 377–403.

Daubechies, I. (1992). *Ten Lectures on Wavelets.* Philadelphia: SIAM.

Dawson, J., C. Soulsby, M. Hrachowitz, and D. Telzlaff (2009). Seasonality of EpCO2 at different scales along an integrated river continuum within the Dee Basin, NE Scotland. *Hydrological Processes 23*(20), 2929–2942.

De Boor, C. (2001). *A Practical Guide to Splines* (Revised ed.). Springer.

Dickey, D. A. and W. A. Fuller (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association 74*(366a), 427–431.

Dwivedi, Y. and S. Subba Rao (2011). A test for second-order stationarity of a time series based on the discrete Fourier transform. *Journal of Time Series Analysis 32*(1), 68–91.

Efron, B. and R. Tibshirani (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science 1*(1), 54–75.

Enders, W. (2004). *Applied Econometric Time Series* (2nd ed.). Wiley series in Probability and Mathematical Statistics. Wiley.

Everitt, B., S. Landau, and M. Leese (2001). *Cluster Analysis* (Fourth ed.). New York: Oxford University Press, Inc.

Fahrmeir, L., T. Kneib, and S. Lang (2004). Penalized additive regression for space-time data: A Bayesian prespective. *Statistica Sinica 14*, 731–761.

Fahrmeir, L. and S. Lang (2001). Bayesian inference for generalized additive mixed models based on Markov random fields priors. *Journal of the Royal Statistical Society: Series C 50*(2), 201–220.

Ferguson, C., A. Bowman, E. Scott, and L. Carvalho (2009). Multivariate varying-coefficient models for an ecological system. *Environmetrics 20*(4), 460–476.

Ferguson, C., L. Carvalho, M. Scott, A. Bowman, and A. Kirika (2008). Assessing ecological responses to environmental change using statistical models. *Journal of Applied Ecology 45*(1), 193–203.

Franco-Villoria, M., M. Scott, T. Hoey, and D. Fischbacher-Smith (2012). Temporal investigation of flow variability in Scottish rivers using wavelet analysis. *Journal of Environmental Statistics 3*(6), 1–20.

Garci-Escudero, L. and A. Gordaliza (2005). A proposal for robust curve clustering. *Journal of Classification 22*(2), 185–201.

Garcia, M., R. Garcia-Rodenas, and A. Gomez (2015). K-means algorithms for functional data. *Neurocomputing 151*(1), 231–245.

Giannitrapani, M., A. Bowman, and E. Scott (2011). Additive models for correlated data with applications to air pollution monitoring. In R. Chandler and E. Scott (Eds.), *Statistical Methods for Trend Detection and Analysis in the Environmental Sciences*, Chapter 7, pp. 267–282. Chichester: Wiley.

Giraitis, L. and R. Leipus (1992). Testing and estimating in the change-point problem of the spectral function. *Lithuanian Mathematical Journal 32*(1), 15–29.

Giraldo, R., P. Delicado, and J. Mateu (2012). Hierarchical clustering of spatially correlated functional data. *Statistica Neerlandica 66*(4), 403–421.

Gordon, A. (1996). Null models in cluster validation. In W.Gaul and D.Pfeifer (Eds.), *From Data to Knowledge*, pp. 32–44. New York: Springer.

Grasse, M., M. Frater, and J. Arnold (2000). Testing VBR video traffic for stationarity. *IEEE International Conference for Transactions on Circuits and Systems for Video Technology 10*(3), 448–457.

Haggarty, R., C. Miller, and E. Scott (2015). Spatially weighted functional clustering of river network data. *Journal of Royal Statistical Society: Series C 64*(3), 491–506.

Haggarty, R., C. Miller, E. Scott, F. Wyllie, and M. Smith (2012a). Functional clustering of water quality data in Scotland. *Environmetrics 23*(8), 685–695.

Haggarty, R., C. Miller, M. Scott, F. Wyllie, and M. Smith (2012b). Functional clustering of water quality data in Scotland. *Environmetics 23*(1), 685–695.

Halliday, D., J. Rosenberg, A. Rigas, and B. Conway (2009). A periodogram-based test for weak stationarity and consistency between sections in time series. *Journal of Neuroscience Methods 180*(1), 138–146.

Hartigan, J. and M. Wong (1978). A K-menas clustering algorithm. *Applied Statistics 28*(1), 100–108.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models* (1st ed.). Number 43 in Monographs on Statistics and Applied Probability. Chapman and Hall.

Henderson, B. (2006). Exploring between site in water quality trends: A functional data analysis approach. *Environmetrics 17*(1), 65–80.

Hitchcock, D., G. Casella, and J. Booth (2006). Improved estimation of dissimilarities by presmoothing functional data. *Journal of the American Statistical Association 101*(473), 211–222.

