# Statistical Analysis of Freshwater Parameters Monitored at Different Temporal Resolutions

Firdaus Mohamad Hamzah

*A Dissertation Submitted to the*
*University of Glasgow*
*for the degree of*
*Doctor of Philosophy*

School of Mathematics and Statistics

April 2012

# Declaration

This thesis has been composed by myself and it has not been submitted in any previous application for a degree. The work reported within was executed by myself, unless otherwise stated.

# Acknowledgements

I would like to thank the Ministry of Higher Education of Malaysia and the National University of Malaysia for supplying the funding for my research work. I also gratefully acknowledge the Scottish Environment Protection Agency (SEPA), the Scottish Centre of Ecology and the Natural Environment (SCENE) and Professor Susan Waldron from the University of Glasgow for providing the data.

Special thanks definitely go to all my supervisors; Professor Marian Scott, Dr. Claire Miller, Professor Susan Waldron and Professor Colin Adams for all the guidance, help, motivation, commitment and enthusiasm, their contributions have always given me the inspiration and encouragement.

Thanks to all in the School of Mathematics and Statistics who have helped me along the way to produce this thesis. My sweet memory at Glasgow will not be depleted.

To my family; my wife, Noredawaty and all my children, Muhammad Faiq Saqif, Nuralya Saqifah and Nuralya Sakinah, and also my parents and brother, this is a special gift for you.

# Abstract

Nowadays, it is of great importance in ecological and environmental studies to investigate some prominent features in environmental determinants using appropriate statistical approaches. The initial motivation of this work was provided by the enthusiasm of the limnologist, biologist and statistician, interested in exploring and investigating certain features of time series data at different temporal resolutions to environmental parameters in freshwater.

This thesis introduces a variety of statistical techniques which are used to provide sufficient information on the features of interest in the environmental variables in freshwater.

Chapter 1 gives the background of the work, explores the details of the locations of the case studies, presents several statistical and ecological issues and outlines the aims and objectives of the thesis.

Chapter 2 provides a review of some commonly used statistical modelling approaches to model trend and seasonality. All the modelling approaches are then applied to low temporal resolution (monthly data) of temperature and chlorophyll measurements from 1987-2005 for the north and south basins of Loch Lomond, Scotland. An investigation into the influence of temperature and nutrients on the variability of log chlorophyll is also carried out.

Chapter 3 extends the modelling for temperature in Chapter 2 with the use of a mixed-effects model with different error structures for temperature data at a moderate temporal resolution (1 and 3 hourly data) in the north, mid and south basins. Three approaches are proposed to estimate the positions of a sharp change in gradient of temperature (thermocline) in deeper basins, using the maximum relative rate of change, changepoint regression and derivatives of a smooth curve.

Chapter 4 investigates several features in semi-continuous environmental variables (15 and 30 minutes data). The temporal pattern of temperature, pH, conductivity and barometric pressure, and the evidence of similarity of the signals of pH and conductivity is determined, using wavelets. The time taken for pH and conductivity to return to 'baseline levels' (recovery period) following extreme discharge is determined for different thresholds of 'extreme discharge' for the Rivers Charr and Drumtee Burn, Scotland and models for the recovery period are proposed and fitted. Model validation is carried out for the River Charr and the occurrence of clusters of extreme discharge for both rivers is investigated using the extremal index.

Chapter 5 summarises the main findings within this thesis and several potential areas for future work are suggested.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Background

In a wide range of environmental fields, the process of explaining the features of the environmental parameters of interest often lead to a detailed statistical analysis. Different temporal resolutions of environmental data sets may reveal certain features that require thorough understanding through the use of the various statistical techniques. Time series measurements of environmental and ecological variables with different temporal resolutions in freshwater offer a vast opportunity for exploring, investigating and developing appropriate statistical models, which extract valuable information at all temporal scales.

## 1.1 Freshwater

Freshwater is important in terms of ecological systems. Freshwaters are predominantly low salt concentration water such as lakes and rivers and therefore, they are often used as the main natural source of water. They are naturally characterized by their biological, chemical and physical determinants that vary temporally and spatially.

However, freshwaters are among the most currently threatened ecosystems in the world (Abell, 2002). The introduction of alien organisms, dam construction, habitat modification and alteration of water chemistry are among the most harmful anthropogenic impacts to the ecosystems (Malmqvist and Rundle, 2002).

### 1.1.1 Lakes and Rivers

Lakes are important for flood control, water supply, cultivation, irrigation, navigation and tourism (Shuijing et al., 2010). Lake ecosystems are rarely influenced by only a single pressure. Eutrophication (a growing deterioration of water quality resulting from the rise of algae concentration from the increase of nutrient loading) (Wetzel, 2001) is commonly observed in lakes. Such a phenomenon is part of the huge ecological problem that has mostly damaged ecosystem condition and resulted in imbalances among different biological processes and a decrease in ecosystem biodiversity (Shuijing et al., 2010). The symptoms of ecosystem damage (Hu et al., 2008) compromise phytoplankton blooms, fish kills and decline in useful water. Chlorophyll$_a$, a major component in phytoplankton, is a primary photosynthetic pigment of all oxygen-evolving photosynthetic organisms and is the main indicator for water quality (Wetzel, 2001). Apart from the deterioration of the water quality caused by phytoplankton abundance, climate change could be another pressure that may give rise to significant impact on water temperature and as a consequence, the ecological system is highly likely to be affected (Ferguson, 2007). Hence, the condition of the ecosystem of the lakes is assessed using several statistical approaches.

While lakes are important due to several factors as mentioned above, rivers are central to the process of economic and social development (John and Michael, 2011). Rivers contribute about 0.0001% of the water on earth and are central to surface water ecosystems (Wetzel, 2001). Despite these low quantities, running waters are hugely significant to human life by providing a rich source of fish and other aquatic life, and are a major source of clean water for drinking and irrigation. Small streams and rivers often join together and subsequently form large river network systems. The water chemistry in rivers is complex and it is influenced by input from the atmosphere and human activities, and geological factors such as type of rock in the river basin. The ecological condition for the animals and plants is influenced by the chemistry of rivers and so, it is of importance to understand the state of the environment of rivers. The condition of the river environment is therefore critical to its management and this thesis contributes to our understanding by examining a variety of statistical techniques to present several features of the environmental determinands in rivers.

## 1.2 Case Studies

The importance of environmental and ecological determinants in lakes and rivers is in characterizing the status of the freshwaters and so, some of their features have been studied. Three case studies from Loch Lomond, and the River Charr and River Drumtee are used to explore and investigate some of the important features that may contribute to the condition of the water body.

### 1.2.1 Loch Lomond

The following descriptions of Loch Lomond are taken from (Krokowski, 2007).

Loch Lomond is located in the north-west of Glasgow and close to the Trossachs National Park (First Scottish Park). Both Loch Lomond and the National Park encompass an area of 1,865km$^2$. The areas of Loch Lomond are characterized as Ramsar sites in terms of rare plants, aquatic invertebrates and wetland plant, and Greenland goose under International and European designations. Loch Lomond is also a Special Area of Conservation for its wood and otters. The loch is occupied by 19 species of freshwater fish such as powan, brook lampreys and Atlantic Salmon (Loch Lomond and Trossachs National Park Authority, 2005).

Loch Lomond is an icy highland loch, which is the largest in the United Kingdom, covering an area of 71.1km$^2$. In addition, the depth of the north basin is 189.9m which is the third deepest in Great Britain (Loch Lomond and Trossachs National Park Authority, 2005). The volume, mean depth and catchment area of the loch are 2628×106m$^3$, 37m and 781km$^2$, respectively. Research has been extensively carried out at the loch in terms of its ecological and geological features (see for example, Slack (1957), Maitland (1981), and Murphy et al. (1994)).

The loch is generally divided into three basins which are distinguished by the differences in their geological features and types of catchment of the main river. The highland boundary divides the loch into northern and southern basins, lying across the loch from the north east to south west. The north basin is generally highland with hard rock and deep, and is surrounded by woodland. The hard rock in the north provides less nutrients with acidic water, and receive less exposure to solar radiation. The main inflow river in the north comes from the river Falloch, flowing through a mountainous catchment. The south basin, which is much shallower than the north basin, is occupied by several islands. The inflow to the loch comes from the river Endrick, flowing through agricultural land before entering the loch. The outflow from the loch comes from the river Leven, which is preserved for water supply. The mid basin of the loch is occupied by a large number of islands (Habib et al., 1997).

### 1.2.2 River Charr

The following descriptions of the river Charr are taken from (Waldron et al., 2009).

The river Charr flows through various habitats, is 10m wide and is located in Glen Dye ($56^o56'27$N, $2^o35'00$W), a headwater subcatchment of the River Dee in north east of Scotland. The altitude of the catchment is in range of 100m up to 580m.

Waldron et al. (2007) describe the topography and sampling points, soil coverage, geology and land use of the Glen Dye catchment. The area of the catchments is dominated by granite. The catchment features include elevated areas above 450m which are covered by carbonized vegetable (up to 5m deep) and leached soil (less than 1m deep). (Waldron et al., 2009) stated ‘The main river valley generally has freely draining alluvial deposits and soils. Regular burning of small areas of moorland may have contributed to some peat degradation and hagging (Thompson et al., 2001) and in places erosion extends to the organomineral interface. A high density of ephemeral drainage channels cover the peat, connecting it to the perennial stream channel network'.

### 1.2.3 River Drumtee

The following descriptions of the river Drumtee are taken from (Waldron et al., 2009).

The river Drumtee Burn is one of the main river networks in Whitelee in the north west region of the biggest wind farm in Europe and encompasses 9.4 $km^2$.

The following description of the Whitelee wind farm is provided by the Environmental Impact Statement (EIS) and has been prepared by Scottish Power for planning consent for the windfarm (CRE Energy, 2002). The wind farm is located in central Scotland and consists of 140 turbines covering an area of 176 $km^2$ across grids of $55^o40^{'}24$N and $4^o16^{'}00$W. (Waldron et al., 2009) stated that 'Land use is predominantly forestry, with rough grazing on open moorland, and more improved pasture and arable land on the northern lower slopes. The windfarm is mostly located in areas of peat, underlain by a clay seal and weakly permeable igneous or moderately permeable sedimentary rocks. Peat depth, measured at 161 locations, ranges from five to over 500cm, mean depth of 190cm (±1 S.D.134.7cm)'.

(Waldron et al., 2009) also stated that ‘All of the peat lands in the development area are blanket bog, but in some locations have features associated with intermediate bogs. There are several large sphagnum-dominated pools and lawns with the peat exceeding 4.5m deep. This contrasts with the surrounding drier, heather-dominated, less species-rich Calluna vulgaris-Eriophorum vaginatum vegetation. Only 35ha (3.5%) of the unforested blanket bog is primary natural bog whilst the remaining area has been impacted, mostly due to the Whitelee forest, a first rotation plantation of 5917ha of mainly Sitka spruce, established between 1962 and 1992 and at altitude from 220m to 376m. Most of the bog exhibits varying degrees of surface damage and drying in the forest area. Bog vegetation has generally been highly modified or lost completely under canopy closure. Acid grassland habitat dominates outwith the forest and unmodified peatland’.

## 1.3 Data

The environmental data used in this thesis comprise three different time resolutions; low, moderate and high frequencies measurements.

Low frequency data consists of monthy temperature, chlorophyll, phosphate and nitrate measurements from 1987-2005 in the north and south basins of Loch Lomond. The data were supplied by the Scottish Environment Protection Agency (SEPA).

The moderate frequency data consists of 3 hourly temperature measurements recorded from thermistor chains at 11 different depths at Cailness (north), Creinch (south), lower and upper chain of Ross Points in Loch Lomond from 1 September 2002 - 31 August 2003. Additionally, 1 hourly temperature measurements collected from a thermistor chain at 11 different depths at the mid basin of Loch Lomond are also available from 17 April 2008 - 27 May 2009. The 1 hourly and 3 hourly data sets were supplied by the Scottish Centre for Ecology and the Natural Environment (SCENE) and Professor Susan Waldron from the University of Glasgow, respectively.

The high frequency data sets collected from monitoring buoys, measured every 15 and 30 minutes over the time period, were also supplied by Professor Waldron. The data sets are temperature, barometric pressure, pH, conductivity and discharge measurements from the rivers Charr and Drumtee Burn in Aberdeen and Whitelee, respectively. The data sets from Aberdeen are measured every 15 minutes from October 2004 - September 2007 whilst in Whitelee, they are recorded every 30 minutes from October 2007 - September 2010.

## 1.4 Ecological Issues

It is of interest to investigate the changes of temperature and chlorophyll at the surface of Loch Lomond since each of them is naturally related to the changes in weather and water quality, respectively. The trend in temperature provides a good indicator of climate change over a particular period and may reflect changes in the ecological processes in the lake. An increase in temperature may result in a phytoplankton bloom. The abundance of phytoplankton resulting from the influence of nutrients may lead to deterioration in water quality.

Temperature monitored at different depths of the loch may provide further information about the ecological process by understanding the temperature changes over depth. However, is depth really required to precisely explain the variability of temperature over the time period in the loch? This question could be answered by fitting an appropriate statistical model to temperature over the time period with depth incorporated as a random effect. Deeper water of the loch may highlight different characteristics of temperature over the time period, compared to shallower water and so, different ecological processes may occur between the north and south basins. A natural feature that often appears in the water column, is where there is a large change in the temperature gradient (a thermocline). This can result in different ecological processes at different depths. Its position is of interest since for a given time point it may divide a particular layer in the water column into different biological and chemical features.

Changes of temperature, barometric pressure, pH and conductivity in rivers over the year may provide certain conditions of dynamic behaviour in large streams. The occurrence of a particular temporal pattern in a single environmental variable over a short time period is important as this may reflect an immediate response in the ecological process. Two environmental variables with a similar temporal pattern may reveal some kind of relationship in winter and summer over different years and so, the coherency of the signals have to be identified. Additionally, two environmental variables with a significant coherence and naturally influenced by hydrological events may provide a particular change in response to the occurrence of such events. Therefore, such a change is identified and an important feature related to the response of both variables is determined. Further investigation on such a prominent feature is carried out to allow prediction in later years.

## 1.5 Statistical Issues

Time series models are often useful to explore and investigate the variability of particular environmental and ecological variables. However, their application is often constrained by the modelling assumptions. In particular, the assumptions of constant mean and variance over the time period often do not hold with real data which manifest nonstationary behaviour. However, the nonstationary nature of the data can be of specific interest to investigate natural features in the time series.

The occurrence of a seasonal pattern in low temperature measurements in freshwater over the time period can be evidence of nonstationary data. The occurrence of a diurnal pattern in moderate and high frequency temperature measurements provides additional information about the nature of the data. Hence, the appearance of the daily and annual cycle need to be removed prior to modelling to avoid any violation of standard model assumptions. This would normally be modelled in the mean/deterministic part of a model to deal with seasonality. The acf and pacf can highlight a possible seasonal pattern but this would be removed before assessing the residuals for remaining correlation.

Missing data often occur in real data monitoring due to several causes. A large number of missing data in a particular period should be given much attention since it may greatly affect the parameters in the model. Missing data (i.e. not occurring at random) often appear in cases where no data are recorded over a sufficiently long period due to failure of the monitoring device. The above issue may greatly affect the fitted model by inflating the bias. For instance, a large number of missing data at the beginning or end of the period may lead to biased estimates of trend in a model of temperature and as a result, incorrect judgement is likely be made on the mean changes of temperature.

Modelling trend and seasonality of low frequency temperature and chlorophyll measurements provided by SEPA over the time period requires flexible models. In particular, models which allow the relationships to be investigated as parametric, nonparametric or varying with respect to another covariate.

Modelling moderate frequency temperature measurements from thermistor chains over the time period at different depths requires an appropriate technique in such a way that the random effect can be incorporated to provide more information. A mixed-effects model could be appropriate for the above case if there is a clear difference of the variability of temperature over the time period for each depth. Higher variance of the random effects in the model may also provide additional information on the potential use of the mixed-effects model. Conversely, a fixed effect model could be more appropriate if similar variability of temperature pattern is exhibited for each depth. Moreover, modelling the moderate frequency temperature measurements with depth at a given time point provides statistical challenges in modelling complex shapes of a curve with a small amount of data, identifying appropriate correlation structure between depths and identifying changepoint and inflection point that represent the position of the thermocline.

The introduction of semi-continuous temperature, barometric pressure, pH and conductivity measurements from monitoring buoys requires a reliable statistical technique to show evidence of the variability of each determinant and the relationship between the determinands. The statistical techniques previously used for the low and moderate frequency data may not be appropriate here since a strong correlation between adjacent data over the time period is expected, leading to nonstationary data. Hence, the technique of wavelets which is not constrained by a stationarity assumption could be appropriate for this case. Furthermore, an appropriate regression model is required to estimate and predict the recovery period of pH and conductivity following an extreme discharge corresponding to particular thresholds. Extreme discharges can be clustered and hence it is necessary to explore this since it can affect the modelling

Modelling low frequency environmental data up until now has been very common and this was what agencies such as SEPA routinely collect. Moreover, with the introduction of moderate frequency data from thermistor chain and semi-continuous data from monitoring buoys, the use of other statistical approaches is required. While current approaches are appropriate for data from thermistor chains and monitoring buoys, the availability of more advanced technology such as remote sensing requires more complex statistical techniques to model the variability in nonstationary data explicitly.

## 1.6 Aims and Objectives

The aims of this thesis are to explore and investigate some important features of environmental determinants in lake and rivers via appropriate statistical approaches. Loch Lomond, the rivers Charr and Drumtee Burn are used as case studies.

Motivated by the importance of prominent features in environmental parameters in freshwaters, the main objectives of this thesis are as follows:

- To investigate trend and seasonal patterns of environmental variables.
- To investigate the stratification of temperature with depth.
- To investigate temporal pattern within and between environmental variable, and recovery period of environmental variable following extreme event in hydrological determinand.
- To develop and apply statistical methodology appropriate to the temporal frequency of environmental monitoring.

The R package (Venables et al., 2011) is used throughout this thesis for analyzing the data sets.

# Chapter 2

# Trends, Seasonality and Relationships

## 2.1 Introduction

Surface water temperature is one of the most important parameters for determining the ecological conditions in lakes (Horne and Glodman, 1994) and can be an indicator of the regional weather and climate near large lakes (Austin and Colman, 2007). The ecological conditions in lakes could be affected resulting from the influence of temperature on the nutrients in lakes (Spears et al., 2008). Several studies have recently been undertaken to determine changes of temperature in lakes. A previous investigation of annual average surface water temperature at Loch Lomond shows an increase of $5^oC$ from 1987 - 2005 (Krokowski, 2007). Such a change is relatively high compared to freshwaters in other countries. For instance, the mean annual temperature of the surface water in Lake Geneva increased by only $1^oC$ from 1983 to 2000 (Gillet and Quetin, 2006).

The use of chlorophyll$_a$ concentrations to represent phytoplankton loading as a general measure of lake water quality is widely adopted around the world as the phytoplankton community will grow in response to available excess nutrients (Nitrogen and Phosphate) a phenomenon known as eutrophication (Smith et al., 1999), resulting in a deterioration in water quality. The use of chlorophyll$_a$ concentrations as a part of the ecological quality determinant in lakes is essential and through the Freshwater Framework Directive most of the European countries have adapted this indicator as a good general measure of ecological impact of eutrophication. There has been very little scientific research examining how such an element in phytoplankton varies naturally, in the absence of nutrient pressures (Carvalho et al., 2008).

The importance of temperature and chlorophyll$_a$ concentrations in characterizing the ecological characteristics in lakes has stimulated interest in investigating their changes over the year in Loch Lomond and hence, two objectives are outlined for this work. Firstly, the trends and seasonal patterns of temperature and chlorophyll$_a$ are investigated and secondly, the evidence of effects of trend, seasonality, temperature and nutrients on the changes in chlorophyll$_a$ concentrations is determined.

## 2.2 Data

Monthly temperature and chlorophyll$_a$ measurements from 1987 to 2005 from the north and south basins of Loch Lomond were supplied by the Scottish Environment Protection Agency (SEPA). The time series for both variables are incomplete in the sense that there are missing values in a number of years.

Figure 2.1 shows plots of temperature (top) and chlorophyll$_a$ (bottom) from 1987-2005 in the north (left) and south (right) basins. The points represent the measurements for each variable over the time period. The variability in the temperature appears mainly constant over the year in both basins. Conversely, the variability in chlorophyll$_a$ is not constant across the 19 years and so, a log transform is used to stabilize the variance.

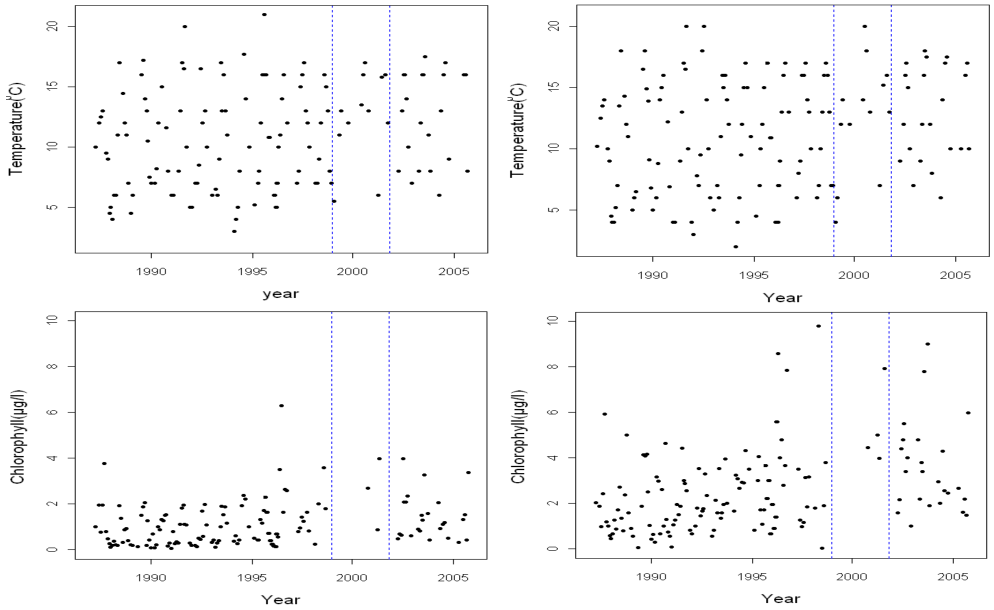

**Figure 2.1.** Plots of temperature (top) and chlorophyll$_a$ (bottom) in the north (left) and south (right) basins from 1987-2005.

Figure 2.2 shows plots of log chlorophyll$_a$ from 1987-2005 in the north (left) and south (right) basins, highlighting more stable variability in the transformed measurements compared to the actual measurements. A large number of missing values are apparent in log chlorophyll$_a$, more so than in temperature, particularly between 1999 and 2001 (between the two vertical dashed lines in each plot).

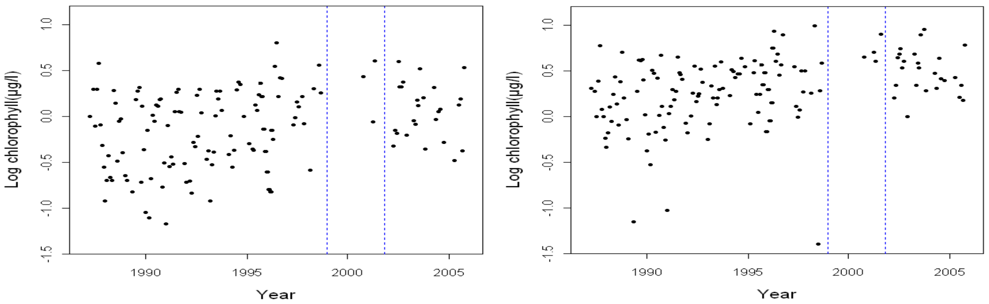

**Figure 2.2.** Plots of log chlorophyll$_a$ in the north (left) and south (right) basins from 1987-2005.

Table 2.1 shows the percentage of missing temperature and chlorophyll$_a$ values for each month across the 19 years. A high number of missing temperature and chlorophyll$_a$ values are observed from December - February and therefore, the missing values are not occurring at random.

| Month | Temperature (%) | | Chlorophyll$_a$ (%) | |
|---|---|---|---|---|
| | North | South | North | South |
| Jan | 68.4 | 68.4 | 68.4 | 68.4 |
| Feb | 63.2 | 63.2 | 68.4 | 73.7 |
| Mar | 63.2 | 47.4 | 63.2 | 57.9 |
| Apr | 26.3 | 21.1 | 47.4 | 36.8 |
| May | 31.6 | 31.6 | 36.8 | 42.1 |
| June | 26.3 | 15.8 | 52.6 | 36.8 |
| Jul | 26.3 | 15.8 | 31.6 | 36.8 |
| Aug | 15.8 | 21.1 | 26.3 | 21.1 |
| Sept | 42.1 | 42.1 | 63.2 | 63.2 |
| Oct | 15.8 | 15.8 | 47.4 | 42.1 |
| Nov | 52.6 | 52.6 | 57.9 | 68.4 |
| Dec | 73.7 | 73.7 | 79.0 | 78.9 |

**Table 2.1.** Percentage of missing data for temperature and chlorophyll$_a$, by month.

Table 2.2 displays the percentage of missing temperature and chlorophyll$_a$ values for each year. A high number of missing temperature and chlorophyll$_a$ values are observed from 1999-2005. The temperature and chlorophyll measurements are mainly missing in winter from 1999-2005. A complete case analysis may result in misleading conclusions and so, statistical methods to impute data will be explored for temperature and chlorophyll.

The missing values in temperature from 1987-2005 are imputed using several approaches described below. Since there is a large period of missing data in chlorophyll from 1999-2001, the unobserved data in this period are not imputed. The missing values in log chlorophyll$_a$ between 1987-1998 and 2002-2005 are imputed using similar approaches to those for temperature.

| Year | Temperature (%) | | Chlorophyll$_a$ (%) | |
|---|---|---|---|---|
| | North | South | North | South |
| 1987 | 41.7 | 41.7 | 33.3 | 33.3 |
| 1988 | 16.6 | 25.0 | 8.3 | 8.3 |
| 1989 | 33.3 | 25.0 | 25.0 | 25.0 |
| 1990 | 41.7 | 33.3 | 0 | 0 |
| 1991 | 25.0 | 25.0 | 8.3 | 8.3 |
| 1992 | 33.3 | 25.0 | 25.0 | 25.0 |
| 1993 | 25.0 | 25.0 | 16.7 | 16.7 |
| 1994 | 41.7 | 25.0 | 33.3 | 33.3 |
| 1995 | 33.3 | 41.7 | 16.7 | 16.7 |
| 1996 | 33.3 | 33.3 | 16.7 | 16.7 |
| 1997 | 33.3 | 25.0 | 50.0 | 41.7 |
| 1998 | 25.0 | 16.7 | 66.7 | 58.3 |
| 1999 | 66.7 | 58.3 | 100.0 | 100.0 |
| 2000 | 66.7 | 66.7 | 91.7 | 91.7 |
| 2001 | 66.7 | 66.7 | 83.3 | 75.0 |
| 2002 | 41.7 | 41.7 | 33.3 | 33.3 |
| 2003 | 41.7 | 41.7 | 41.7 | 41.7 |
| 2004 | 58.3 | 58.3 | 58.3 | 58.3 |
| 2005 | 75.0 | 66.7 | 58.3 | 58.3 |

**Table 2.2.** Percentage of missing data for temperature and chlorophyll$_a$, by year.

### 2.2.1 Data Imputation Procedures

**Temperature**

The problem of missing data has been recognized and increasingly debated in the statistical literature (for example, see Little and Rubin (2002); Allison (2002)). Missing data may affect statistical power by reducing sample size or, in more serious case, estimates of statistics derived by deleting cases with missing values may be biased, particularly if the missing values are different from those with complete data (Diane et al., 2010).

Hence, four potential imputation approaches are used to impute data for temperature. The mean square errors (MSE), defined as $\sum_{i=1}^{n} \frac{(y_i - \hat{y_i})^2}{n}$, is used for comparing the observed measurements $(y_i)$ and the corresponding fitted values $(\hat{y_i})$, and the approach that contributes to the lowest mean square errors is chosen.

The first approach is mean substitution in which the mean of the temperature for a given month is determined and used to replace all the missing values in that month.

In the second approach, a harmonic model which includes a trend and seasonal term (equation 3.3) is used,

$$y_i = \beta_0 + \beta_1 t_i + \gamma \cos \left\{ \frac{2\pi (t_i - \theta)}{p} \right\} + \epsilon_i; i = 1, 2, \ldots, n \tag{2.1}$$

where $y_i$ is the $i$th temperature, $t_i$ is the $i$th month, $\beta_0$ and $\beta_1$ are the intercept and slope, $\epsilon_i$ are random errors which are identically and normally distributed with zero mean and constant variance, $p$ is the number of months in a year, $\gamma$ is amplitude and $\theta$ is the phase angle. $\gamma$ and $\theta$ are defined as $\sqrt{\beta_1^2 + \beta_2^2}$ and $pcos^{-1}\left(\frac{\beta_2^2}{\beta_1^2}\right)$, respectively.

For computational simplicity model 3.3 can be written to be linear as defined in equation 3.5.

$$y_i = \beta_0 + \beta_1 t_i + \beta_2 cos\left(\frac{2\pi t_i}{p}\right) + \beta_3 sin\left(\frac{2\pi t_i}{p}\right) + \epsilon_i \qquad (2.2)$$

where $\beta_0$, $\beta_1$, $\beta_2$ and $\beta_3$ are the coefficients of the model.

The steps for imputing the missing values are as follow:

1. The harmonic model from equation 3.5 is fitted to the monthly temperature measurements from 1987-2005.
2. The assumptions of linearity and constant variance of the erros are checked.
3. The estimates from the fitted model are used to impute the missing data from 1987 to 2005.
4. The imputed value is added to an estimated error, generated from a Normal distribution with zero mean and variance based on the residuals from the fitted model in step (1).

In the third approach, a model which decomposes the time series into trend, seasonal, cyclical, and irregular (error) terms (equation 2.3) is used,

$$y_t = TR_t + SN_t + CL_t + IR_t; t = 1, 2, \dots, n \tag{2.3}$$

where $TR_t, SN_t, CL_t, IR_t$ are trend, seasonal, cyclical and irregular components at time $t$ and $n$ is the number of measurements (Bowerman and O'Connell, 1993). The steps of imputation are as follow:

- All the missing values in the first year (1987) are imputed by substituting the mean value for the specific missing month's data since a complete data set is required in the first year prior to decomposing the time series.
- The estimates of $TR_t + CL_t$ are determined using the centered moving average $CMA_t$ from 1987-2005.
- The estimates of $SN_t+IR_t$, defined by the difference between measurements and $CMA_t$ at each time $t$, are computed and they are grouped by month. The mean of these estimates is determined for each month.
- The seasonal factors $SN_t$ are obtained by normalizing the $\overline{SN}_t$ values and the estimate of $SN_t$ is $SN_t = \overline{SN}_t - (\frac{\sum_{i=1}^{L} \overline{SN}_t}{12})$.
- The deseasonalized measurements at time $t$ are determined by the difference between the measurements and $SN_t$.
- The estimate of trend at time $t$, $TR_t$ is determined by fitting a linear regression model to the above deseasonalized measurements.
- The estimate of $CL_t + IR_t$ is $y_t - (TR_t + SN_t)$ and the moving average of order $p$, $CL_t = \frac{(CL_{t-1}+IR_{t-1})+(CL_t+IR_t)+(CL_{t+1}+IR_{t+1})}{p}$ is used for estimation.

- $IR_t$ are assumed to be the errors and they are generated from a Normal distribution with zero mean and variance equal to the mean square of deviation of the residuals and added to the imputed values.

In the final approach, a linear regression model of water temperature in the loch is fitted to air temperature from Paisley (which hosts the nearest MET office station) as defined in equation 3.1 and the fitted values are used to impute the missing values,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i; i = 1, 2, \ldots, n \tag{2.4}$$

where $y_i$ is the $i$th water temperature in Loch Lomond, $x_i$ is the $i$th air temperature at Paisley and $\epsilon_i$ are random errors which are assumed to be identically and normally distributed with zero mean and constant variance. The use of air temperature from a nearby location is due to the fact that temperature between two local areas is highly correlated. The minimum and maximum monthly air temperatures at Paisley are available from 1987-2005 and so, the mean of these two values is computed for each of the months over the year and its correlation with water temperature in the north and south basins are initially checked prior to model fitting. Evidence of a strong linear relationship highlighted by the correlation indicates the appropriateness of the use of a linear regression model of water temperature on air temperature to impute the missing data.

**Chlorophyll$_a$**

The first three approaches used for imputing missing values in temperature are applied to log chlorophyll$_a$, however, the fourth approach used for temperature is not considered due to the fact that the air temperature from Paisley and log chlorophyll$_a$ in Loch Lomond are not highly correlated. The imputation process is initially carried out for the log chlorophyll$_a$ measurements in the first period (1987-1998) and the best approach with the lowest mean square errors is used for imputing missing values in the second period (2002-2005).

## 2.3 Modelling Temperature and Chlorophyll$_a$

A variety of regression approaches are presented to determine the appropriate model that can best describe the variability in temperature and chlorophyll$_a$. The details of each of the models are as follows:

### 2.3.1 Parametric Regression

The first approach considered uses a parametric model including a sinusoidal/ harmonic function, which consists of trend and a seasonal pattern (equation 3.3). This harmonic model is fitted to temperature and log chlorophyll$_a$ for each of the basins.

### 2.3.2 Nonparametric Regression

Modelling via nonparametric regression is more flexible than the first approach as it is not constrained by any particular functional form i.e. the trend and seasonal pattern are not constrained to be linear or monotonic or to have a particular functional form.

Let equation 2.5 be a model of $i$th response $Y_i$ on an unknown smooth function $f$ of a covariate $X_i$,

$$Y_i = f(X_i) + \epsilon_i; i = 1, 2, \ldots, n \tag{2.5}$$

where $\epsilon_i$ are independent random errors with zero mean and constant variance $\sigma^2$. The estimate of the smooth function $f(X_i)$ can often be conveniently expressed in vector-matrix notation as $SY$, where $S$ denotes a smoothing matrix whose rows consist of weights appropriate to estimation at each evaluation point and $Y$ denotes the response in vector form.

The smooth function $f(X)$ from the above equation can be determined via a number of approaches. Nadaraya (1964) and Watson (1964) proposed a simple kernel approach by constructing the local mean estimator. Cleveland (1979) fitted a local linear regression as an alternative to the construction of a local mean for the data. Fan (1993) and Fan and Gijbels (1992) showed the advantage of the local linear estimator compared to the local mean estimator since better theoretical properties are highlighted, particularly features near the edges of the region of the data. Instead of a local linear estimator, (Fan and Gijbels, 1996) used polynomial regression of degree $p$ for estimating the smooth function.

A local polynomial estimator of $p$th order as defined in equation 2.6 is minimized,

$$\min_{\beta_j} \sum_{i=1}^{n} \Big(Y_i - \sum_{j=o}^{p} \beta_j (X_i - x)^j\Big)^2 w\Big(\frac{X_i - x}{h}\Big) \tag{2.6}$$

where $w\Big(\frac{X_i - x}{h}\Big)$ is the kernel density function which is generally a symmetric probability density function with finite second derivative (Simonoff, 1996). The smoothing parameter $h$, which is also known as the span or bandwidth, determines the size of the neighbourhood of $x$ and controls the smoothness of the final estimates (Cleveland and Devlin, 1988); (Cleveland, 1979). A poor choice of bandwidth may result in a low quality estimate (Wand and Jones, 1995). The solution of equation 2.6 at a local point $x$ is obtained as follows,

$$\hat{\beta} = (X^T W X)^{-1} X^T W Y$$

where $Y$ is the $n$-vector with $i$th element of $Y_i$, X is an $n \times (p+1)$ matrix with $(i, j)$th element of $(X_i - x)^j$ and $W$ is the $n \times n$ diagonal matrix with $(i, j)$th element $w\Big(\frac{X_i - x}{h}\Big)$. $\beta_0$ is the estimate at $x$ since this intercept term denotes the position of the local regression line at the local point $x$.

Two different kernel density functions; gaussian and tricube weight functions are used for the curve fitting in this thesis. The gaussian kernel density function used with local linear regression is defined as $\exp\left\{-\frac{1}{2}\frac{(X_i-x)^2}{h}\right\}$ where $h$ is the standard deviation while the tricube weight function used with lowess (or loess) (Cleveland, 1979) is defined as follows,

$$w\Big(\frac{|X_i - x|}{h}\Big) = \begin{cases} \Big[1 - \Big(\frac{|X_i - x|}{h}\Big)^3\Big]^3; w\Big(\frac{|X_i - x|}{h}\Big) \leq 1 \\ \\ 0; w\Big(\frac{|X_i - x|}{h}\Big) > 1 \end{cases}$$

Obviously, the order of the polynomial terms in equation 2.6 can be increased to derive other types of polynomial estimators. Typical choices of $p$ are 0, 1, 2 and 3, with a better asymptotic and boundary bias correction on the local linear ($p$=1) and local cubic ($p$=3) compared to local constant ($p$=0) and local quadratic ($p$=2) estimators, respectively (Clifford et al., 1998). The local linear approach, however, has a very attractive property of smoothing since the bias component does not depend on the pattern of the design points (Bowman and Azzalini, 1997); Fan (1993). Another attractive theoretical property is the good behaviour near the extreme of the design points (Fan and Gijbels, 1996).

Another class of nonparametric regression estimators is based on continuous piecewise polynomials or smoothing splines (Eubank, 1988). One way of estimating the smoothing spline function is by minimizing the penalized criterion (equation 2.7),

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i - f(x_i)\}^2 + h\int_a^b \{f(x)''\}^2 dx \tag{2.7}$$

where $h$ is the global smoothing parameter that determine the trade-off between goodness of fit (in terms of mean square error) and smoothness (in terms of the integrated squared second derivative of $f(x)$). The unique minimizer for the above penalized least squares (equation 2.7) is a natural cubic spline with knots at the distinct values of $x_i$ and spline coefficients that are shrunk according to $h$ (Wahba, 1990). The knot sequence must be carefully chosen since a poor choice can have adverse effects on the estimates (de Boor, 1978).

If the smooth function $f$ is restricted to be a periodic function, then the solution of equation 2.7 is defined by a periodic spline basis expansion,

$$\hat{f}(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i)$$

where $K(x, x_i)$ is the reproducing kernel $\sum_{k=1}^{\infty} \frac{2cos(2\pi k(x-x_i))}{(2\pi k)^4}$. This enables a cyclic smoother to be used for the variable.

### Smoothing parameter

The smooth function produced from nonparametric regression will be close to a straight line as the amount of smoothing increases whilst the line will begin to interpolate the data as the smoothing decreases. Choosing the right $h$ is a crucial step in estimating the smooth function $f$. There are various methods for choosing the smoothing parameters. Akaike Information Criterion (AIC) (Akaike, 1973), cross validation ($CV$) (Stone, 1974), generalized cross validation (GCV) (Craven and Wahba, 1979) and Improved Akaike Information Criterion (AIC$_c$) (Hurvich et al., 1998) are among the typical ways to do so.

The cross-validation ($CV$) function, defined in equation 2.8, is among the most popular criteria used in local linear regression,

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}_{h(i)}(x_i)\right)^2 \tag{2.8}$$

where $n$ is the number of data points, $y_i$ is the $i$th response, $\hat{f}_{h(i)}$ indicates the fitted value at $x_i$ and is computed by leaving out the $i$th observation. The $h$, that minimizes this criterion is adopted. This smoothing selection is also used in the smooth spline (Rice and Silverman, 1991), however, generalized cross-validation $GCV$ (equation 2.9), is often used as an approximation to $CV$ in penalized smoothing splines to avoid the computational complexity of $CV$,

$$GCV(h) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{f}_h(x_i)}{1 - \frac{tr(S)}{n}}\right\}^2 \tag{2.9}$$

where $tr(S)$ is the trace of smoother matrix. The trace is often interpreted by the number of linearly independent explanatory variables in the model (Hastie and Tibshirani, 1990). The generalized maximum likelihood and unbiased risk estimation are other alternatives used in smoothing splines (Cari et al., 2005).

The automatic smoothing selection described above may not always work well especially when the data are correlated in time. Alternatively, a graphical approach is commonly used for selecting smoothing parameters by specifying the degree of freedoms of a model component. Hence, the latter approach is used as a guidance to the required smoothing parameter. Hastie and Tibshirani (1990) define the degrees of freedom for the parameter, variance and error as follows:

$$df_{par} = tr(S)$$

$$df_{var} = tr(SS^T)$$

$$df_{err} = n - tr(2S - SS^T)$$

The above definitions are used throughout this thesis for all models that involve the smoothing parameter.

### 2.3.3 Additive Models

An additive model, introduced by Stone (1985), which is a generalization of a nonparametric regression model, taking more than one explanatory variable into account, is considered in the third approach. This model, which is additive in smoothing functions, is defined in equation 2.10,

$$y = a_0 + \sum_{j=1}^{p} f_j(x_j) + \varepsilon \tag{2.10}$$

where $a_o$ is the intercept, $f_j(x_j)$ is the univariate smoothing function of the $j$th explanatory variable, $\varepsilon$ are the random errors with $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$. The mean of the response $y$ is the sum of smoothing functions of $f_j(x_j)$, defined in equation 2.11.

$$E\left(y|x\right) = a_0 + \sum_{j=1}^{p} f_j\left(x_j\right); j = 1, 2, \ldots, p \tag{2.11}$$

The backfitting procedure (Hastie and Tibshirani, 1990) is one approach that can be used to estimate the smooth functions in the above model and the algorithm is as follows.

1. Initialize: $f_j = f_j^{(0)}, j = 1, 2, \ldots, p$
2. Cycle: $j = 1, 2, \ldots, p, 1, 2, \ldots, p, \ldots$
   $f_j = S_j\left(y - \sum(f_k/x_j)\right)$
3. Continue (2) until the individual smooth functions converge,

where $f_j = \{f_j(x_{1j}), \ldots, f_j(x_{nj})\}^T$.

The additive model for each of temperature and log chlorophyll$_a$ is fitted for two distinct smooth functions of month and year (equation 2.12),

$$y = a_0 + f_1(x_1) + f_2(x_2) + \epsilon \tag{2.12}$$

where $y$ is temperature and log chlorophyll$_a$, respectively, $f(x_1)$ and $f(x_2)$ are the univariate smoothing functions of month and year, and $\epsilon$ are the random errors with zero mean and constant variance.

The `gam` function in the `mgcv` library (Wood, 2005), is used to fit the additive model. The degree of smoothness of model terms is estimated as part of the fitting. The `mgcv` implementation of gam represents the smooth functions using penalized regression splines, and by default uses basis functions for these splines that are designed to be optimal, given the number of basis functions used. The smooths of noncyclical and cyclical terms are controlled by cubic regression spline and a circular smoothing spline, respectively where both splines are determined by a number of knots that leads to the appropriate trend and cyclical patterns.

### 2.3.4 Bivariate Models

The second approach considered the extension of model (2.5) to two dimensions (Bivariate Model), defined by equation (2.13),

$$Y_i = f(X_{1i}, X_{2i}) + \epsilon_i \tag{2.13}$$

where $\epsilon_i$ are the random errors with zero mean and constant variance.

If $X$ denotes an $n \times 3$ design matrix in which its $i$th rows consist of all the elements $\{1(x_{1i} - x_1)(x_{2i} - x_2)\}$, and $W$ denotes a matrix of 0s with the weights $w(\frac{x_{1i}-x_1}{h_1})w(\frac{x_{2i}-x_2}{h_2})$ for each observation down the diagonal, the local linear estimator can be written as the first element of the solution by weighted least squares

$(X^TWX)^{-1}X^TWY$, where $Y$ is the column matrix of response for each observation (Ruppert et al., 2003).

The estimate at point $(x_1, x_2)$ can be carried out via weighted least squares as defined in equation (2.14),

$$\min_{a_0,b_1,b_2} \sum_{i=1}^{n} \{y_i - a_0 - b_1(x_{1i} - x_1) - b_2(x_{2i} - x_2)\}^2 w(x_{1i} - x_1; h_1) w(x_{2i} - x_2; h_2) \quad (2.14)$$

where $w(x_{1i} - x_1; h_1)$ and $w(x_{2i} - x_2; h_2)$ are weights for $x_1$ and $x_2$ which are formed by the two-dimensional kernel function $w\left(\frac{x_{1i}-x_1}{h_1}, \frac{x_{2i}-x_2}{h_2}\right)$.

The bivariate model 2.13 for each of temperature and log chlorophyll$_a$ is fitted to the bivariate smooth function of month and year, where $Y_i$ are temperature and log chlorophyll$_a$, $f(X_{1i}, X_{2i})$ is a bivariate smoothing functions of month $(X_{1i})$ and year $(X_{2i})$.

The `sm.regression` function from the `sm` library (Bowman and Azzalini, 2003), is used to fit the bivariate model 2.13. The smoothing parameter is defined by the degrees of freedom and the value that results in the appropriate trend and seasonal pattern is chosen. The smoothing parameter of model 2.13 is determined by setting $df = tr(S) = 13$ to give a reasonable amount of smoothing in both directions.

### 2.3.5 Semiparametric Models

The semiparametric model which consists of both parametric terms and smooth functions of explanatory variables as defined in equation 2.15, is used to quantify the influence of covariates on the response variable in the fourth approach,

$$y = a_0 + \beta_1 x_1 + f_2(x_2) + \epsilon \tag{2.15}$$

where $\beta(x_1)$ is the parametric term of first explanatory variable $x_1$, $f_2$ is the univariate smoothing function of the second explanatory variable $x_2$ and $\epsilon$ are the random errors with E($\epsilon$)=0 and Var($\epsilon$)=$\sigma^2$.

Cleveland (1979), Silverman (1985), Hastie and Tibshirani (1986), Green (1987) and Speckman (1988) discussed this type of model for independent responses whilst Zeger and Diggle (1994) discuss the semiparametric model for longitudinal data.

This model is fitted via a backfitting algorithm (Moyeed and Diggle, 1994), which can be thought of as having two distinct smoothers $S_1$ and $S_2$. Let $X$ be the full-rank design matrix of explanatory variables and $\beta$ consist of intercept and slope components. The projection of $S_1 = X(X^TX)^{-1}X^T$ produces a least squares fit of X$\beta$ simply denoted by a smooth function $\hat{f}_1$ whilst S$_2$ gives a projection of estimate $\hat{f}_2$. The steps of the backfitting are as follows:

$$\hat{f}_1 = S_1(Y - \hat{f}_2) = X(X^TX)^{-1}X^T(Y - \hat{f}_2) = X\hat{\beta}$$

$$\hat{f}_2 = S_2\left(Y - X\hat{\beta}\right)$$

Hastie and Tibshirani (1986) show that the estimates of $\hat{\beta}$ and $\hat{f}_2$ can be solved explicitly using equations 2.16 and 2.17, respectively.

$$\hat{\beta} = \{X^T(I - S_2)X\}^{-1}X^T(I - S_2)Y \tag{2.16}$$

$$\hat{f}_2 = S_2(Y - X\hat{\beta}) \tag{2.17}$$

The semiparametric model (equation 2.18) of temperature and log chlorophyll$_a$, respectively, is fitted,

$$y = a_0 + \beta_0(t) + \beta_1 \cos\left\{\frac{2\pi\,(x - \theta)}{p}\right\} + \epsilon \tag{2.18}$$

where $y$ is temperature and log chlorophyll$_a$, respectively, $x$ is the month, $\beta_0(t)$ is the nonparametric trend which depends on time $t$, $\beta_1$ is the amplitude, $\theta$ is the phase angle, $p$ is the number of month in a year and $\epsilon$ are the random errors which have zero mean and constant variance.