Hormann, S., L. Kidzinski, and M. Hallin (2014). Dynamic functional principal components. *Journal of Royal Statistical Society 77*(2), 319–348.

Hormann, S. and P. Kokoszka (2010). *Handbook of Statistics - Time Series Analysis - Methods and Applications*, Chapter Functional time series, pp. 157–186. Amsterdam - Elsevier.

Horváth, L., M. Hušková, and P. Kokoszka (2010). Testing the stability of the functional autoregressive process. *Journal of Multivariate Analysis 101*(2), 352–367.

Horváth, L., P. Kokoszka, and G. Rice (2014). Testing stationarity of functional time series. *Journal of Econometrics 179*(1), 66–82.

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification 12*(1), 193–218.

Hyndman, R. and H. Shang (2010). Rainbow plots, bagplots, and boxplots for function data. *Journal of computational and Graphical Statistics 19*(1), 29–45.

Ignaccolo, R., S. Ghigo, and E. Giovenali (2008). Analysis of air quality monitoring networks by functional clustering. *Environmetrics 19*(7), 672–686.

Intergovernmental Panel on Climate Change IPCC (2001). The scientific basis. In C. A. Johnson (Ed.), *Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, pp. 882. Cambirdge University Press, New York.

Intergovernmental Panel on Climate Change IPCC (2013). Climate change 2013: The physical science basis.

Jacques, J. and C. Preda (2014). Functional data clustering: A survey. *Advanced Data Analysis Classification 8*(3), 231–255.

Jaimungal, S. and K. Eddie (2007, April). Consistent functional PCA for financial time series. In *FEA 07 Proceedings of the Fourth IASTED International Conference on Financial Engineering and Applications*, pp. 103–108.

James, G. and C. Sugar (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association 98*(462), 397–408.

Jeong, Y., B. Sanders, and S. Grant (2006). The information content of high-frequency environmental monitoring data signals pollution events in the coastal ocean. *Environmental Science Technology 40*(20), 6215–6220.

Jolliffe, J. (2005). *Principal Component Analysis.* Chichester: Wiley.

Kannan, S. and S. Ghosh (2011). Prediction of daily rainfall state in a river basin using statistical downscaling from GCM output. *Stochastic Environmental Research and Risk Assessment 25*(4), 457–474.

Kaufman, L. and P. Rousseeuw (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* New York: Wiley.

Kelsall, J., J. Samer, S. Zeger, and J. Xu (1997). Air pollution and mortality in Philadelphia, 1974-1988. *American Journal of Epidemiology 146*(9), 750–761.

Kirchner, J., X. Fang, C. Neal, and A. Robson (2004). The fine structure of water quality dynamics: The (high-frequency) wave of the future. *Hydrological Processes 18*(7), 1353–1359.

Krzanowski, W. and Y. Lai (1988). A criterion for determining the number of groups in a data set using sum of squares clustering. *Biometrics 44*(1), 23–34.

Kwiatkowski, D., P. Phillips, P. Schmidt, and Y. Shin (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics 54*(1-3), 159–178.

Labat, D. (2005). Recent advances in wavelet analyses: Part 1. A review of concepts. *Journal of Hydrology 314*(1-4).

Lamon, E., K. Reckhow, and K. Havens (1996). Using generalized additive models for prediction Chlorophyll-a in Lake Okeechobee, Florida. *Lakes and Reservoirs: Research and Management 2*(1-2), 37–46.

Li, S., X. Lu, and R. Bush (2013). CO2 partial pressure and CO2 emission in the Lower Mekong River. *Journal of Hydrology 504*, 40–56.

Li, S., X. Lu, M. He, Y. Zhou, and A. Ziegler (2012). Daily CO2 partial pressure and CO2 outgassing in the upper Yangtze River basin: A case study of Longchuanjiang. *Journal of Hydrology 466-467*, 141–150.

Lin, N., J. Jiang, S. Guo, and M. Xiong (2015). Functional principal component analysis and randomized sparse clustering algorithm for medical image analysis. *PLoS ONE 10*(7), 1–17.

Lin, X. and D. Zhang (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society: Series B 55*(2), 381–400.

Liu, C., S. Ray, and G. Hooker (2016). Functional principal component analysis of spatially correlated data. *Statistics and Computing*, 1–16.

Loperfido, J., C. Just, and J. Schnoor (2009). High-frequency diel dissolved oxygen stream data modeled for variable temperature and scale. *Environmental Engineering 10*(12), 1250–1256.

Lopez-Pintado, S. and J. Romo (2009). On the concept of depth for functional data. *Journal of the American Statistical Association 104*(486), 718–734.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. Volume 1 of *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297. University of California Press.

Mandlebrot, B. (1972). Statistical methodology for non-periodic cycles: From the covariance to R/S analysis. *Annals of Economic and Social Measurement 1*(3), 259–290.