The steps of fitting the semiparametric model 2.18 are as follows. The `sm.weight` function (Bowman and Azzalini, 2003) is used to estimate the smoothing matrix $S_2$ for the index of month from 1987-2005. The smooth matrix $S_2$ is used in equation 2.16 to estimate $\hat{\beta}$. The fitted values for the semiparametric model is computed as follows;

$$\hat{y} = \hat{a_0} + \hat{f_2} + \hat{\beta}_1 \cos\left\{\frac{2\pi\,(x-\theta)}{12}\right\}$$

where $\hat{\beta}_1$ is the slope component from $\hat{\beta}$ and $\theta$ is determined from the previous harmonic model 3.3.

### 2.3.6 Varying Coefficient Models

A varying coefficient model is an extension of a generalized linear model and is useful for longitudinal studies where the effect of explanatory variables on the response may change over time. Hoover et al. (1998) and Fan and Zhang (2000) discuss and show a few examples of this model in different fields.

This model is generally defined in equation 2.19 where the regression coefficients are allowed to depend on certain explanatory variables (Hastie and Tibshirani, 1986),

$$y_i = \beta_{i0}(r_i) + \sum_{j=1}^{p} \beta_{i0}(r_i)x_{ij} + \epsilon_i \tag{2.19}$$

and the above model can be simplified in a matrix-vector notation form as in equation 2.20,

$$y_i = x_i^T \beta(r_i) + \epsilon_i; i = 1, 2, \ldots, n \tag{2.20}$$

where $y_i$ is the $i$th response, $x_i = (x_{i0}, x_{i1}, \ldots, x_{ip})^T$ are $i$th explanatory variables with $x_{i0} = 1$ which depend on another covariate $r_i$, $\beta_i = (\beta_{i0}, \beta_{i1}, \ldots, \beta_{ip})^T$ are the $i$th coefficients and $\epsilon_i$ are the $i$th random errors which have zero mean and constant variance.

A nonparametric time varying coefficient time series model with a time trend (equation 2.19) is fitted when $r_i$ is denoted by time $\{t_i\}$, with $E(\epsilon_i|x_i) = 0$ and $E(\epsilon_i^2|x_i) = \sigma_i^2(x_i)$ (Cai et al., 2000a). Cai et al. (2000b) highlight the advantage of this model as it allows an increase of predictive utility over a parametric model and the bias of estimates can be reduced significantly.

The following two varying coefficient models are fitted for the fifth and final approaches:

1. Nonparametric trend with varying amplitude and fixed phase angle.
2. Nonparametric trend with varying amplitude and phase angle.

and the above models are defined by equations 2.21 and 2.22, respectively,

$$y = \beta_0(t) + \beta_1(t)\cos\left\{\frac{2\pi\,(x-\theta)}{p}\right\} + \epsilon \tag{2.21}$$

$$y = \beta_0(t) + \beta_1(t)\cos\left\{\frac{2\pi\,(x-\theta(t))}{p}\right\} + \epsilon \tag{2.22}$$

where $y$ are temperature and log chlorophyll$_a$, $x$ is the month, $\beta_0(t)$ is the nonparametric trend which depends on time $t$, $\beta_1(t)$ is the amplitude which also depends on time $t$, $\theta$ is the fixed phase angle and $\theta(t)$ is the varying phase angle which depends on time $t$, $p$ is the number of month in a year and $\epsilon$ are the random errors which have zero mean and constant variance.

The steps of fitting the first varying coefficient model 2.21 are as follows.

1. A bivariate model of each temperature and log chlorophyll is fitted on $\cos\left\{\frac{2\pi(month-\theta)}{12}\right\}$ and index of month; $\hat{y} = \hat{f}(\cos\left\{\frac{2\pi(month-\theta)}{12}\right\}, month)$.
2. The weight function for a matrix $W = [\hat{f}(\cos\left\{\frac{2\pi(month-\theta)}{12}\right\}, month]$ is determined using `sm.weight2` function (Bowman and Azzalini, 2003).
3. A linear model is fitted to $\hat{y}$ (step (1)) on the local point (evaluation point) $(x)$ for $\cos\left\{\frac{2\pi(month-\theta)}{12}\right\}$ (step (2)) for each year as follows; $\hat{y} = \beta_0 + \beta_1 x$.
4. The fitted values for each year are determined using the following varying coefficient model; $\hat{z} = \beta_0 + \beta_1 \cos\left\{\frac{2\pi(month-\theta)}{12}\right\}$ where $\beta_0$ and $\beta_1$ are the varying intercept and slope for each year from step (3).

Finally, the steps of fitting the second varying coefficient model 2.22 are as follows.

1. The weight function for each index of month $(1, 2, \ldots, n)$ is defined using a Gaussian weight function as follows, $\exp\left\{-\frac{1}{2}\frac{(x_i-x)^2}{h}\right\}$ where $x_i$ is the index of month and $x$ is the local point.
2. A nonlinear model of temperature and log chlorophyll, respectively, is fitted using the following model; $\hat{y} = \beta_0 + \beta_1 \cos\left\{\frac{2\pi(month-\theta)}{12}\right\}$, where $\beta_0$, $\beta_1$ and $\beta_2$ are the initial parameters determined from the simple harmonic model 3.3 and the weight function corresponding to the first month from step (1) is incorporated in the model.

3. The estimates of the parameters in the first month from step (2), denoted by $\beta_0(1)$, $\beta_1(1)$ and $\theta(1)$, are used to determine the fitted values for the first time point using the following model; $\hat{z} = \beta_0(1) + \beta_1(1)\cos\left\{\frac{2\pi(month-\theta(1))}{12}\right\}$.

4. Steps (2) - (3) are repeated for the next time points $2, 3, \ldots, n$.

### Correlation Structure of the Errors

In addition to the trends and seasonality in time series data, autocorrelation is likely to be present in the errors. After fitting a model to account for trend and seasonality, an investigation of the relationship of the residuals over time could then be carried out accordingly. In particular, a correlation structure of the residuals can be assessed with both autocorrelation and partial autocorrelation functions. Typically, models of Autoregressive, Moving Average or Autoregressive Moving Average may represent the correlation structure of the errors.

The sample autocorrelation, $r_k$ at lag $k > 0$ is defined in equation 2.23,

$$r_k = \frac{\sum_{t=k+1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2} \tag{2.23}$$

where $t$ is time, $T$ is total number of times, $k$ is lag, $y_t$ is the error at time t, $y_{t+k}$ is the error at time $t + k$, and $\bar{y}$ is the mean of the errors.

A plot of $r(k)$ with $k \geq 0$ is known as a correlogram and the uncertainty of such correlations is often displayed at a certain level of significance. Such an uncertainty is subject to the mean and variance of the sample autocorrelation at lag $k$ as follows.

Let $e_1, e_2, ..., e_N$ be a series of errors which are independent and identically distributed with arbitrary mean and $N$ is the number of errors. Kendall et al. (1983) show that the mean and variance of the autocorrelation of the errors at lag $k$, $r_k$ could be approximated by equations 2.24 and 2.25, respectively,

$$E(r_k) \simeq \frac{-1}{N} \tag{2.24}$$

$$Var(r_k) \simeq \frac{1}{N} \tag{2.25}$$

where $r_k$ is asymptotically normally distributed under weak conditions. The approximate 95% confidence interval of $r_k$ are $\frac{-1}{N} \pm \frac{2}{\sqrt{N}}$, which is often further approximated to $\pm 2/\sqrt{N}$ (Chatfield, 1996).

A partial autocorrelation as defined in equation 2.26 is used as a complement to the autocorrelation function. The uncertainty of the correlation is determined by comparing the coefficients of the partial autocorrelation against the critical region with lower and upper limits, given by $\frac{\pm 2}{\sqrt{n}}$.

$$\Phi_{kk} = \frac{r_k - \sum_{j=1}^{k-1} \Phi_{k-1,j} r_{k-1}}{1 - \sum_{j=1}^{k-1} \Phi_{k-1,j} r_k} \tag{2.26}$$

## 2.4 The Approximate F-Test

The following models have been evaluated for temperature and log chlorophyll$_a$ and the most appropriate model is determined by comparing the models using the approximate F-test.

- Model 1: Harmonic Model (Linear Trend and Constant Seasonal Pattern)
- Model 2: Semiparametric Model (Nonparametric Trend and Constant Seasonal Pattern)
- Model 3: Nonparametric Trend with Varying Amplitude model
- Model 4: Nonparametric Trend with Varying Amplitude and Phase Angle Model.
- Model 5: Additive Model (Nonparametric Trend and Constant Seasonal Pattern)
- Model 6: Bivarate Model (Nonparametric Trend and Varying Seasonal Pattern)

The approximate F-test proposed by Hastie and Tibshirani (1990) is used to assess the model components in each of the six models and the F-statistic of the nested models is defined in equation 2.27,

$$F = \frac{(RSS_o - RSS_1)/(df_o - df_1)}{RSS_1/df_1} \tag{2.27}$$

where this value is F-distributed with degrees of freedom for errors $df_o - df_1$ and $df_1$, $RSS_o$ and $RSS_1$ are residual sum of squares from the full and reduced models, respectively. The residual sum of squares for model $k$ with independent errors is defined in equation 2.28.

$$RSS_k = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.28}$$

## 2.5 Application to Loch Lomond

### 2.5.1 Imputed Values

Table 2.3 summarizes the mean square errors for each of the imputation approaches for temperature and log chlorophyll$_a$. The lowest mean squares errors for temperature and log chlorophyll$_a$ in both basins are given by the second approach and so, the imputed values from the harmonic model are used to replace the missing temperature and log chlorophyll$_a$ measurements.

| Approach | Temperature | | Log chlorophyll$_a$ | |
|---|---|---|---|---|
| | North | South | North | South |
| 1 | 3.91 | 4.12 | 0.22 | 0.35 |
| 2 | 2.71 | 2.83 | 0.06 | 0.12 |
| 3 | 3.14 | 3.43 | 0.13 | 0.20 |
| 4 | 2.91 | 3.12 | – | – |

**Table 2.3.** Comparison of the Mean Square Errors of different imputation approaches on temperature and log chlorophyll$_a$.

### 2.5.2 Modelling on Temperature and Log chlorophyll$_a$

Figures 2.3 and 2.4 present the models for temperature for the north (left) and south (right) basins. The fitted models are displayed and the points represent the measurements including the imputed values. Generally, the fitted models for temperature are very similar for each modelling approach.

Fitted Model 1 for the north and south basins are defined in equations 2.29 and 2.30, respectively, and they are illustrated in Figure 2.3 (top). The standard error for each coefficient is marked in parentheses underneath the coefficient. There is evidence of positive trend in the north but no trend in the south basin, indicating a rise in temperature in the north but no statistically significant change in temperature in the south basin during 1987 - 2005. The amplitude of the seasonal cycle in the south is slightly larger than the north basin, suggesting higher variability in temperature is exhibited in the shallower basin. The adjusted coefficients of determination of the models with respect to the north and south basins are 58.4% and 71.9%, indicating that a moderate percentage of temperature variability is explained by the trend and seasonality terms in both basins. The models indicate an increase of temperature in the north basin of approximately $1.6^oC$ and no increase in the south basin.

$$temp_{(t)} = \underset{(0.42)}{9.43} + \underset{(0.003)}{0.007t} - \underset{(0.29)}{5.12}cos\left\{2\pi\left(\frac{t-\underset{(0.10)}{45.47^o}}{12}\right)\right\} \tag{2.29}$$

$$temp_{(t)} = \underset{(0.17)}{10.28} - \underset{(0.25)}{6.00}cos\left\{2\pi\left(\frac{t-\underset{(0.07)}{44.59^o}}{12}\right)\right\} \tag{2.30}$$

Figure 2.3 (centre) shows Model 2 where smooth trends and constant seasonal patterns are exhibited in both basins. The smooth trend rises gradually in the north whilst it is approximately constant in the south basin. Figure 2.3 (bottom) displays Model 3 with varying trends and small changes in seasonal patterns which are more apparent in the north than in the south basin. The temperature in the north increases slowly over the year, however, an approximately constant temperature is evident in the south basin.

Figure 2.4 (top) shows Model 4 where the varying trend in the north rises gradually but there is an approximately constant temperature in the south basin. The varying seasonal pattern in the south is almost constant but larger than that in the north basin. Figure 2.4 (centre) shows Model 5 where smooth trends and constant seasonal patterns are apparent in both basins. The smooth trend increases gradually in the north whilst no clear changes in trend are highlighted in the south basin. Figure 2.4 (bottom) shows Model 6 in which the smooth trends highlight a small increase from 1995 until middle of the period and starting to decrease up to 2005 in both basins.

For log chlorophyll$_a$, all of the fitted models for the north and south basins are shown in Figures 2.5 and 2.6. The fitted models and points represent the fitted values and measurements including imputed values, respectively. The differences between the fitted models are very small and from the pictures, Model 1 appears to represent the data as well as the others.

Model 1 for the north and south basins is defined in equations 2.31 and 2.32 and is depicted in Figure 2.5 (top). There is evidence of positive trends and constant seasonal patterns in both basins, indicating the rise of log chlorophyll$_a$ with an annual cycle from 1987-2005. The amplitude of the seasonal pattern in the north is relatively larger than the south basin, indicating a higher variability in log chlorophyll$_a$ in the deeper water of the north basin. The adjusted coefficient of determination of the models with respect to the north and south basins are 64% and 23%, showing that a moderate variability of log chlorophyll$_a$ is explained by the explanatory variables for the north basin whilst considerably lower variability is apparent in the model for the south basin.

$$log(chl_a)_t = \underset{(0.03)}{-0.33} + \underset{(0.0002)}{0.002t} - \underset{(0.02)}{0.41}cos\left\{2\pi\underset{(0.11)}{\left(\frac{t-63.54^o}{12}\right)}\right\} \qquad (2.31)$$

$$log(chl_a)_t = \underset{(0.03)}{0.09} + \underset{(0.0002)}{0.002t} - \underset{(0.02)}{0.18}cos\left\{2\pi\underset{(0.31)}{\left(\frac{t-49.26^o}{12}\right)}\right\} \qquad (2.32)$$

Figure 2.5 (centre) shows model 2 where smooth trends and constant seasonal patterns are exhibited in both basins. The smooth trends in both basins increase from 1987-1998 but become approximately constant from 2002-2005. Figure 2.5 (bottom) displays Model 3 which highlights an increase of levels over time in both basins. Similar seasonal patterns are clearly shown in the north than in the south basin from 1987-1998, however, differences in seasonality between 2002 and 2005 in both basins are apparent.

Figure 2.6 (top) shows Model 4 where the varying trends slightly increase in both basins. A constant seasonal pattern is highlighted in the north whilst in the south basin, a change in the seasonality is observed from 1987-1998 but this become more constant from 2002-2005. Figure 2.6 (centre) shows Model 5, with similarity in smooth trends and constant seasonal patterns evident in both basins. Figure 2.6 (bottom) shows Model 6 in which the smooth trends gradually increase in both basins. In the north basin, the seasonal pattern is approximately constant from 1987-1998 but tends to change in the latter part (2002-2005) of the periods. Conversely, the seasonal pattern in the south gradually changes in the first part of the period (1987-1998) but then become more constant from 2002-2005.

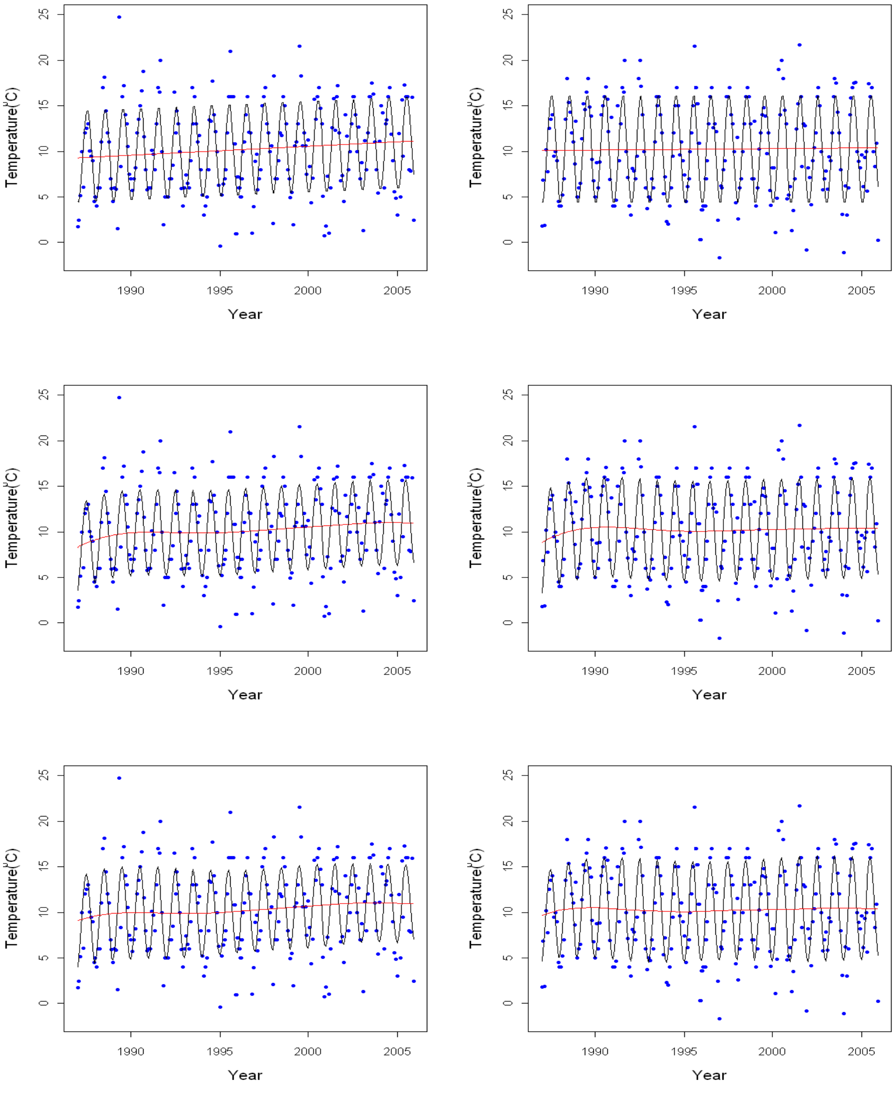

**Figure 2.3.** Fitted models of surface water temperature from 1987 - 2005 via Model 1 (top), Model 2 (centre) and Model 3 (bottom) for the north (left) and south (right) basins.

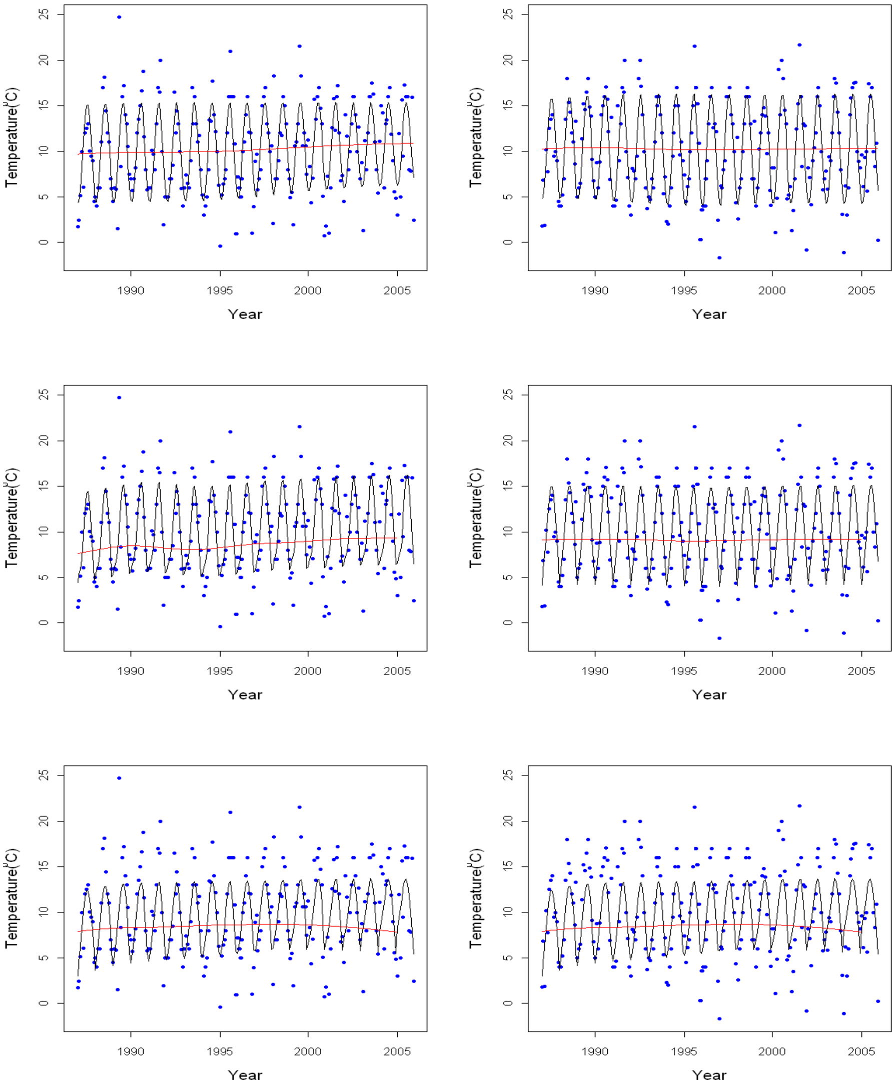

**Figure 2.4.** Fitted models of surface water temperature from 1987 - 2005 via Model 4 (top), Model 5 (centre) and Model 6 (bottom) for the north (left) and south (right) basins.

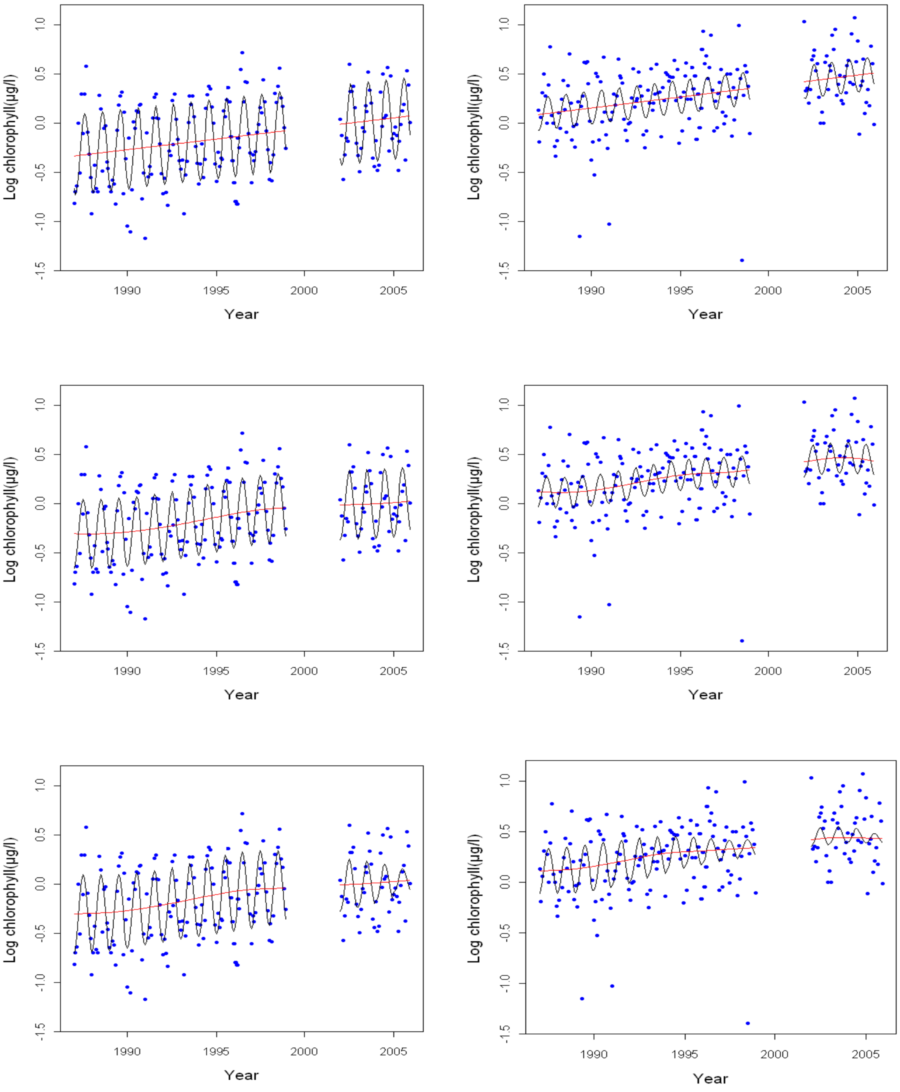

**Figure 2.5.** Fitted models of log chlorophyll$_a$ from 1987 - 2005 via Model 1 (top), Model 2 (centre) and Model 3 (bottom) for the north (left) and south (right) basins.

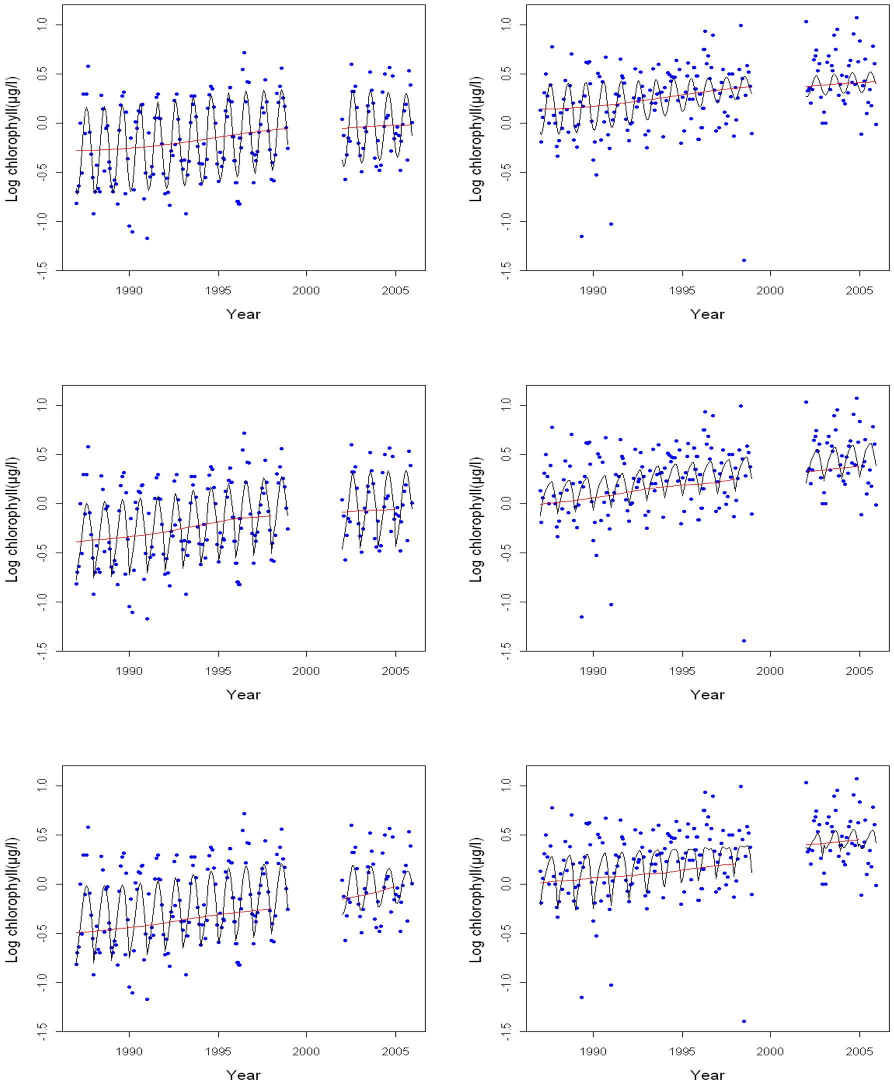

**Figure 2.6.** Fitted models of log chlorophyll$_a$ from 1987 - 2005 via Model 4 (top), Model 5 (centre) and Model 6 (bottom) for the north (left) and south (right) basins.

### 2.5.3 Comparison of Models of Temperature and Log Chlorophyll

Tables 2.4 and 2.5 show the residual sum of squares (RSS) and degrees of freedom (df) for different models of temperature and log chlorophyll$_a$, respectively. The lowest RSS is produced by fitted Model 5 for temperature for both basins but for log chlorophyll$_a$, the fitted Models 4 and 5 provide similarly low RSS. A formal statistical comparison is essential to determine the most appropriate model. In particular, the approximate F-test is required to show any evidence of parametric trend and seasonality in temperature over the year whilst for log chlorophyll$_a$, the smooth trend and seasonality for 19 years needs to be investigated.

The approximate F-test allows a comparison between a complex and simple model, with the lower and higher RSS, respectively. For temperature in the north and south basins, Model 5 shows the lowest RSS followed by Models 4 and 1, suggesting that they are the only models that can be reasonably compared. Similar models can be compared for log chlorophyll$_a$ in the north basin, however, a greater number of models for log chlorophyll$_a$ in the south basin can be compared.

| Model | RSS | | df | |
|---|---|---|---|---|
| | North | South | North | South |
| 1 | 2135.26 | 1593.57 | 224 | 224 |
| 2 | 2497.24 | 2116.55 | 222.8 | 222.8 |
| 3 | 2506.41 | 2121.78 | 221.1 | 221.1 |
| 4 | 2049.82 | 1544.76 | 217.7 | 217.7 |
| 5 | 1924.41 | 1451.84 | 217.00 | 217.1 |
| 6 | 2249.99 | 2273.98 | 215 | 215 |

**Table 2.4.** Residual Sum of Squares (RSS) and degree of freedoms (df) of different models of temperature for the north and south basins.

| Model | RSS | | df | |
|---|---|---|---|---|
| | North | South | North | South |
| 1 | 10.61 | 18.49 | 188 | 188 |
| 2 | 14.47 | 18.56 | 187.3 | 187.3 |
| 3 | 14.02 | 18.00 | 186.6 | 186.6 |
| 4 | 10.02 | 17.80 | 181.8 | 181.8 |
| 5 | 10.06 | 17.84 | 181.3 | 181.1 |
| 6 | 10.80 | 18.38 | 179 | 179 |

**Table 2.5.** Residual Sum of Squares (RSS) and degree of freedoms (df) of different models of log chlorophyll$_a$ for the north and south basins.

The approximate F-statistic and its p-values for models of temperature and log chlorophyll$_a$ are tabulated in Tables 2.6 and 2.7, respectively. The results in Table 2.6 show the appropriateness of Model 5 to explain the variability of temperature in the north and south basins. Table 2.7, however, shows evidence of a parametric trend and constant seasonal pattern for log chlorophyll$_a$ for both basins, highlighting the appropriateness of the harmonic model (Model 1) to explain the variability of log chlorophyll$_a$ from 1987-1998 and 2002-2005.

The above results show that the additive model appears to be the most appropriate for temperature whilst the harmonic model with positive trend is more appropriate for log chlorophyll$_a$.

| Model | | F-value | | p-value | |
|---|---|---|---|---|---|
| North | South | North | South | North | South |
| 1 and 4 | 1 and 4 | 1.45 | 1.09 | 0.152 | 0.370 |
| 1 and 5 | 1 and 5 | 3.39 | 3.02 | 0.01 | 0.030 |

**Table 2.6.** The approximate F-test for the models of temperature for the north and south basins.

| Model | | F value | | p-value | |
|---|---|---|---|---|---|
| North | South | North | South | North | South |
| 1 and 4 | 1 and 3 | 1.73 | 3.17 | 0.1163 | 0.078 |
| – | 1 and 4 | – | 1.13 | – | 0.346 |
| – | 1 and 5 | – | 0.98 | – | 0.457 |
| – | 1 and 6 | – | 0.11 | – | 0.988 |

**Table 2.7.** The approximate F-test for the models of log chlorophyll$_a$ for the north and south basins.

The assumptions of linearity and constant variance of the residuals from the additive models of temperature in both basins are checked. For illustration, the diagnostic plots for the north basin are illustrated in Figure 2.7. The plot of the residuals against fitted values (top left) highlights linearity and constant variance of the residuals whilst the plots of acf (top right) and pacf (bottom left) suggest no correlation structure of the residuals since the correlation coefficients lie within the 95% confidence intervals (horizontal dashed lines).

Similarly, the assumptions of linearity and constant variance of the residuals are checked for the harmonic model of log chlorophyll$_a$ in both basins. Additionally, the normality plot is produced to determine the distribution of the residuals from the parametric model 1. The diagnostic plots of the residuals for model 1 in the north basin are displayed in Figure 2.8. The plot of residuals (top left) suggests the linearity and constant variance of the residuals and the plots of acf (top right) and pacf (bottom left) suggest no correlation structure for the residuals. The normality plot (bottom right) suggests that the residuals are normally distributed.

Similar results are given for temperature and log chlorophyll$_a$ in the south basins, satisfying the model assumptions.

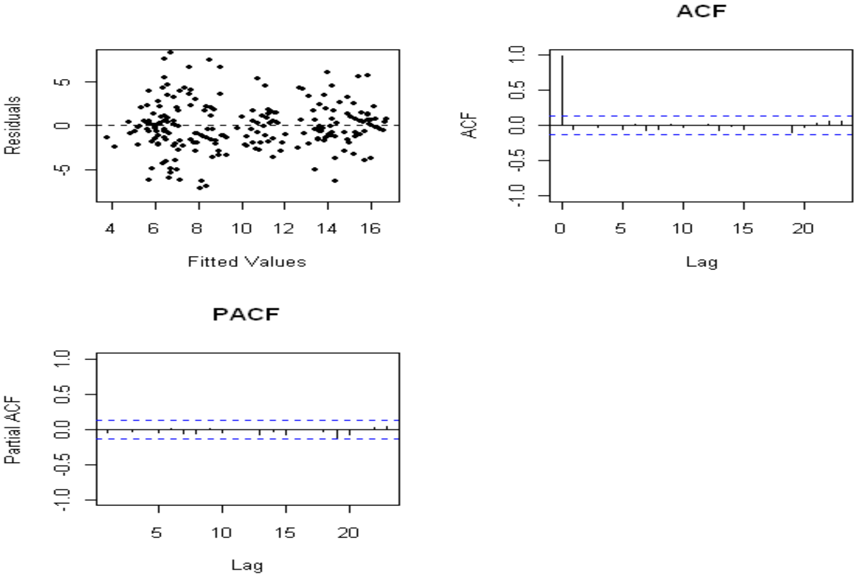

**Figure 2.7.** Diagnostic plots for the additive model of temperature (model 5) for the north basin.

Hence, the additive models for temperature in both basins are defined in equation 2.33, respectively, whilst the harmonic models of log chlorophyll$_a$ for the north and south basins are defined in equations 2.31 and 2.32, respectively.

$$temp = f_1(month) + f_2(year) + \epsilon \tag{2.33}$$

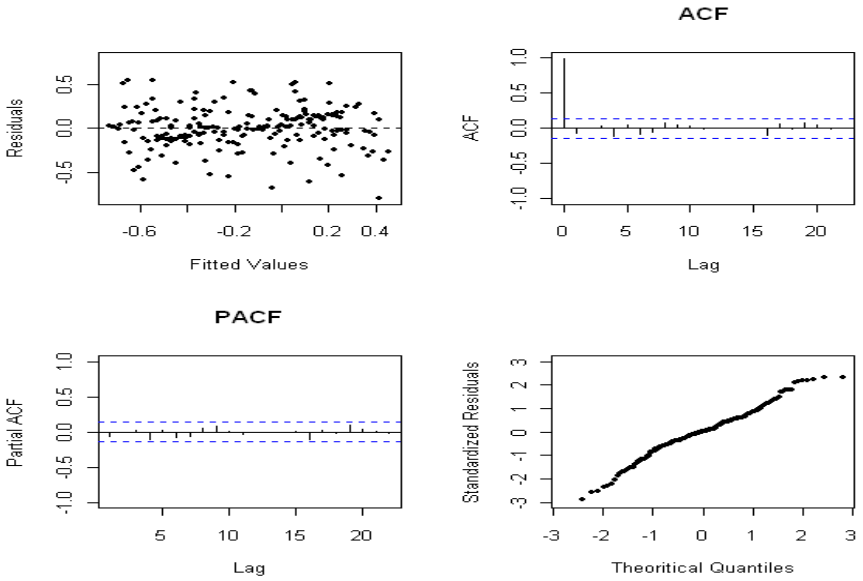

**Figure 2.8.** Diagnostic plots for the harmonic model of log chlorophyll$_a$ (model 1) for the north basin.

## 2.6 Assessing Ecological Relationships

Increasing temperature in lakes may enhance the photosynthesis in the phytoplankton (Helmut and Thomas, 2008) and so, the rise of chlorophyll$_a$ concentrations could be observed. Additionally, changes in phytoplankton abundance is likely to occur following the rise of nutrients in lakes. For instance, Smith et al. (1999) shows that the increase of nitrogen (N) and phosphate (P) concentrations results in the bloom of the phytoplankton community structure in lakes. Since the previous findings show the evidence of trends and seasonal patterns in log chlorophyll$_a$ in the north and south basins of Loch Lomond, it is of interest to investigate the influence of temperature, nitrate and phosphate in the model of log chlorophyll$_a$.

In the following, the relationships between log chlorophyll$_a$ and temperature, P and N are initially explored and their patterns of relationships are used as a basis to model log chlorophyll$_a$. The correlation structure of log chlorophyll$_a$ is then investigated and finally, model testing is carried out via approximate F-tests to determine the appropriate models for the north and south basins.

### 2.6.1 Plots of relationships of log chlorophyll$_a$ on temperature and nutrients.

For illustration, Figure 2.9 displays the plots of relationships between log chlorophyll$_a$ and each of temperature, phosphate and nitrate in the south basin. The log chlorophyll$_a$ gradually increases as temperature rises up to $10^oC$ and this phenomenon is expected as they are often related naturally. However, the unexpected decline in log chlorophyll$_a$ as the temperature continues to increase from $10^oC$ - $20^oC$ is exhibited and this feature is likely due to unidentified factors. While an approximate quadratic relationship is shown between the log chlorophyll$_a$ and temperature, no clear relationships are highlighted between log chlorophyll$_a$ and nitrate or phosphate. Rather, log chlorophyll$_a$ remains constant as both phosphate (top right) and nitrate (bottom left) increase, contradicting the natural behaviour of their relationships in lakes. The constant relationships may suggest the inadequacy of the nitrate and phosphate concentrations to reflect the rise of chlorophyll concentrations naturally.

Higher variability of each of the nutrients on log chlorophyll$_a$ is also highlighted in the above plots and so, a natural log transformation is used to stabilize the variance. Figure 2.10 presents the plots of relationships between log chlorophyll$_a$ with each of log phosphate and log nitrate, highlighting greater stability in the variance of log phosphate compared to the actual measurements. The variability

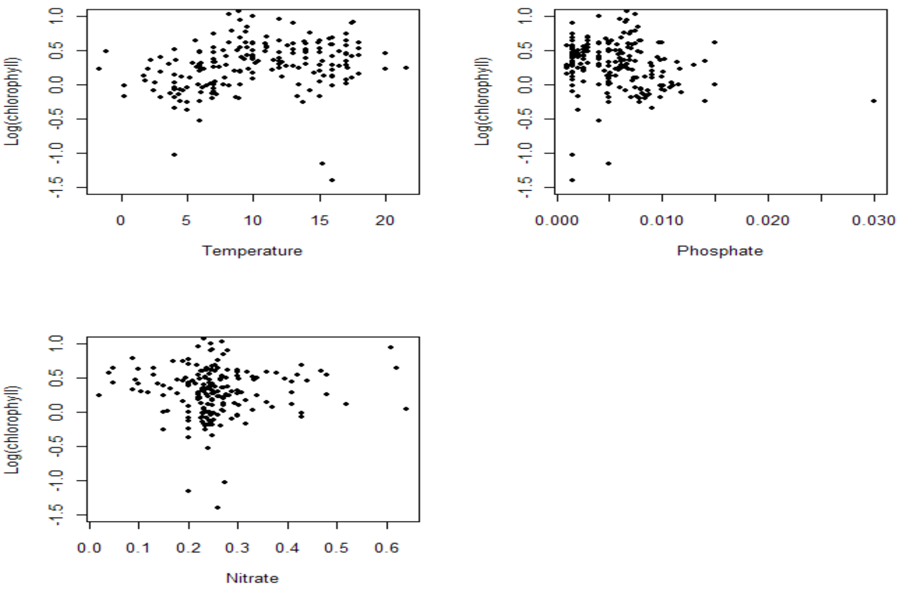

**Figure 2.9.** Plots of the relationships between log chlorophyll$_a$ and each of the temperature (top left), phosphate (top right) and nitrate (bottom left) at Creinch.

of log nitrate, however, is similar to the actual measurements, suggesting that both the actual and transformed nitrate measurements show no clear differences in term of their variability with log chlorophyll$_a$.

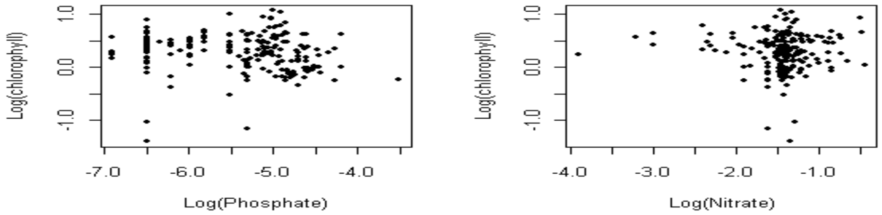

**Figure 2.10.** Plots of the relationships of log chlorophyll$_a$ on log phosphate (left) and log nitrate (right) at Creinch.

### 2.6.2 Correlation Structure of the Residuals

The previous investigation shows the appropriateness of a harmonic model (Model 1) for log chlorophyll$_a$ on trend and seasonal pattern for the north and south basins. In addition, Figures 2.9 and 2.10 above highlight the approximate quadratic pattern between log chlorophyll$_a$ and temperature but no clear patterns are shown between log chlorophyll$_a$ and each of log phosphate and log nitrate. A formal investigation on the influence of all potential predictors on log chlorophyll$_a$ is then carried out via a statistical modelling approach. In the initial part of the modelling strategy, the residuals from a parametric model 2.34 are extracted for testing the correlation structure in the time series.

$$\begin{aligned} log(chl_a)_t &= \beta_0 + \beta_1 year_t + \beta_2 cos\left\{2\pi\left(\frac{year_t - \theta}{12}\right)\right\} + \\ &\quad \beta_3 temp_t + \beta_4 temp_t^2 + \beta_5 log(P)_t + \\ &\quad \beta_6 log(N)_t + \varepsilon_t \end{aligned} \tag{2.34}$$

For illustration, diagnostic plots of the above model for the south basin are shown in Figure 2.11. The plot of residuals against the fitted values (top left) shows the linearity and slightly constant variance of the residuals. Both plots of the autocorrelation (top right) and partial autocorrelation (bottom left) highlight that almost all of the correlations are within 95% confidence intervals. The normality plot (bottom right) indicates the normality of the residuals is satisfied. Hence, the diagnostic plots suggest that the residuals are essentially white noise which is identically and normally distributed with zero mean and constant variance.

Similar results are given in the north basin, highlighting no evidence of correlation in the residuals time series.

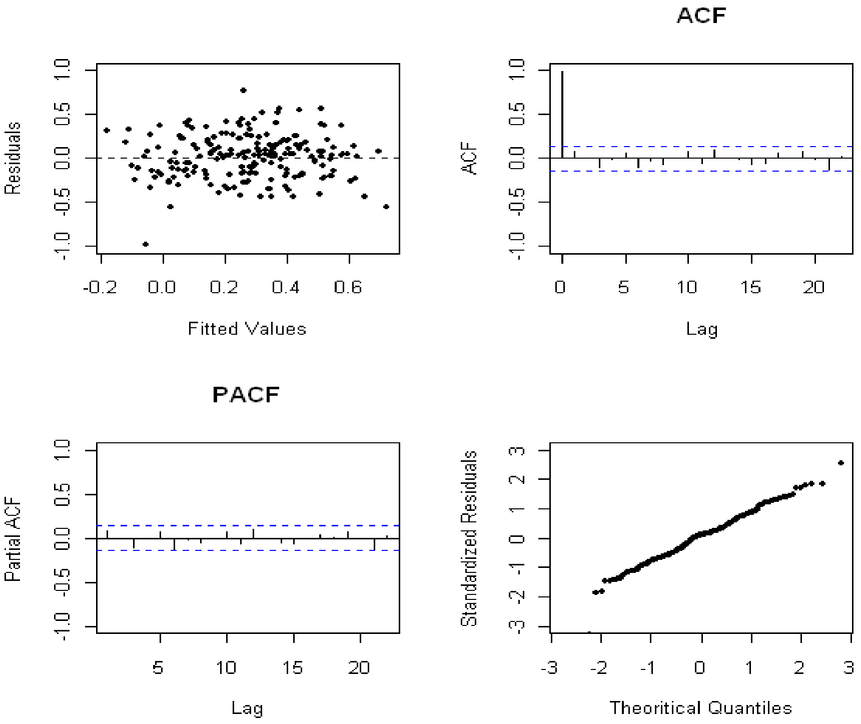

**Figure 2.11.** Plot of residuals against fitted values (top left), autocorrelation (top right) and partial autocorrelation (bottom left) functions of the residuals and normality plots of the standardised residuals (bottom right), from Model 2.34 at Creinch.

### 2.6.3 Comparison between Models of Log Chlorophyll on Temperature and Nutrients

As an alternative to Model 2.34, the nonparametric Model 2.35 of log chlorophyll$_a$ is fitted to the log chlorophyll for the north and south basins to allow smooth functions of each predictors. In particular, this model allows smoothing functions of trend, seasonality, temperature, log nitrate and log phosphate on the log chlorophyll,

$$log(chl_a)_t \quad = \quad \beta_0 + m_1(year_t) + m_2(month_t) + m_3(temp_t)$$

$$+m_4\{(log(P)_t)\} + m_5\{(log(N)_t)\} + \varepsilon_t \qquad (2.35)$$

where $\varepsilon_t$ are random errors with zero mean and constant variance. Model 2.35 is fitted using the `gam` function in the `mgcv` library of R. The cubic spline basis is used to construct weights for year, temperature, log P and log N components whilst a circular smoother is used for constructing the weights for month. The number of chosen knots allows a reasonable amount of smoothness for each component in the additive model 2.35 and results in the approximate degree of freedom of 15 for model 2.35. For illustration, the smooth model for log chlorophyll with its component functions from model 2.35 are displayed in Figure 2.12.