Mardia, K., J. Kent, and J. Bibby (1979). *Multivariate Analysis.* New York: Academic Press.

McMullan, A., A. Bowman, and E. Scott (2007). Water quality in the River Clyde: A case study of additive and interaction models. *Environmetrics 18*(5), 527–539.

Melnikov, O., L. H. Raun, and K. B. Ensor (2016). Dynamic principal component analysis: Identifying the relationship between multiple air pollutants. *arXiv preprint arXiv:1608.03022*.

Miller, C. and A. Bowman (2012). Smooth principal components for investigating changes in covariance over time. *Journal of the Royal Statistical Society: Series C 61*(5), 693–714.

Milligan, G. and M. Cooper (1985). An examination of the procedures for determining the number of clusters in a data set. *Psychometrika 50*(2), 159–179.

Milligan, G. and M. Cooper (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research 21*(4), 441–458.

Moraetis, D., D. Efstathiou, F. Stamati, O. Tzoraki, N. Nikolaidis, J. Schnoor, and K. Vozinakis (2010). High-frequency monitoring for the identification of hydrological and bio-geochemical processes in a Mediterranean river basin. *Journal of Hydrology 389*(1), 127–136.

Nason, G. (2008). *Wavelets Methods in Statistics With R* (1st ed.). Use R! Springer.

Neal, C. (1988). Determination of dissolved CO2 in upland streamwater. *Journal of Hydrology 99*(1-2), 127–142.

Neal, C., B. Reynolds, J. Kirchner, P. Rowland, D. Norris, D. Sleep, A. Lawlor, C. Woods, S. Thacker, H. Guyatt, C. Vincent, K. Lehto, S. Grant, J. Williams, M. Neal, H. Wickham, S. Harman, and L. Armstrong (2013). High-frequency precipitation and stream water quality time series from Plynlimon, Wales: An openly accessible data resource spanning the periodic table. *Hydrological Processes 27*(17), 2531–2539.

Neal, C., B. Reynolds, D. Rowland, P.and Norris, J. Kirchner, M. Neal, D. Sleep, A. Lawlor, C. Woods, S. Thacker, H. Guyatt, C. Vincent, K. Hockenhull, H. Wickham, S. Harman, and L. Armstrong (2012). High-frequency water quality time series in precipitation and streamflow: From fragmentary signals to scientific challenge. *Science of the Total Environment 434*, 3–12.

Ockendena, M., C. Deasya, C. Benskina, K. Bevena, S. Burked, A. Collinse, R. Evansf, P. Falloong, K. Forbera, K. Hiscockh, M. Hollawaya, R. Kahanag, C. Macleodi, S. Reaneyb, M. Snella, M. Villamizarj, C. Wearinga, P. Withersk, J. Zhouj, and P. Haygarth (2016). Changing climate and nutrient transfers: Evidence from high temporal resolution concentration-flow dynamics in headwater catchments. *Science of The Total Environment 548-549*, 325–339.

Ombao, H. and M. Ho (2005). Time-dependent frequency domain principal components analysis of multichannel non-stationary signals. *Computational Statistics and Data Analysis 50*(9), 2339–2360.

Panaretos, V. and S. Tavakoli (2013). Cramer-Karhunen-Loeve representation and harmonic principal component analysis of functional time series. *Stochastic Processes Applications 123*(7), 2779–2807.

Paparoditis, E. (2009). Testing temporal constancy of the spectral structure of a time series. *Bernoulli 15*(4), 1190–1221.

Pearce, J., J. Beringer, N. Nicholls, R. Hyndman, and N. Tapper (2011). Quantifying the influence of local meteorology on air quality using generalized additive models. *Atmospheric Environment 45*(6), 1328–1336.

Peng, J. and H. Muller (2008). Distance-based clustering of sparsely observed stochastic processes with applications to online auctions. *Annals of Applied Statistics 2*(3), 1056–2077.

Percival, D. and A. Walden (2006). *Wavelets Methods for Time Series Analysis* (1st Edition ed.). Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Phineiro, J. and D. Bates (2000). *Mixed-Effects Models in S and S-Plus* (1st ed.). Statistics and Computing. Springer.

Politis, D. and J. Romano (1994). The stationary bootstrap. *Journal of the American Statistical Association 89*(428), 1303–1313.

Priestley, M. and T. Subba Rao (1969). A test for non-stationarity of time-series. *Royal Statistical Society 31*(1), 140–149.

Ramsay, J. and S. Graves (2009). *Functional Data Analysis with R and MATLAB (Use R).* Springer.

Ramsay, J., G. Hooker, and S. Graves (2009). *Functional Data Analysis with R and MATLAB.* Springer Series in Statistics. Springer.

Ramsay, J. and B. Silverman (1997). *Functional Data Analysis* (1st ed.). Springer Series in Statistics. Springer.