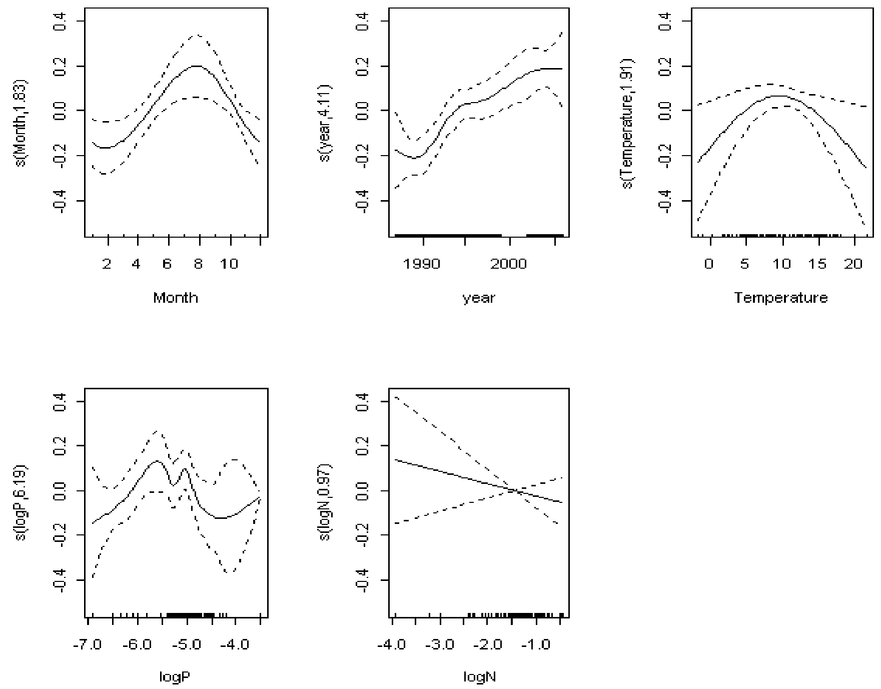

**Figure 2.12.** Plot of smooth function of log chlorophyll against each of the predictors for the south basin.

Models 2.34 and 2.35 are statistically compared using the approximate F-test to determine the most appropriate model and the results are tabulated in Table 2.8, suggesting the appropriateness of Model 2.34 for the north and south basins since the p-values are greater than 0.05 and hence, no complex model is required.

| Model | | F value | | p-value | |
|---|---|---|---|---|---|
| North | South | North | South | North | South |
| 2.34 and 2.35 | 2.34 and 2.35 | 1.364 | 2.746 | 0.183 | 0.100 |

**Table 2.8.** The approximate F-test for the models of log chlorophyll$_a$ on the year, month, temperature, log P and log N.

Thus, the significance of each predictor in model 2.34 for both basins is determined using F-test. In particular, the non-significant terms are removed and the model which consists of only significant predictors is refitted. The appropriate models of log chlorophyll$_a$ for the north and south basins are defined in equations 2.31 and 2.36, respectively.

The trend and seasonality are the only significant predictors in the model of log chlorophyll$_a$ for the north basin, indicating a similar model as previously fitted on year and month (equation 2.31). The model indicates that the non-significant influence of temperature and nutrients on variability of log chlorophyll$_a$ in the deeper location of the loch. In the south basin, however, trend, seasonality and temperature are significant predictors in the model of log chlorophyll$_a$. The addition of temperature in the model results in the increase of information on the variability of log chlorophyll over the year from 23% (model 2.32) to 33% (model 2.36).

$$log(chl_a)_t \quad = \quad 0.02_{(0.11)} + 0.002_{(0.0003)}(year)_t$$

$$-0.20_{(0.05)}cos\left\{2\pi\left(\frac{year_t - 50^o_{(0.21)}}{12}\right)\right\}$$
$$+0.04_{(0.01)}(temp)_t - 0.003_{(0.0008)}(temp)^2_t + \varepsilon_t \qquad (2.36)$$

The diagnostic plots of Model 2.36 for the south basin are illustrated in Figure 2.13, suggesting that the residuals are identically and normally distributed with zero mean and constant variance. Since all the statistical assumptions for models of log chlorophyll$_a$ for both basins are satisfied, they could be used for predicting phytoplankton in the deeper (north) and shallower (south) locations of the loch.

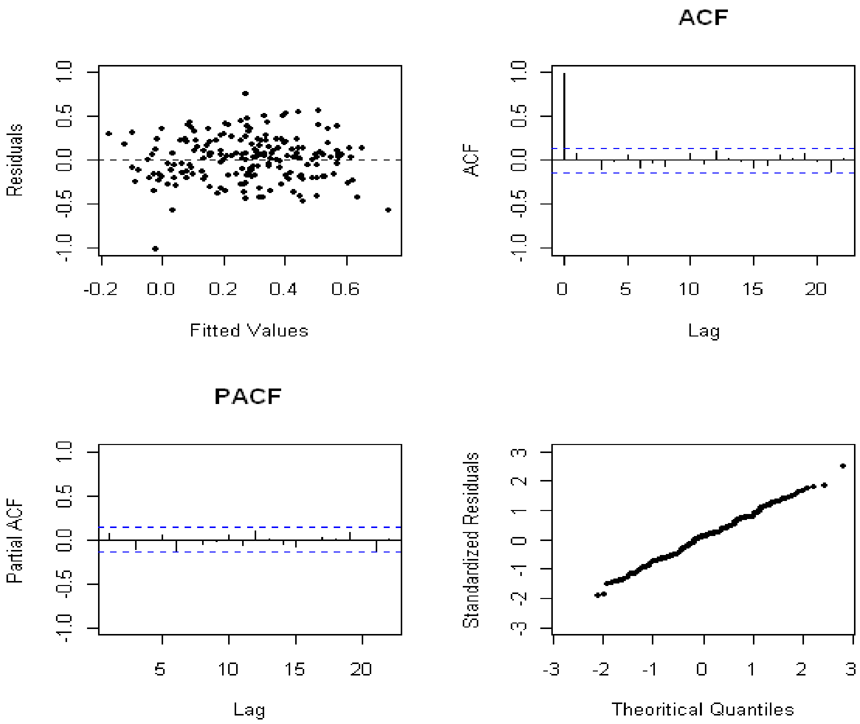

**Figure 2.13.** Plot of residuals against fitted values (top left), autocorrelation (top right) and partial autocorrelation (bottom left) functions of the residuals and normality plots of the standardised residuals (bottom right), from Model 2.36 at Creinch.

## 2.7 Summary

A series of statistical models fitted to temperature and log chlorophyll$_a$ for the north and south basins in Loch Lomond highlight constant seasonal patterns for both sites. Temperature rises smoothly whilst log chlorophyll$_a$ increases linearly in both basins.

The analysis of the 19 year temperature time series highlights smooth trends of temperature in the north and south basins. In general, there is an increase of temperature in the north but approximately constant in the south basin. There are no particular large changes over the year for both basins which is quite different to the 4-5$^oC$ increase proposed by Krokowski (2007). This difference might be due to the consideration of only a parametric models with trend and using a complete case analysis, which ignores the missing values in winter in the later years. Constant and smooth seasonal patterns are exhibited for both basins, suggesting constant annual cycle over the 19 years.

In log chlorophyll$_a$, there are increases in the north and south basins of about 0.29 $\mu g/l$ from 1987 to 1998 and 0.1 $\mu g/l$ from 2002 to 2005, in broad agreement (after transformation) with that reported by Krokowski (2007). Constant seasonal patterns are highlighted in both basins for both periods.

The addition of temperature, log nitrate and log phosphate in the time series models of log chlorophyll$_a$ leads to the evidence of linear trend and seasonal pattern in the north but for the south basin, the evidence of temperature, linear trend and seasonal pattern are observed. The changes in temperature have significant impact on the bloom of phytoplankton in the south basin. Nitrate and phosphate appear to be in significant in the north and south basins, suggesting that they are not part of the primary controls on the changes of log chlorophyll$_a$. As an implication, it is likely that eutrophication has not occurred in the north and south basins for 19 years, indicating a good quality water can be obtained from the loch.

# Chapter 3

# Temporal Temperature Patterns with Depth

## 3.1 Introduction

The investigation of temporal patterns in surface temperature in Loch Lomond is extended by considering temperature data from thermistor chains with a moderate temporal frequency of 1 and 3 hourly data, at 11 different depths. The investigation of higher frequency temperature data, recorded at several locations and at different depths in the loch may provide further insights into any apparent changes in ecological processes. Two questions of interest are as follows:

- Are changes in temperature over time consistent at different depths and different sites in the loch?
- How does the temperature profile stratify with depth?

For the first question, a statistical model is an appropriate way to explain the mean change in temperature over the year at different depths.

For the second question, the temperature profiles with depth for each time point over the year may have particular patterns and therefore, a certain prominent and natural feature in the lake could be identified (that of a thermocline).

A well known natural phenomenon that occurs within most lakes during the summer is the development of temperature stratification. Conversely, no apparent deposition of layers in lakes appears over winter. The stratification of temperature in the water column in summer results in different characteristics of temperature profiles with depth over time. In particular, the temperature profiles with depth in summer, can be essentially viewed as a smooth and continuous curve with two bends of different degrees of curvature as illustrated in Figure 3.1 (Victor and Robin, 2005). The two bends divide the profile into 3 zones, which are an upper warmer zone, an intermediate zone and a bottom colder zone which are known as epilimnion, metalimnion and hypolimnion, respectively. The upper stratum (the epilimnion) is more or less uniformly warm, circulating and fairly turbulent. The lowest stratum (the hypolimnion) is relatively calm and cold. The metalimnion exhibits a marked thermal discontinuity. The metalimnion is defined as 'the water stratum of steep thermal gradient, bounded by the intersections of the nearby zones i.e. epilimnion and hypolimnion' (Wetzel, 2001).

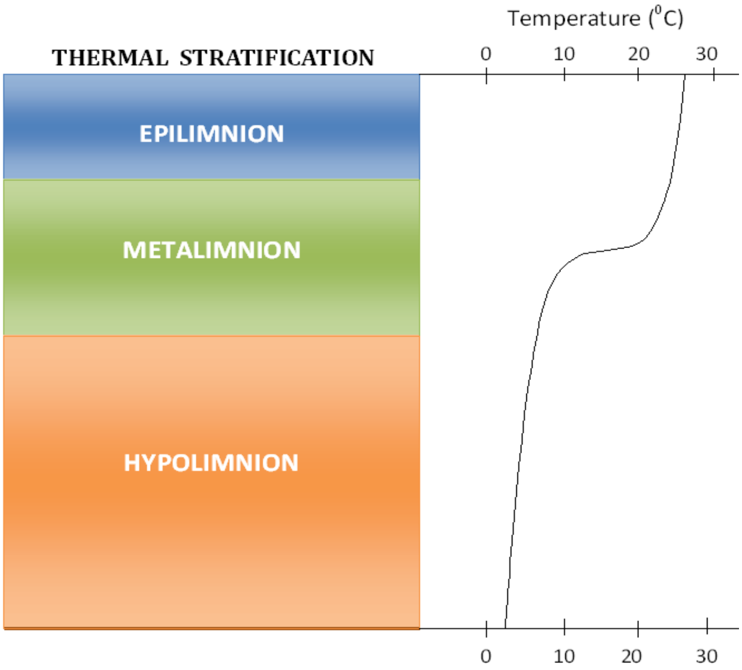

**Figure 3.1.** Thermal stratification in lakes during summer season

Early analysis by Ricker (1937) and Hutchinson (1937) has shown that the heating of a stratified lake is a result of several factors such as solar radiation, turbulent conduction and biological processes. The radiation and turbulence from the epilimnion transfers the heat to the strata underneath, particularly in warmer periods of the year (Imberger and Patterson, 1990). The heat conduction in the epilimnion and metalimnion decreases as the stratification process reaches completion. In hypolimnion, the heat conduction is very small, resulting from the biological oxidation during decomposition process in lakes (Wetzel, 2001).

The above process is often referred to as the formation of the thermocline with its conceptual basis extensively discussed by Hutchinson (1957). According to widely accepted limnological convention, the thermocline is defined as an imaginary plane located at the depth where the rate of change of decrease (temperature gradient) in the temperature profile is maximum (Victor and Robin, 2005). This phenomenon is depicted in Figure 3.2, where the thermocline plane is somewhere in between the mixed layer and deep water. Mathematically, the thermocline depth is the inflection point of the temperature curve with depth where the temperature gradient changes sign of second derivatives (Victor and Robin, 2005).

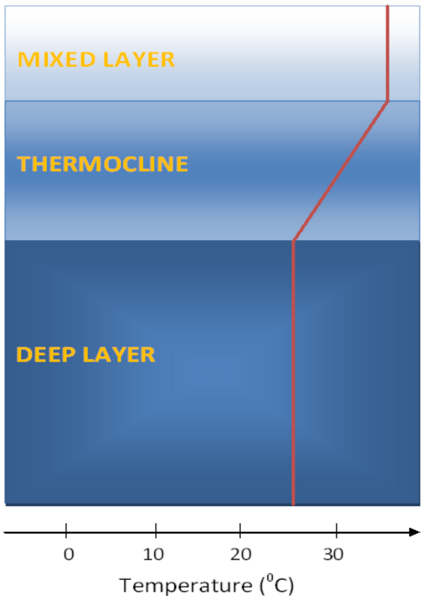

**Figure 3.2.** The thermocline zone in the water column

Therefore, the objectives of this chapter are twofold:

- to model the temperature over time with depth at different sites in the loch.
- to investigate the position of the thermocline since it partitions the water column into two strata with different biological and chemical features, reflecting the ecological process in lakes.

## 3.2 Data

The study is carried out on thermistor data which consist of temperature measurements at four different sites in Loch Lomond. The temperature was recorded at 3-hourly intervals at 11 different depths, from 1 September 2002 until 31 August 2003, at Cailness (north basin), Creinch (south basin) and Ross Point. From 17 April 2008 to 27 May 2009, 1-hourly temperature measurements ($^oC$) were recorded at 11 different depths in the mid basin. Each of these sites has different depths of temperature measurements.

The data of 3-hourly temperature measurements were supplied by the Scottish Centre for Ecology and the Natural Environment (SCENE) whilst the 1-hourly temperature measurements were contributed by the Scottish Centre for Ecology and the Natural Environment (SCENE) and Prof. Susan Waldron from the University of Glasgow.

## 3.3 Exploratory Analysis

The hydrological year is a 12-month period, usually selected to begin and end during a relatively dry season and used as a basis for processing streamflow and other hydrological data. The periods from September - August or October - September are often used in Britain to represent the hydrological year. Since the temperature in the loch could be influenced by hydrological events, the patterns of 3-hourly temperature measurements over the hydrological year, with depths, at Cailness, Creinch and Ross Point are explored. The plots of temperature over the hydrological year at different depths are used to provide an early impression of the temperature profiles. The occurrence of different patterns in temperature over the year at different depths may highlight different features down the depth profile of the loch.

Figures 3.3 and 3.4 show the temperature pattern over the hydrological year at different depths at Cailness, Creinch and for the lower and upper chains of Ross Points, respectively. The depths vary between sites and some of the measurements are missing.

In the north basin (Figure 3.3) (top), the temperature shows strong patterns in the shallower depths while at greater depths, the temperature exhibits only a weak fluctuation. These characteristics might be due to constant solar radiation over the surface water, resulting in heat transfer via a conduction process down to a certain depth. However, the deeper water only gains heat from the upper layer as it is not directly affected by the solar radiation and therefore, remains cooler with less variability.

In the south basin (Figure 3.3) (bottom), the temperature shows a strong pattern at each depth. This might be due to the fact that the south is shallower than the north basin and therefore, the patterns of the temperature at all depths do not change much since all depths could be affected by a similar amount of heat transfer from radiation and so, the temperature profile at a series of depths seems very similar.

The lower chain of Ross Point (Figure 3.4) (top) and upper chain of Ross Point (Figure 3.4) (bottom) highlight similar step functions in the temperature pattern. The strange pattern is mostly likely due to instrumental problems and so, no further analyses are carried out for this site.

Figure 3.5 shows an alternative view of the temperature patterns over the year, with depths in the north (top) and south (bottom) basins. The red and green curves represent the measurements in 2002 and 2003, respectively. In the north basin, there is an apparent difference in the variability of the temperature pattern for each depth from September to November 2002 and April to August 2003. However, the temperature pattern for each depth exhibits less variability from December 2002 - March 2003. This is due to the fact that heat conduction, beginning from the water surface down the depth profile is not largely affected by the radiation process and so, the conduction rate from the top down to the bottom of the lake slowed considerably. The depth profile indicates weak heat conduction. In the south basin, the variability of the temperature pattern with depth is similar from September 2002 - May 2003 and starts to diverge thereafter. The similarity of the temperature over these months, with depth, could be a result of similar heat conduction from solar radiation over the surface down to 12 metres depths. A slightly difference in the variability of temperature over the remaining months could be reflected by the continuous solar radiation in a longer daylight period.

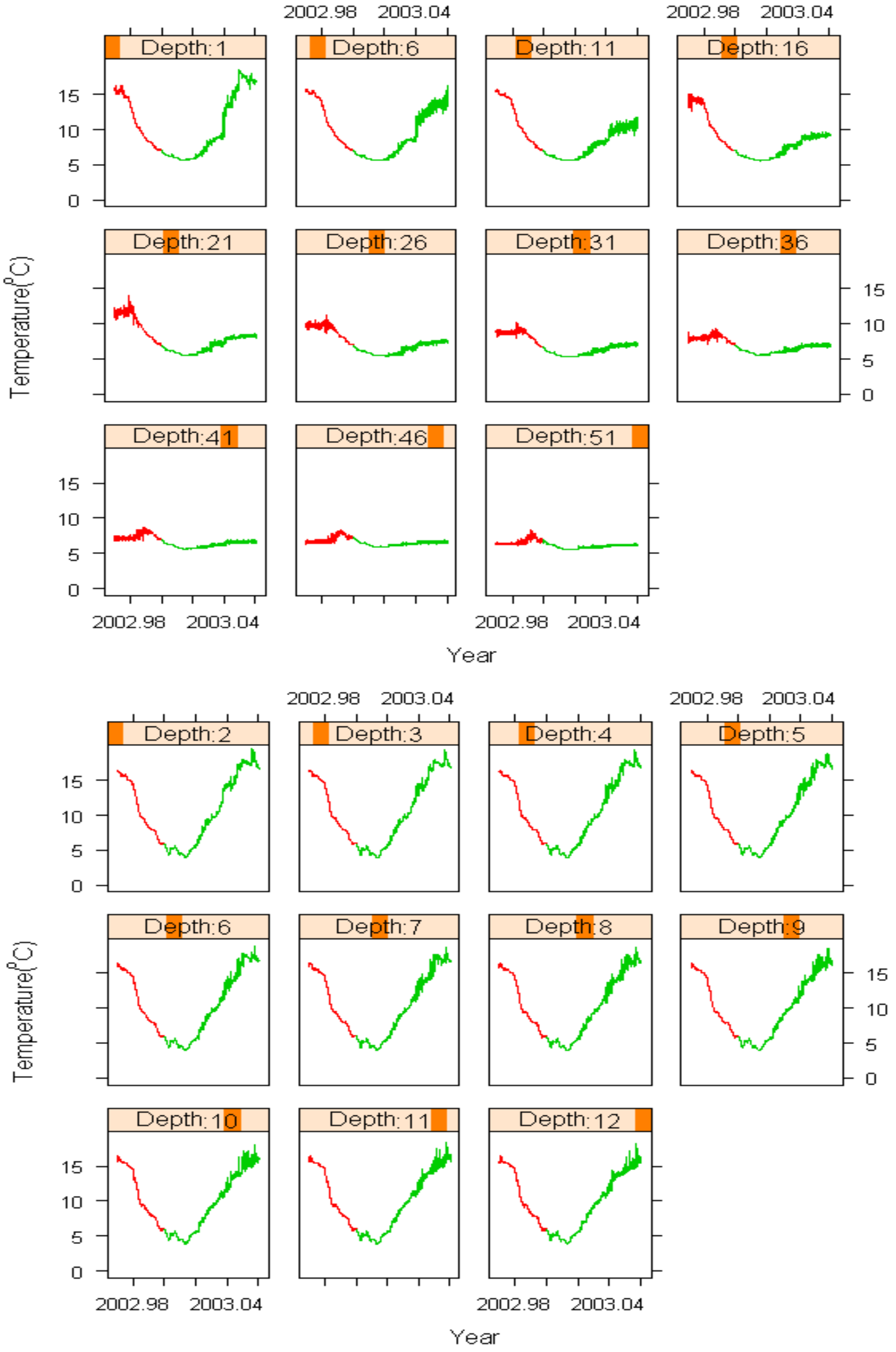

**Figure 3.3.** The 3-hourly temperature measurements over the year (1 Sept 2002 - 31 August 2003) with depths, in the north (top) and south (bottom) basins with the red and green curves represent the measurements in 2002 and 2003, respectively.

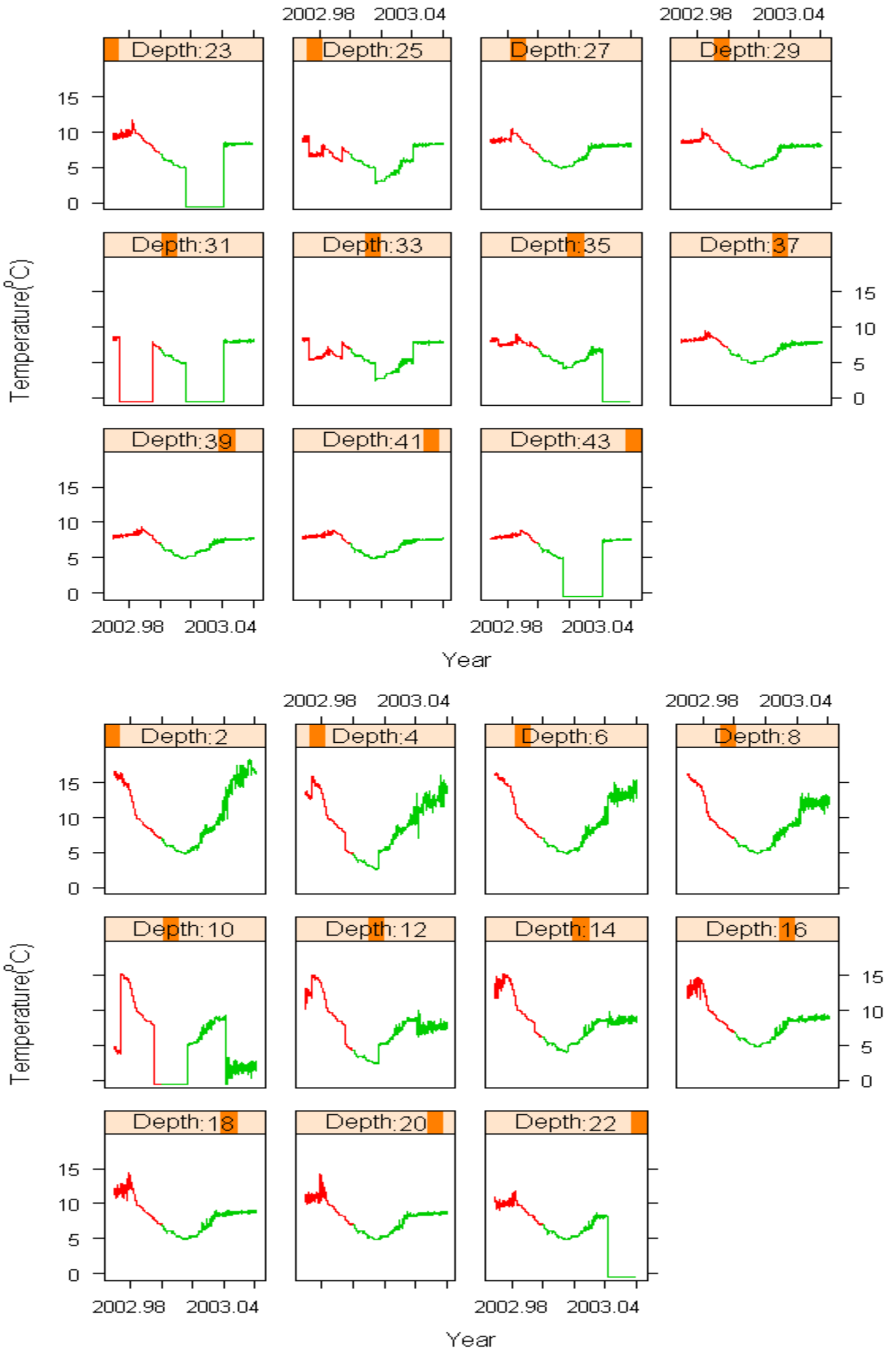

**Figure 3.4.** Series of 3-hourly temperature measurements (1 Sept 2002 - 31 August 2003) with depth, at lower chain (top) and upper chain (bottom) of Ross Point with the red and green curves represent the measurements in 2002 and 2003, respectively.

The temperature pattern over the year in the loch could be naturally described by a seasonal pattern. However, since the temperature measurements in the north and south basins are recorded in only one year period, the temperature patterns over the year are approximately quadratic for most of the depths, despite unclear patterns for the depths closed to the bottom of the water body in the north basin. Hence, the temperature measurements throughout the year at the deeper (north) and shallower (south) locations in Loch Lomond can be modelled, accordingly.

A greater variability in the temperature profile is exhibited over the warmer months in 2002 and 2003 in the north than the south basin. Such a feature may suggest the development of the thermocline in the north basin but this natural feature in lakes could not be identified graphically in the south basin. This is highly likely due to the fact that the thermocline is only developed in deep water over the warmer months and so, the south basin which is shallower, is not given much attention for further exploration and investigation of the thermocline.

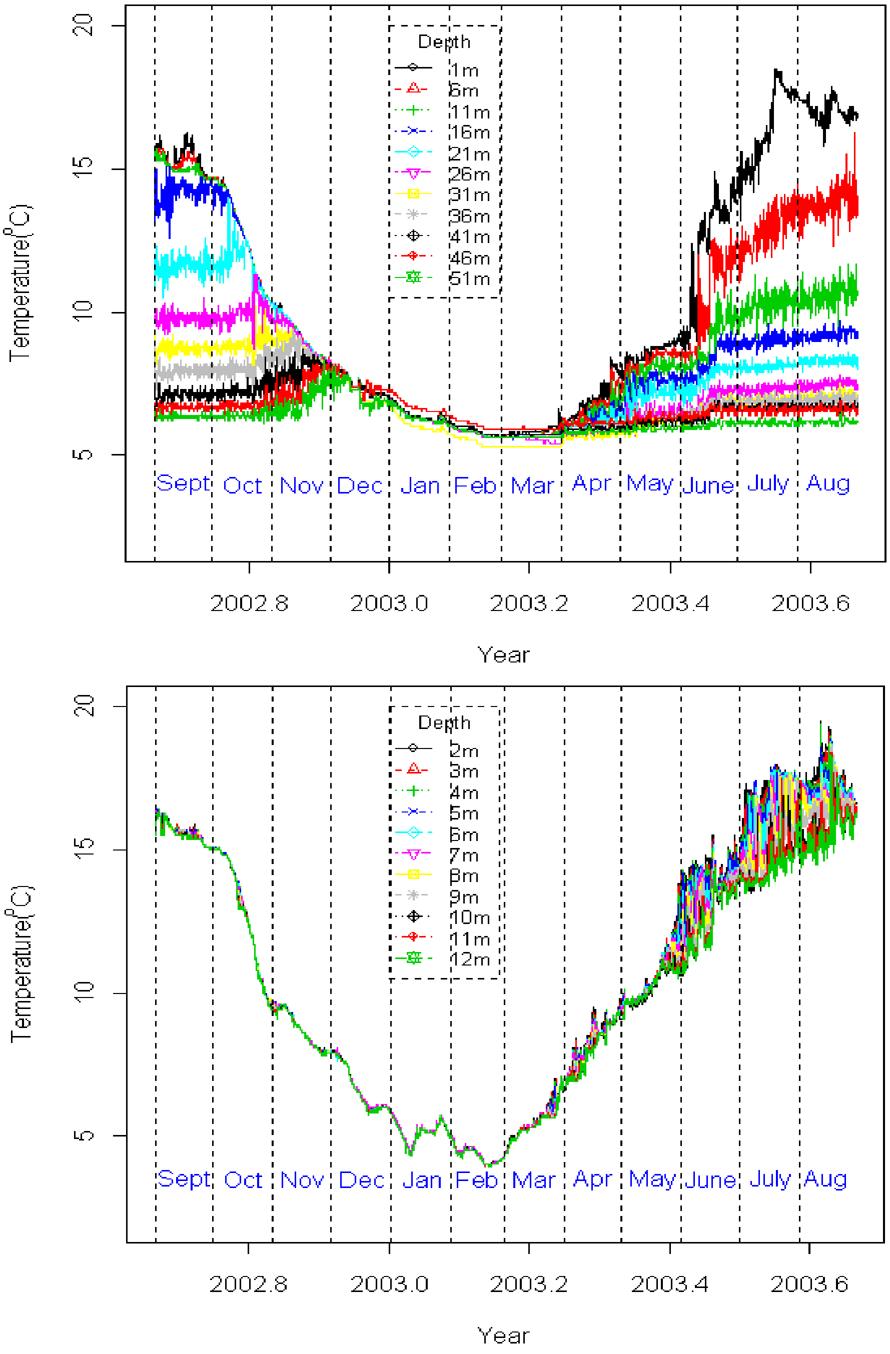

**Figure 3.5.** The 3-hourly temperature measurements at 11 different depths in the north (top) and south (bottom) basins from 1 September 2002 to 31 August 2003.

Table 3.1 shows the summary for temperature for the north and south basins. Since the possible number of measurements for each depth is 2920, 44 and 15 measurements are missing in the north and south basins, respectively. The 1.5% and 0.5% missing values with respect to the north and south basins represent a sufficiently small percentage of unobserved measurements that they may have little effect on the model and so, imputing the missing values is not required. The lowest and highest values of the sites are observed at Creinch, indicating that the shallow water body keeps and releases more heat in summer and winter, respectively, than the deep water body (north). As expected, the average temperature in the south is higher than the north basin.

| Site | Sample for each depth | Minimum | Maximum | Mean |
|---|---|---|---|---|
| North | 2876 | 5.30 | 18.50 | 7.80 |
| South | 2905 | 3.90 | 19.50 | 10.30 |

**Table 3.1.** Summary statistics for the temperature measurements in the north and south basins from 1 September 2002 to 31 August 2003.

**Correlation Structure**

The correlation structures of the residuals over the time period for each depth in the north and south basins are explored. Observations at different depths for a given time are assumed independent since there are only 11 observations for each time point. The temperature considered at a given depth features a diurnal pattern and so, a moving average of order 8 is used to remove the daily pattern.

A moving average is a smoothing technique that allow a clear view of the trend of a time series. It could be used to remove the periodic fluctuation and random noise variation within time series (Spyros et al., 1997).

The moving average (MA) of order $m$ is defined as $z_i$ in equation 3.1,

$$z_i = \sum_{j=i-(m-1)}^{i} \frac{y_j}{m}; j = m, m+1, \ldots, n \tag{3.1}$$

where $n$ is the number of observations, $y_j$ is the temperature at jth time point and $z_i$ is the ith moving average.

Since there are 8 sets of 3 hours measurements for each day in the north and south basins, the diurnal pattern is removed using $m = 8$ and hence, the first value of the MA(8) is defined as $z_1$ in equation 3.2,

$$z_1 = \sum_{j=1}^{8} \frac{y_j}{8} \tag{3.2}$$

Each of the temperature measurements could be denoted by a model for temperature over the time period with a quadratic trend and diurnal pattern at a given depth for each of the north and south basins (equation 3.3),

$$y_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \gamma \cos\left\{\frac{2\pi\left(t_i - \theta\right)}{p}\right\} + \epsilon_i; i = 1, 2, \ldots, n \tag{3.3}$$

where $y_i$ is the ith measurements, $t_i$ is the ith time, $p$ is the number of measurements in 24 hours and $n$ is the number of measurements in the time series. The use of moving average of order 8 on a series of temperature measurements (model

3.3) results in equation 3.4.

$$z_i = \beta_0 + \beta_1 t_i + \beta_1 t_i^2 + \epsilon_i; i = 8, 9, \ldots, n \tag{3.4}$$

A modified series which consist of the random noise and diurnal cycle components are extracted by subtracting the new value of the moving average (equation 3.4) from the actual measurements (equation 3.3) and is defined in equation 3.5.

$$y_i - z_i = \gamma \cos\left\{\frac{2\pi\left(t_i - \theta\right)}{8}\right\} + \epsilon_i \tag{3.5}$$

The above method may have implications for autocorrelation in the modified series and so, 9 values of $y_1, y_2, \ldots, y_9$ from a Normal distribution with zero mean and constant variance are generated. The moving average of order 8 is applied on this series resulting to two moving average values $z_1$ and $z_2$, and the difference between each of these moving average values and the actual values from the previous Normal distribution are computed. The correlation between the two differenced values are determined. The above process is repeated for 10000 times and the mean of the correlation for the modified series are determined. The results shows a low correlation of -0.10, indicating the appropriateness of this method to reduce the correlation in the modified time series.

Plots of autocorrelation and partial autocorrelation functions of the residuals after deseasonalizing are used to highlight any possible correlation structures for both basins. The 95% confidence intervals of the correlations for both functions are also displayed and the correlation structure is identified.

Figures 3.6 and 3.7 show the autocorrelation functions (acf) for the deseasonalised residuals for each depth in the north and mid basins. The acfs for both basins exhibit an exponential decreasing pattern, suggesting the appropriateness of an Autoregressive (AR) model for the correlation structure of the residual series. The evidence of correlation at lags one and two for each depth is apparent in the north. In the south basin, the significant correlation is highlighted at lags one, two and three. The low correlation at lags two and three in the north and south basins, respectively, indicate that the linear relationships of the deseasonalised residuals with respect to 6 and 9 hour time shifts are weak and so, they could be ignored. The pacf for the deseasonalised residuals at each depth for the north and south basins are displayed in Figures 3.8 and 3.9, respectively, showing a cut-off after lag 1. The acf suggests the appropriateness of AR(1) and AR(2) models for the north and south basins, respectively. The pacf for the deseasonalised residuals suggests similar AR(1) model for the north basin but contradicts the models of residuals for the mid and south basins resulting from the acf. Since the errors in real environmental time series are often and reasonably defined by the AR process, the AR(1) and AR(2) models for deseasonalised residuals for the north and south basins could be plausible for the error structures to be incorporated in the further modelling.

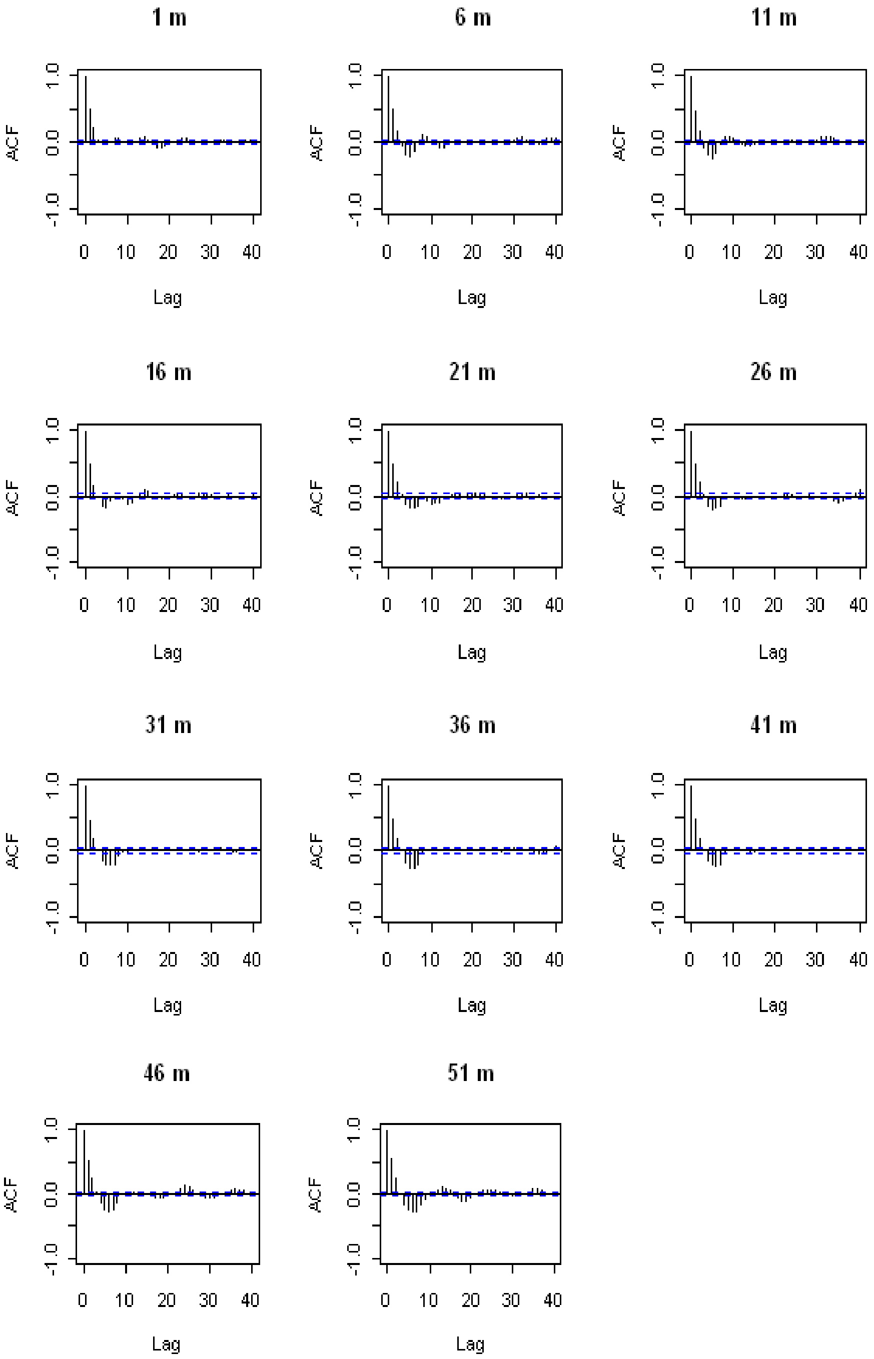

**Figure 3.6.** The acf of the deseasonalised residuals for 11 different depths in the north basin

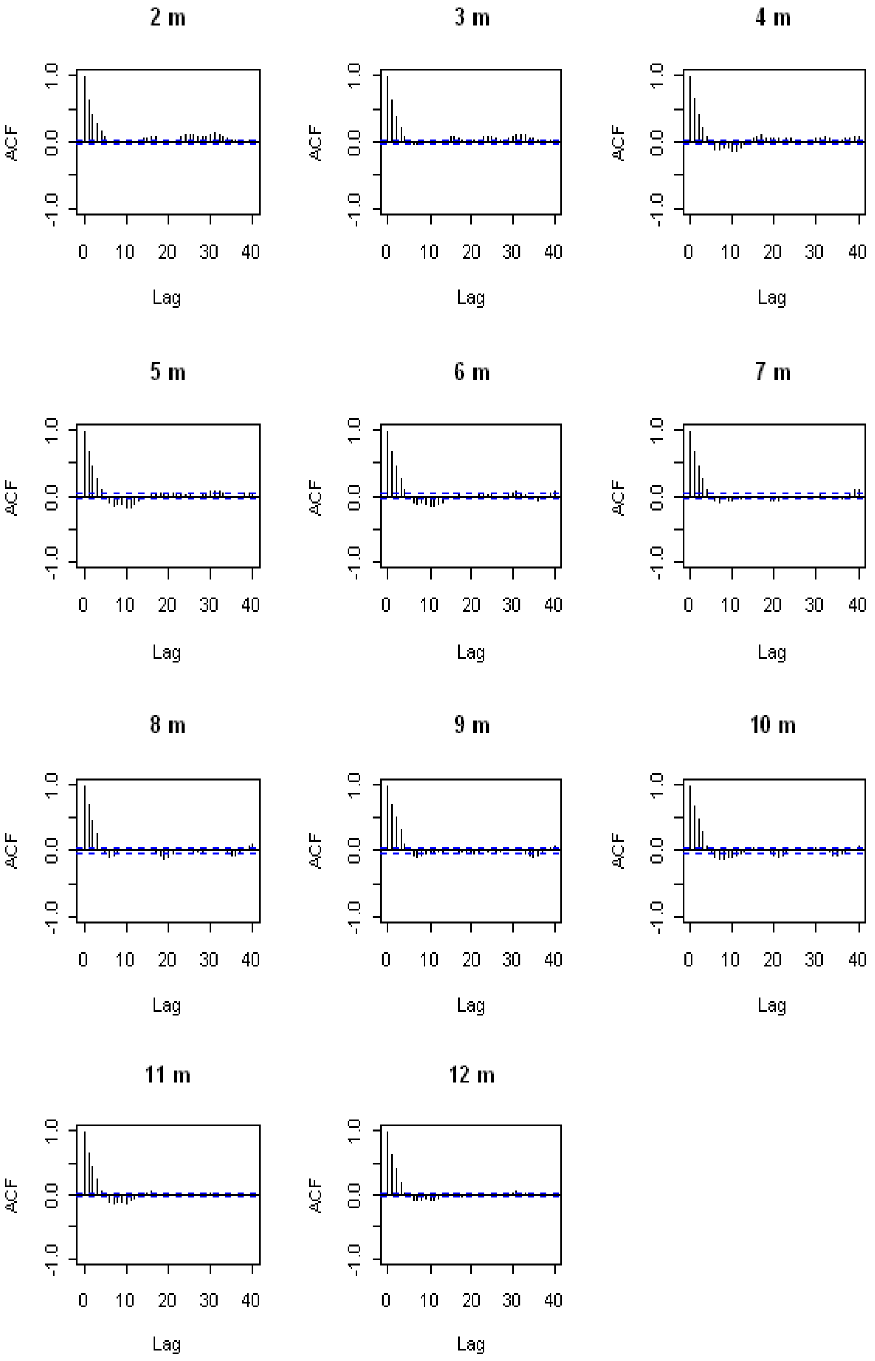

**Figure 3.7.** The acf of the deseasonalised residuals for 11 different depths in the south basin

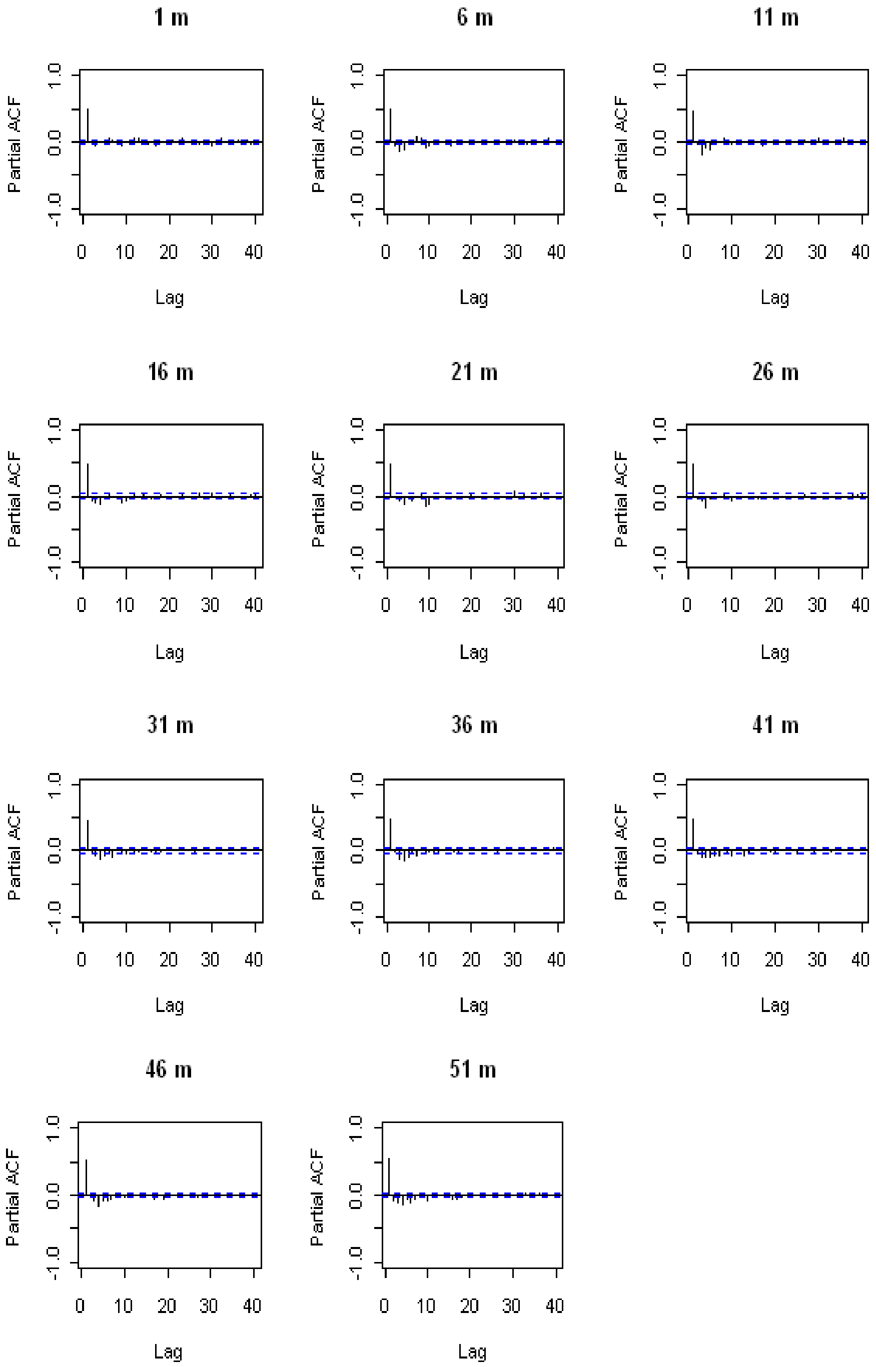

**Figure 3.8.** The pacf of the deseasonalised residuals for 11 different depths in the north basin

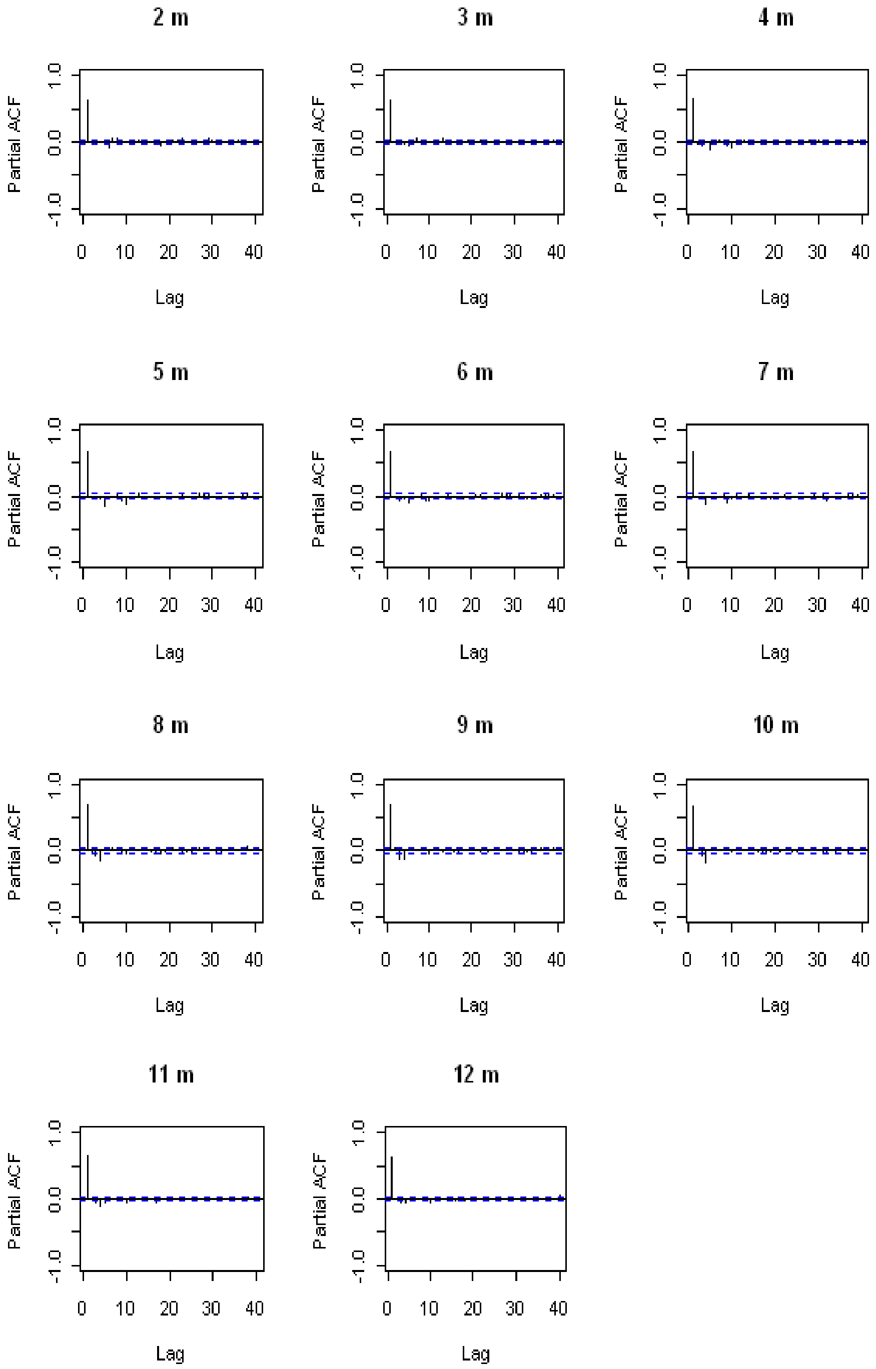

**Figure 3.9.** The pacf of the deseasonalised residuals for 11 different depths in the south basin

The apparent differences between the two basins suggest it is worth exploring the temperature measurements in the mid basin.