Ramsay, J. and B. Silverman (2005). *Functional Data Analysis* (2nd ed.). Springer Series in Statistics. Springer.

Ramsay, J. O. and C. J. Dalzell (1991). Some tools for functional data analysis. *Journal of Royal Statistical Society: Series B 53*(3), 539–572.

Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association 66*(336), 846–850.

Raymond, P., N. Caraco, and J. Cole (1997). Carbon dioxide concentration and atmospheric flux in the Hudson River. *Estuaries 20*(2), 381–390.

Richey, J., J. Melack, A. Aufdenkampe, V. Ballester, and L. Hess (2002). Outgassing from Amazonian rivers and wetlands as a large tropical source of atmospheric CO2. *Nature 416*(6881), 617–620.

Ruppert, D., M. Wand, and R. Carroll (2003). *Semiparametric Regression* (1st ed.). Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Schwartz, J. (1994). Nonparametric smoothing in the analysis of air pollution and respiratory illness. *The Canadian Journal of Statistics 22*(4), 471–487.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464.

Sen, A. (2009). Spectral temporal characterization of river flow variability in England and Wales for the period 18652002. *Hydrological Processes 23*(8), 1147–1157.

Shen, Q. and J. Faraway (2004). An F test for linear models with functional responses. *Statistica Sinica 14*(4), 12391257.

Shumway, R. and D. Stoffer (2011). *Time Series Anaylsis and its Applications: With R Examples* (3rd ed.). Springer.

Spezia, L., M. Futter, and M. Brewer (2011). Periodic multivariate normal hidden Markov models for the analysis of water quality time series. *Environmetrics 22*(3), 304–317.

Sugiura, N. (1978). Further analysis of the data by Akaike' s information criterion and the finite corrections. *Communications in Statistics - Theory and Methods 7*(1), 13–26.

Sun, Y. and M. Genton (2011a). Adjusted functional boxplots for spatio-temporal data visualization and outlier detection. *Environmentrics 23*(1), 54–64.

Sun, Y. and M. Genton (2011b). Functional boxplots. *Journal of Computational and Graphical Statistics 20*(2), 316–334.

Tibshirani, R., G. Walther, and T. Hastie (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of Royal Statistical Society: Series B 63*(2), 411–423.

Trapey, T. and K. Kinateder (2003). Clustering functional data. *The Journal of Classification 20*(1), 93–114.

Underwood, F. (2009). Describing long-term trends in precipitation using generalized additive models. *Journal of Hydrology 364*(3), 285–297.

United States Environmental Protection Agency EPA (2012). Water: Monitoring and assessment.

Waldron, S., M. Scott, and C. Soulsby (2007). Stable isotope analysis reveals lower-order river dissolved inorganic carbon pools are highly dynamic. *Enviornmental Science Technology 41*(17), 6156–6162.

Waldron, S., M. Scott, L. Vihermaa, and J. Newton (2014). Quantifying precision and accuracy of measurements of dissolved inorganic carbon stable isotopic composition using continuous-flow isotope-ratio mass spectrometry. *Rapid Communications in Mass Spectrometry 28*(10), 1117–1126.

Ward, J. and H. Joe (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association 58*(301), 236–244.

Welch, P. (1967). The use of Fast Fourier Transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics 15*(2), 70–73.

White, M., J. Schmidt, and D. Topping (2005). Application of wavelet analysis for monitoring the hydrologic effects of dam operation: Glen Canyon Dam and the Colorado River at Lees Ferry. *River Research and Applications 21*(5), 551–565.

Whiting, J. (2006). *Identification and modelling of hydrological persistence with hidden Markov models*. Ph. D. thesis, University of Adelaide, Australia.

Wood, S., Y. Goude, and S. Shaw (2015). Generalized additive models for large data sets. *Journal of Royal Statistical Society: Series C 64*(1), 139–155.

Wood, S. N. (2006). *Generalized Additive Models - An Introduction with R* (1st ed.). Text in Statistical Science Series. Chapman and Hall.

Wood, S. N. (2011). Fast stable REML and ML estimation of semiparametric GLMs. *Journal of Royal Statistical Society: Series B 73*(1), 3–36.

Yao, G., Q. Gao, Z. Wang, X. Huang, T. He, Y. Zhang, S. Jiao, and J. Ding (2007). Dynamics of CO2 partial pressure and CO2 outgassing in the lower reaches of the Xijiang River, a subtropical monsoon river in China. *Science of the Total Environment 376*(1), 255–266.

Yick, J., B. Mukherjee, and D. Ghosal (2008). Wireless sensor network survey. *Computer Networks 52*(12), 2292–2230.

Zivot, E. and J. Wang (2006). *Modeling Financial Time Series with S-plus* (2nd ed.)., Chapter Vector Autoregressive Models for Multivariate Time Series, pp. 385–429. Springer New York.