Table 3.2 shows the summary of temperature in the mid basin. The possible number of measurements for each depth from 17 April 2008 - 27 May 2009 is 9719, indicating 622 missing values. However, the incomplete data sets, with approximately 6.7% missing at each depth is fairly small and so, imputed values are not necessary for modelling the temperature over the time period. The minimum and maximum values are $1.5^oC$ and $19.8^oC$, and the mean temperature is $8.36^oC$. The mean temperature in the mid basin is lower than the south but slightly greater than the north basin as expected. However, the lowest temperature observed in the mid basin compared to much shallower (south) and deeper (north) locations of the loch is unexpected and is likely due to an error in the measurements.

| Sample for Each Depth | Minimum | Maximum | Mean |
|---|---|---|---|
| 9097 | 1.50 | 19.80 | 8.36 |

**Table 3.2.** Summary statistics for the temperature measurements in the mid basin from 17 April 2008 - 27 May 2009.

Figure 3.10 shows the temperature patterns over the year, with depths, in the mid basin. The variability of the temperature for each depth could be distinguished in the warmer months from April - October 2008 and March - May 2009. Conversely, the temperature for each depth in the colder months displays similar variability from November 2008 - February 2009. Since the mid basin is predominantly a deep water body, the thermocline development is likely to be identified in summer.

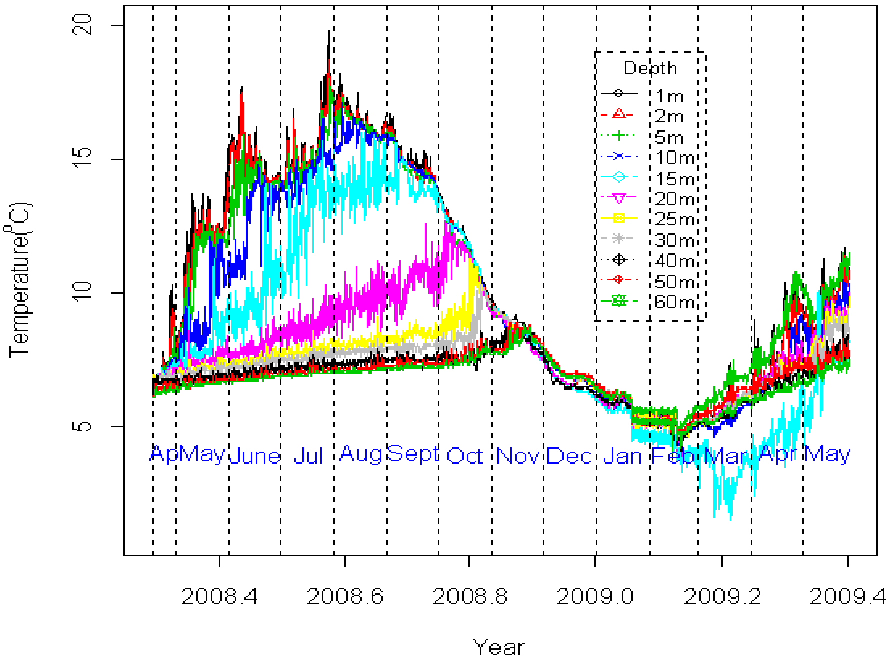

**Figure 3.10.** The 1-hourly temperature measurements at 11 different depths in the mid basin from 17 April 2008 - 27 May 2009.

The data from the mid basin is then explored for the occurrence of any correlation structure of the deseasonalised residuals. Firstly, the 1 hourly temperature measurements are aggregated to 3 hourly values to allow adequate comparison with the previous two basins and the moving average of order 8 is used to remove the diurnal cycle. The same lag of time between two adjacent measurements may allow consistent description of the correlation structure of the deseasonalised residuals for each basin.

Figures 3.11 and 3.12 show the acf and pacf for the deseasonalised residuals at each depth in the mid basin. The acf for each depth exhibits an exponential decay pattern, suggesting the appropriateness of an Autoregressive (AR) model for the correlation structure. The evidence of the correlation at lags one, two and three are noticeable, however, the low correlation at lag three indicates the weak linear relationships of the residuals separated by 9 hours time shift and so, they are not given much attention. The pacf, on the other hand, does not match the acf. However, the AR(2) model for the deseasonalised residuals could be a reasonable error structure as it provide more meaningful explanation on the relationships of the residuals, ecologically.

While the small percentage of missing values in the north and mid basins may not affect the fitted model of temperature over the year very much, with depths, such missing values may greatly affect the estimates of the thermocline since only 11 temperature measurements are recorded with depths at each time point. This is due to the fact that the missing values at particular depths may result in a large uncertainty in the estimates of the positions of the thermocline. Hence, complete data sets are required to achieve the second objective and the imputation of the missing values in the north and mid basins are carried out via appropriate approaches as follows.

For the north basin, the missing values are imputed by substituting the mean of the temperature measurements at the same time on the day before and after the missing observation.

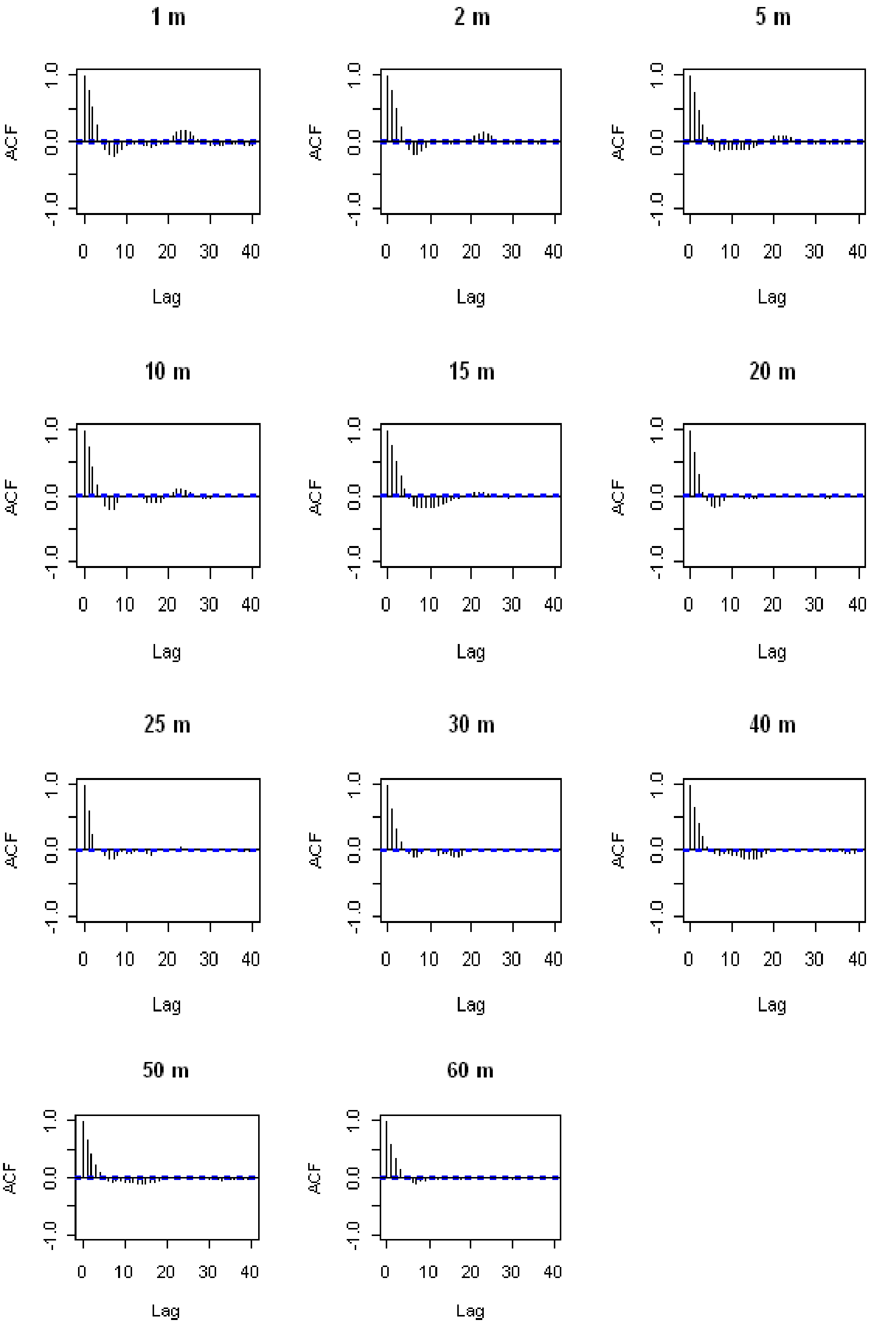

**Figure 3.11.** The acf of the deseasonalised residuals for 11 different depths in the mid basin

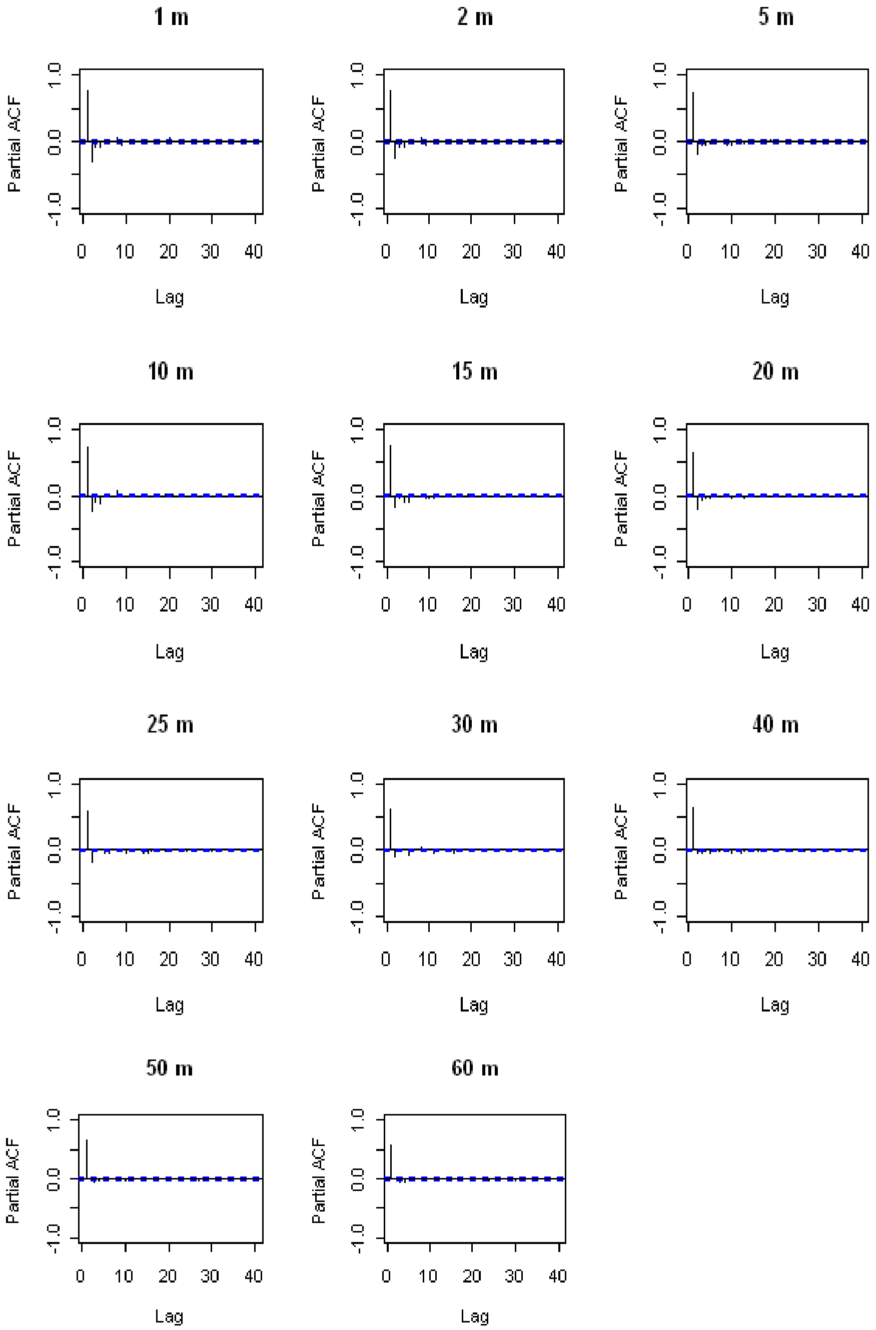

**Figure 3.12.** The pacf of the deseasonalised residuals for 11 different depths in the mid basin

For the mid basin, the number of missing data in summer and autumn are smaller than in winter. Since a small number of missing data occurs in summer and autumn, the same imputation technique as in the north basin is used whilst a different imputation techniques is applied to the winter. In winter, a harmonic model is fitted to the temperature measurements over the time period for each depth since a natural sinusoidal pattern is observed over the year from the exploratory plots. The amplitude of the harmonic patterns at each of 1 to 15 metres from the surface is larger than those below 15 metres and so, two harmonic models with different sizes of amplitude are fitted. Since the missing values occur in the mid winter, the observed temperature measurements in this season are divided into two parts, before and after the period of the missing data. The mean of the observed temperature measurements in winter is added to each of the fitted values within the period of missing data in mid winter. For the missing data at depths below 15 metres, however, the same imputation method as used in the north basin is carried out.

The random errors generated from a normal distribution with zero mean and constant variance are then added to each of the imputed values from the north and mid basins.

Scatter plots of depth against temperature at each of the time points grouped by month, are used to obtain an initial impression of the temperature profiles with depth, in the water column. A large change in temperature between two contiguous depths may indicate the thermocline.

Figures 3.13 and 3.14 depict scatter plots of the depth(m) against 3-hourly temperature($^oC$) measurements in each month, in the north and mid basins, respectively. The approximate constant temperature with depth is portrayed over the colder months in both basins, highlighting no particular feature in the water column.

The temperature profiles with depth in the north basin are constant from December 2002 to March 2003 and appear to follow a linear pattern, whilst a similar feature in the temperature profile is also shown in the mid basin from November 2008 to February 2009. Conversely, non constant temperature profiles with depth are shown in the remaining months of the year in both basins. The non constant temperature profiles with depth appear to highlight some features in the data measurements as described below. In particular, the quadratic and cubic patterns in the temperature profiles with depth in the warmer months may suggest the development of the thermocline.

In the north basin, the approximate cubic pattern of temperature with depth is shown from September - November 2002. The approximate quadratic pattern of the temperature with depth (Figure 3.13) is more apparent from April to August in the following year. The thermocline may appear at the positions of 11 and 26 metres below the water surface in the north basin.

In the mid basin (Figure 3.14), both linear and quadratic patterns are evident from April - May 2008 whilst a combination of quadratic and cubic patterns are shown in June 2008. The approximate cubic pattern seems plausible for the temperature pattern with depth from July to October 2008. It is likely that the thermocline is developed between 10 and 30 metres below the surface. However, no specific pattern is highlighted from March to May 2009. Such a pattern may suggest different features in the water column, however, the apparent temperature changes between 5 and 20m below the surface is likely due to an unexpected ecological disturbance.

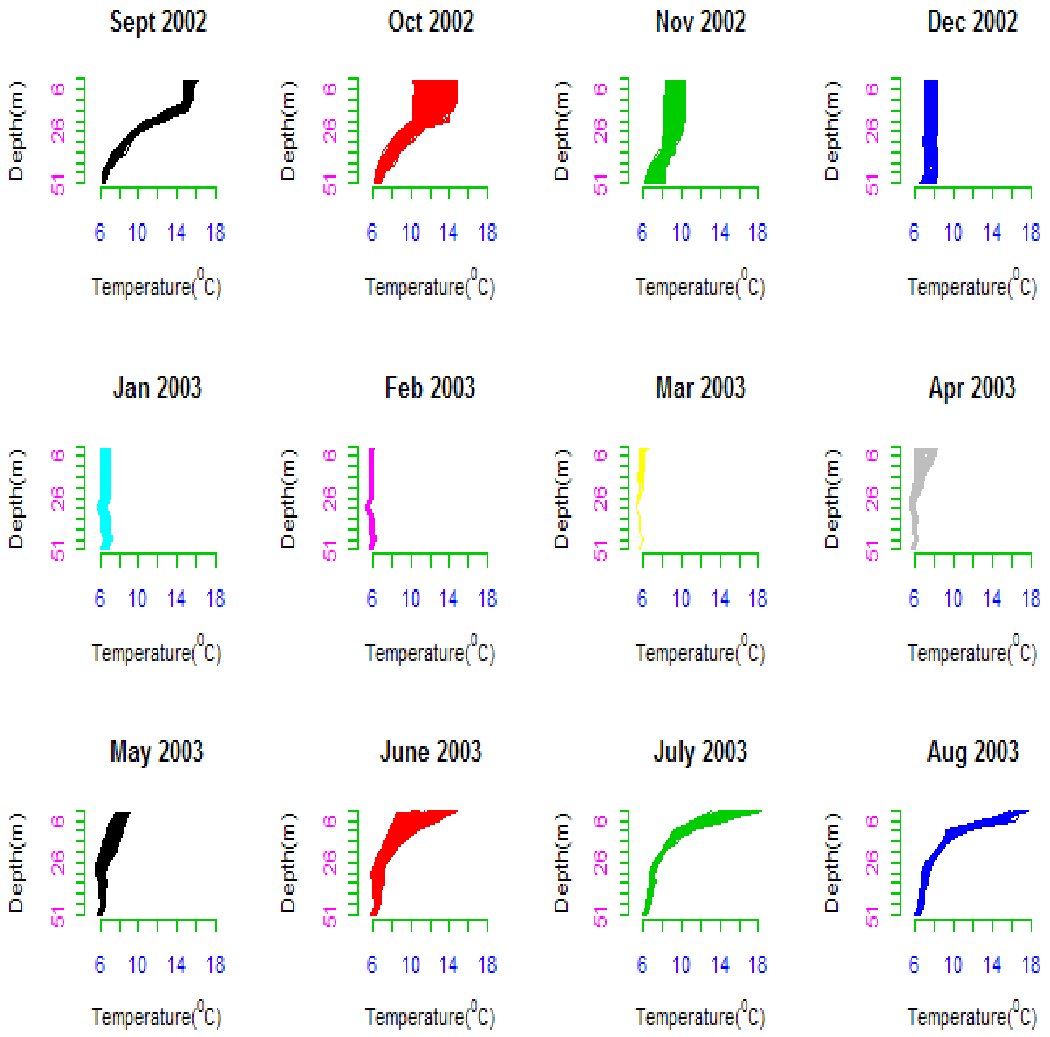

**Figure 3.13.** Scatter plots of depth(m) against temperature($^o$C) at each time point, grouped by month, for the north basin.

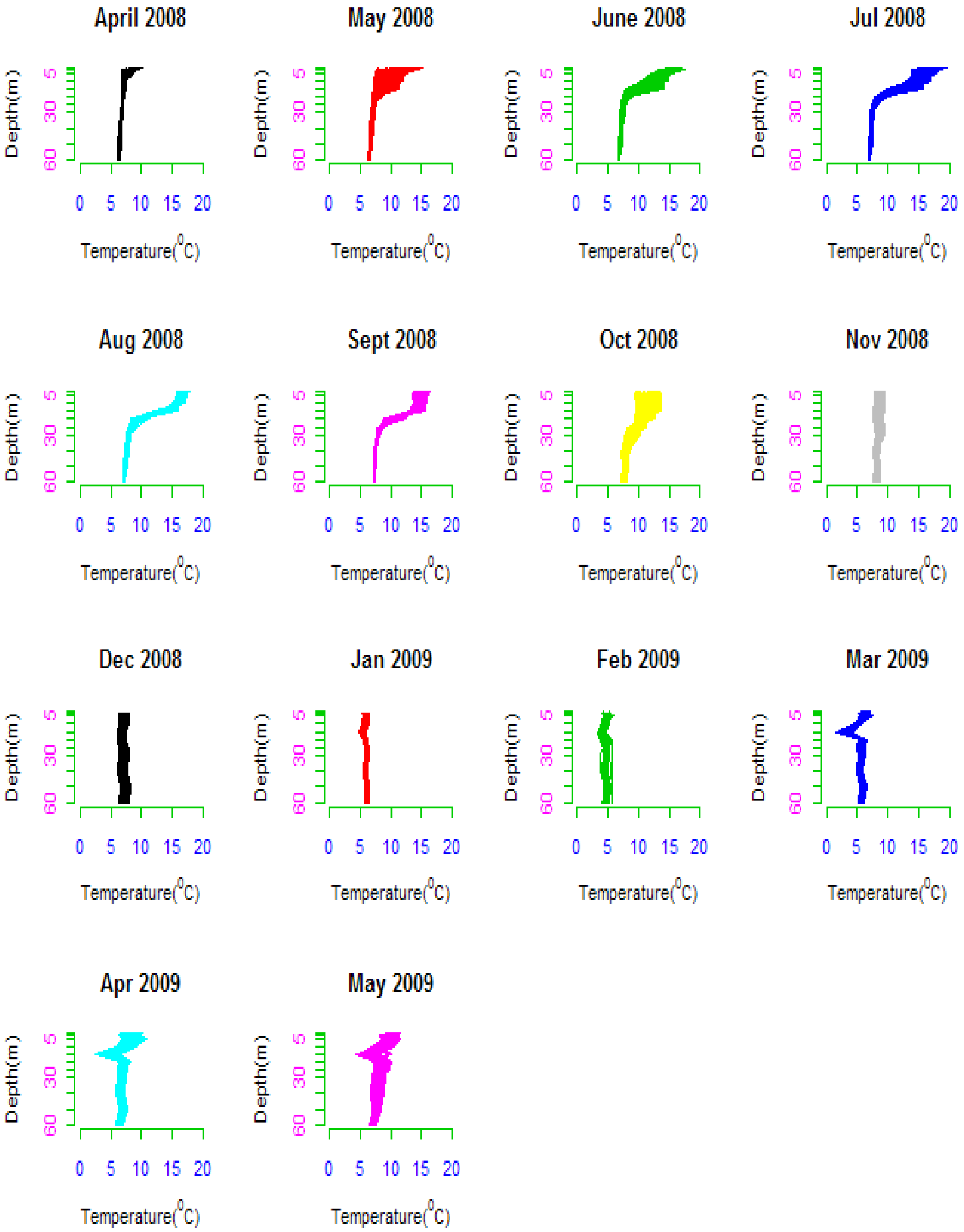

**Figure 3.14.** Scatter plots of depth(m) against temperature($^o$C) at each time point, grouped by month, for the mid basin.

Finally, a contour plot of temperature across depth and year, with a contour step of $1^oC$, is also displayed to show the entire temperature profile and the possibility of the formation of the thermocline. The contour plots of temperature for the north (top) and mid (bottom) basins are shown in Figure 3.15, with a contour step of $1^oC$, highlighting the homogeneous temperature in the water column between December 2002 and March 2003, and between December 2008 and February 2009, respectively, but greater changes are evident in the other months.

Areas of sharp temperature gradients, which are shown by several contours close to each other, may indicate the position of the thermocline. In the north basin, this feature is observed from September to November 2002 at about 16 - 26 metres from the surface and at the shallower levels of less than 16 metres down the depth profile from June - August 2003. In the mid basin, such a characteristic is exhibited at approximately 25 metres down the depth profile from May - October 2008, and between 10 and 20 metres down the depth profile from March to May 2009. However, the sharp temperature gradient in 2009 may suggest a different characteristic in the water body due to a similar degree of temperature as highlighted by indistinct colours.

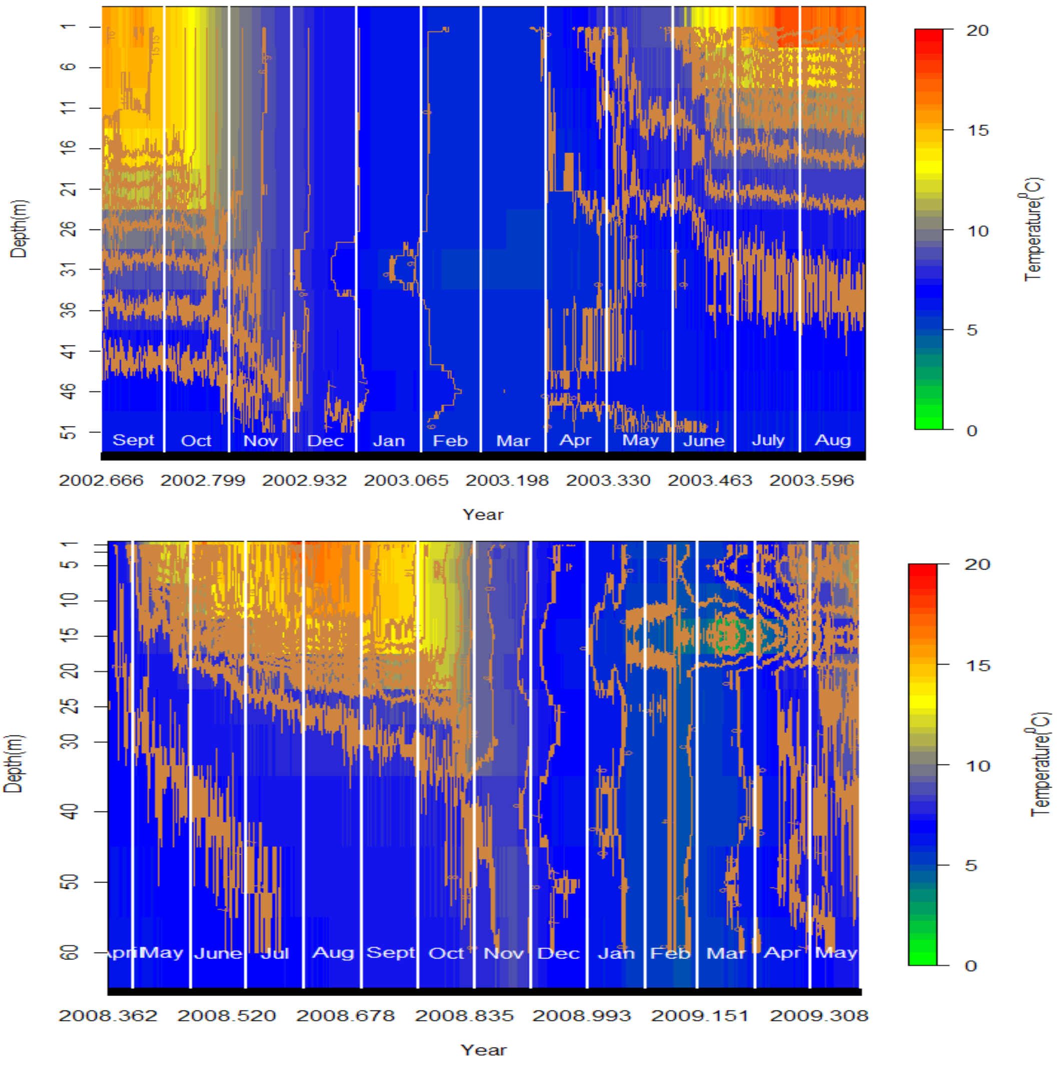

**Figure 3.15.** Contour plot of temperature across depth and year for the north (top) and mid (bottom) basins.

## 3.4 Methods

The first question of interest is answered by fitting a linear mixed-effects model to the temperature over the year, with depth, in the north and south basins (September 2002 - August 2003) and mid basin (17 April 2008 - 27 May 2009). Model comparison is carried out to determine the appropriate model for each basin.

For the second question, three approaches are used to investigate the possible position of the thermocline development in the north and mid basins from 1 Sept 2002 - 31 August 2003 and 17 April 2008 to 27 May 2009, respectively; the maximum relative rate of change of the temperature curve, changepoint regression and estimation of the derivative of a smooth temperature curve.

The details of the above approaches are as follows:

### 3.4.1 Linear Mixed-Effects Model

A general representation of a linear mixed-effects model is defined in equation 3.6.

$$y_{ij} = X_{ij}\beta + Z_{ij}b_j + \varepsilon_{ij}; \begin{cases} i = 1, 2, \ldots, \text{n time points} \\ j = 1, 2, \ldots, \text{m levels of a grouping factor} \end{cases} \tag{3.6}$$

$$b_j \sim N(0, D), \varepsilon_{ij} \sim N(0, \Sigma)$$

where $y_{ij}$ is a matrix of response at the $i$th time point and $j$th level, $X_{ij}$ is $(X_{ij1}, X_{ij2}, \ldots X_{ijp})$ is the design matrix for $p$ fixed effects at $i$th time point and $j$th level of the grouping factor, $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})^T$ denotes the vector corresponding to the $(p-1)$ fixed effects, $Z_{ij} = (Z_{ij1}, Z_{ij2}, \ldots, Z_{ijq})$ is the design matrix for $q$ random effects at $i$th time point and $j$th level of the grouping factor, $b_j = (b_{j1}, b_{j2}, \ldots, b_{jq})$ is the vector of random effects at $j$th level and $\varepsilon_{ij} = (\varepsilon_{1j}, \varepsilon_{2j}, \ldots, \varepsilon_{nj})^T$ is the vector of random error at $i$th time point and $j$th level

Now having defined the linear mixed-effects model, the appropriate explanatory variables are determined prior to fitting the mixed model. Plots of temperature over the time period at different depths (Figure 3.3) indicate that a model of temperature over the year with a quadratic pattern (equation 3.7) could be appropriate at most of the depths $j$ in the north and south basins. The plot of temperature across the year for each depth in the mid basin (Figure 3.10), however, indicates the adequacy of a cubic pattern to be incorporated in the model of temperature over the year (equation 3.8),

$$y_i = \beta_o + \beta_1 t_i + \beta_2 t_i^2 + \epsilon_i \tag{3.7}$$

$$y_i = \beta_o + \beta_1 t_i + \beta_2 t_i^2 + \beta_3 t_i^3 + \epsilon_i \tag{3.8}$$

where $y_i$ is the temperature at time $t_i$, $\beta_o$, $\beta_1$, $\beta_2$ and $\beta_3$ are fixed effects and $\epsilon_i$ are random errors which follow a particular correlation structure. Equations 3.7

and 3.8 allow a separate temperature-time model for each of the depths. The incorporation of the temperature-time model for all depths results in fitting the linear mixed effects model. Hence, the linear mixed effects Model 3.9 is fitted for the north and south basins whilst Model 3.10 is fitted to the temperature in the mid basin,

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + b_{0j} + b_{1j} t_{ij} + b_{2j} t_{ij}^2 + \epsilon_{ij} \tag{3.9}$$

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 t_{ij}^3 + b_{0j} + b_{1j} t_{ij} + b_{2j} t_{ij}^2 + b_{3j} t_{ij}^3 + \epsilon_{ij} \tag{3.10}$$

where $i$ is the time point, $j$ is the level of depth, $y_{ij}$ and $t_{ij}$ denotes vector of temperature measurements and time at $i$th time point and $j$th level, respectively, $\beta_o, \beta_1$, $\beta_2$ and $\beta_3$ are fixed effects, $b_{oj}, b_{1j}$, $b_{2j}$ and $b_{3j}$ are random effects for depth $j$ and $\epsilon_{ij}$ are random errors at $i$th time point and $j$th level. The reason for incorporating all the random effects $b_{oj}$, $b_{1j}$, $b_{2j}$ and $b_{3j}$ corresponding to the fixed effects $\beta_o$, $\beta_1$, $\beta_2$ and $\beta_3$ is due to the fact that the coefficients in the model may change with depth.

The following are the details of the linear mixed effects model as explained by Pinheiro and Bates (2000). The random effect, $b_j$ and the random error, $\varepsilon_{ij}$ are assumed to be identically and normally distributed with zero mean and variance-covariance D and $\Sigma$, respectively. In the simple case, the variance-covariance of the random error at any level $j$, $\Sigma_j$ is assumed to be $\sigma^2 I$ as follows

$$\Sigma_j = \sigma^2 \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & 1 & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

and the variance-covariance matrix of the random effect D at all levels j is as follows.

$$D = \begin{bmatrix} d_{11}^2 & d_{12} & \dots & d_{1q} \\ d_{21} & d_{22}^2 & \dots & d_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ d_{q1} & d_{q2} & \dots & d_{qq}^2 \end{bmatrix}$$

The likelihood maximised of the parameters $\beta$, D and $\Sigma$ is formed from the marginal distribution of $y_{ij}$, where $y_{ij}$ is identically and normally distributed with mean $X_{ij}\beta$ and variance $V_j$, $y_{ij} \sim N(X_{ij}\beta, V_j)$ where $V_j$ is a covariance matrix and $y_{ij}$ has a marginal Normal distribution with mean $X_{ij}\beta$ and variance $V_j$. The estimate of $V_j$ is defined by $\hat{V}_j = Z_{ij}\hat{D}_j Z_{ij}^T + \hat{\Sigma}_{ij}$, which incorporates the variance from random errors and random effects.

The log likelihood of the marginal distribution of $y_{ij}$ is defined in equation 3.11,

$$l_i = -\frac{1}{2}(y_{ij} - X_{ij}\beta)^T V_j^{-1}(y_{ij} - X_{ij}\beta) - \frac{1}{2}\log|V_j| \tag{3.11}$$

where $|V_j|$ denotes the determinant of $V_j$. In fact, given the variance-covariance matrices D and $\Sigma$, the maximum likelihood estimator of $\beta$ can be computed by using the best linear unbiased estimator (BLUE) as defined in equation 3.12.

$$\hat{\beta} = (\sum_{i=1}^{m}(X_{ij}^T V_j^{-1} X_{ij})^{-1})(\sum_{i=1}^{m}(X_{ij}^T V_j^{-1} y_{ij})) \tag{3.12}$$

The components of the random effects, $b_i$ are needed to predict a future response and can be predicted using the best linear unbiased prediction (BLUP), as defined in equation 3.13.

$$\hat{b_{ij}} = \hat{D} Z_{ij}^T \hat{V}^{-1}(y_{ij} - X_{ij}\beta) \tag{3.13}$$

Pinheiro and Bates (2000) show that the profile log likelihood function of the variance-covariance matrices D and $\Sigma$ is produced by substituting the BLUE of $\beta$ from equation 3.12 into the log likelihood of the marginal distribution (equation 3.11) and the variance components can be obtained by maximizing this profile log likelihood. A restricted maximum likelihood estimator (REML) method is preferred as the maximum likelihood method tends to underestimate the variance components (Pinheiro and Bates, 2000).

Optimization of the profile log likelihood of a linear mixed-effects model is typically done by the use of EM iterations or via Newton-Raphson iterations (Laird and Ware (1982); Lindstrom and Bates (1988); Longford (1993)). The EM algorithm (Dempster et al., 1977) is a popular iterative algorithm for likelihood estimation in the presence of incomplete data, however, the Newton-Raphson algorithm (Thisted, 1988) is one of the most widely used optimization procedure.

The previous exploration of the deseasonalised residuals suggest particular error structures for the north, mid and south basins. In particular, the AR(1) correlation structure for $\epsilon_{ij}$ is incorporated in the model for the north whilst AR(2) error structure is used for the mid and south basins. The `lme` function from `nlme` library in R (Pinheiro and Bates, 2000), is used for fitting the linear mixed-effects models 3.9 and 3.10.

Linear mixed-effects models are used here since they have widely been used in environmental studies when repeated measurements are made on groups of related statistical units. Lai and Helser (2004) compared Atlantic scallops by using simple linear regression (group factor is ignored) and linear mixed-effects model (group factor is incorporated as fixed-effect) approaches in the field of biology and proposed that the linear mixed-effects model was an effective way to analyze and compare weight-length relationship of scallops between groups. Meng et al. (2007) used linear fixed-effects and linear mixed-effects models in their forestry study to fit the relationship between either forest biomass or volume of trees and normalized difference vegetation index (NDVIsa), in which NDVISa was used as the predictor which implies the area of trees, whilst biomass or volume was used as the response and discovered that the linear mixed-effects model was the best modelling approach.

**Model Selection**

Two approaches for model selection are considered in this study.

The first approach is based on model selection tools; Akaike Information Criteria

(AIC) and Bayesian Information Criteria (BIC) as defined in equations 3.14 and 3.15, respectively. Both criteria consist of two terms, which consider the measure of fit and the complexity of a model,

$$AIC = -2L(\theta) + 2p \tag{3.14}$$

$$BIC = -2L(\theta) + 2p\log(N^{'}) \tag{3.15}$$

where $L(\theta)$ is either likelihood, ML or REML, $p$ is the number of parameters, $N$ is the number of observations. In ML, $N^{'} = N$, but for REML, $N^{'} = N - p$ (Alain et al., 2009). The log likelihood function from REML or ML can be used to measure the fit, whilst the number of parameters indicate the complexity of the model. Unlike AIC, the number of observations is also taken into account in BIC.

The second approach for assessing the model is a formal statistical analysis on the comparison of two nested models. A general approach for comparing two models fit by maximum likelihood is via a likelihood ratio test (Lehmann, 1986), however, formal comparison can also be carried out on models fit by REML. The nested models can be compared if they are fitted by REML and the fixed-effects for both models are the same (Pinheiro and Bates, 2000). The likelihood ratio test (LRT) statistic is defined as,

$$2\log\left(\frac{L_2}{L_1}\right) = 2\left[\log(L_2) - \log(L_1)\right]$$

where $L_1$ and $L_2$ are the likelihood of restricted model (simple model) and general model (full model), respectively, and the value of the above equation is positive. If $n_k$ is the number of parameters to be estimated in model $k$, then the asymptotic distribution of LRT statistic under the null hypothesis that the restricted model is more appropriate is $\chi^2$ distributed with $n_2 - n_1$ degrees of freedom (Pinheiro and Bates, 2000).

Since a model comparison requires two nested models, the second model of temperature over the time period, with depths, for each basin is fitted with a simple correlation structure than that used in the first model. In particular, the use of AR(1) for the north and AR(2) for the mid and south basins for the correlation structures in the model of temperature over the year, with depths, results in defining no correlation structure for the north but AR(1) for the mid and south basins in the second mixed-effects model.

### 3.4.2 Maximum Relative Rate of Change

By referring to the recent thermocline definition among limnologists, it is of interest to determine the rate of change in the temperature with depth. A meaningful and easily interpreted normalization approach would be through the ratio of rate of change in temperature between three adjacent depths and total rate of change in temperature across all depths. The maximum of the ratio is used as the basis of the results.

Let $x_{ij}$ and $y_{ij}$ denote depth and temperature at time $i$ from 1 to 2920 and position $j$ of 2 to 10, respectively. The absolute rate of change (RC) in temperature $y_{ij}$ at a particular depth $x_{ij}$ is the absolute difference in slopes between the temperature at such a depth and the depths immediately above and below, as defined in equation 3.16

$$RC_j = \left| \left( \frac{y_{ij+1} - y_{ij}}{x_{ij+1} - x_{ij}} \right) - \left( \frac{y_{ij} - y_{ij-1}}{x_{ij} - x_{ij-1}} \right) \right| \begin{cases} i = 1, 2, \ldots, 2920, \\ j = 2, 3, \ldots, 10 \end{cases} \tag{3.16}$$

where $n$ is the number of time points over the year.

There are 9 values of the RC in temperature, produced at each of the time points. Figure 3.16 highlights one of the RC at depth of 6m from the water surface.

The ratio of the absolute rate of change at a given depth $j$ and the total absolute rate of change, denoted by the relative rate of change (RRC) as defined in equation 3.17 is used to normalize the value.

$$RRC_j = \frac{RC_j}{\sum_{j=2}^{10} RC_j}; j = 2, \ldots, 10 \tag{3.17}$$

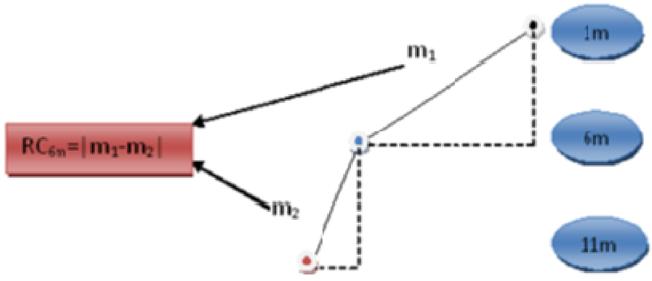

**Figure 3.16.** The absolute rate of change in temperature at depth of 6m

The RRC for each time gives a value, ranging from 0 to 1. The depth corresponding to the highest value of RRC may informally indicate the position of a thermocline.

### 3.4.3 Changepoint Regression

Changepoint regression can be used in situations where the regression slope is not constant but could change rapidly at a given point (see for example, Quandt (1958), Hudson (1966), Krisnaiah and Miao (1988), and Julious (2001)). A changepoint may highlight the position on a curve of temperature with depth for a given time point where a rapid change of temperature has occurred. Hence, such a point may represent the thermocline depth.

Estimation of the parameters in the model is straightforward if the position of the changepoint is known, however, the changepoint must be estimated if the position is unknown and so, a numerical optimization is required to estimate the parameters in the model (Julious, 2001).

The determination of the changepoint on the curve of temperature with depth for a given time point is adapted from (Julious, 2001) as follows:

For any interval of depths $(X_o, X_1)$, two lines are defined by equation 3.18:

$$f(x_i) = \begin{cases} f_1(x_i; \beta_1); X_o \leq x_i \leq \tau, \\ f_2(x_i; \beta_2); \tau \leq x_i \leq X_1 \end{cases} \tag{3.18}$$

where $f(x_i)$ is the temperature at depth $x_i$ where $f_1(x_i;\beta_1) = f_1(x_i;\beta_2)$ at $\tau$. For a simple two-line regression this is equivalent to equation 3.19:

$$f(x_i) = \begin{cases} \alpha_1 + \beta_1 x_i; X_o \leq x_i \leq \tau, \\ \alpha_2 + \beta_2 x_i; \tau \leq x_i \leq X_1 \end{cases} \tag{3.19}$$

where the parameters in the two models are constrained by $\alpha_1 + \beta_1 x_i = \alpha_2 + \beta_2 x_i$ at $\tau$. The parameters for each half of the two models can be estimated from equation 3.20.

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \beta_1^* \\ \beta_2^* \end{pmatrix} - \frac{s}{t} C^{-1} q \tag{3.20}$$

where,

$$\begin{aligned} \beta_1^* &= (X_o' X_o)^{-1} X_o' Y_0 \\ \beta_2^* &= (X_1' X_1)^{-1} X_1' Y_1 \\ s &= (\beta_1^{*'}, \beta_2^{*'}) q \\ t &= q' C^{-1} q \end{aligned}$$

$$\begin{aligned} C^{-1} &= \begin{pmatrix} (X_o' X_o)^{-1} & 0 \\ 0 & (X_1' X_1)^{-1} \end{pmatrix} \\ q &= (1, \tau, -1, -\tau)' \end{aligned}$$

$\hat{\beta}_1$ and $\hat{\beta}_2$ are the unconstrained estimates of the parameters from the two linear models in the domains of $(X_o, \tau)$ and $(\tau, X_1)$, respectively. However, the problem is not linear if the changepoint is unknown and so, the parameters can be estimated via numerical optimization (Julious, 2001), which is simplified in equation 3.20 with the following considerations (Hudson, 1966).

For a given time point, two unconstrained linear models of temperature are fitted on a series of depths $x_1, x_2, \ldots, x_t$ and $x_{t+1}, x_{t+2}, \ldots, x_T$, respectively.

1. If the two lines from the unconstrained models join between the adjacent depths $x_t$ and $x_{t+1}$, then the residual sum of squares from this model is less than any constrained model for these two depths that is forced to meet between $x_t$ and $x_{t+1}$.

2. Conversely, if the two lines do not join between the adjacent depths $x_t$ and $x_{t+1}$, then the constrained model with the smallest residual sum of squares will have a changepoint at either $x_t$ or $x_{t+1}$

3. The linear model that is constrained to meet at a particular point will not reduce the residual sum of squares.

The algorithm to estimate all the parameters in the model is derived by Julious (2001) and simplified as follows,

- For a given time point, all the unconstrained two-line models of temperature are fitted on depths $(X_1, X_t)$ and $(X_{t+1}, X_n)$, respectively, where $t = 2, 3, \ldots, 9$.

- The two lines from the unconstrained models that meet within adjacent depths $X_t$ and $X_{t+1}$ are recoded as constrained model.
- If the residual sum of squares from the best constrained model $RSS_c$ is smaller than the best unconstrained model $RSS_{uc}$, the algorithm stops and the parameters in such a constrained model are chosen.
- However, if $RSS_c \geq RSS_{uc}$, than the best fitting unconstrained model is constrained by forcing the two lines to meet at each $X_t$ and $X_{t+1}$ using equation 3.20 and the best fitted model with the lower $RSS$ is added to the previous recoded constrained models.
- If the new $RSS_c$ is smaller than the $RSS_{uc}$ from the best unconstrained model, the algorithm stops and the parameters from the new constrained model are chosen.
- However, if new $RSS_c \geq RSS_{uc}$, either no changepoint or more than one changepoint is assumed to have occurred.

### 3.4.4 Derivative of a Smooth Curve

By referring to the mathematical definition of the thermocline, an inflection point on a smooth curve of temperature with depth can be determined from the second derivative of a temperature function with respect to depth. The inflection point can be mathematically defined as a point at which the second derivative or the curvature of the smooth curve is zero. Such an inflection point is a position on a curve at which the curvature changes sign, meaning that the curve changes from being concave upwards (positive curvature) to concave downwards (negative curvature), or vice versa.

Let $Y$ and $X$ denote the temperature and depth, respectively. The thermocline depth $x_o$ is defined as the depth where:

$$\left.\frac{d^2Y}{dX^2}\right|_{x_o} = 0$$

A nonparametric regression model as defined in equation 3.21, is used to estimate a smooth curve for each temperature profile with depth, at each of the time points.

$$Y_i = f(X_i) + \varepsilon_i \tag{3.21}$$

where the properties of the above model has been discussed in the early chapter.

The `smooth.lf` function in `locfit` library in R (Loader, 1999) is used to fit the smooth curve of temperature profiles with depth, where the Gaussian kernel density function is used as a weight function. The estimates of the derivative for each of the local points are determined from the above package. The derivative of a smooth function is widely discussed and used. The bandwidth is chosen to be 70% of the local nearest neighbourhood for each local point. The above percentage is used since it could sufficiently provide an appropriate smooth curve of temperature with depth for each time point since about 8 out of 11 measurements are taken to estimate the derivatives at each of the depths.

A cubic trend seems to be appropriate to represent the temperature profiles with depth in a number of the warmer months. Therefore, a local polynomial regression of order 3 is used. The local/evaluation points used, ranging from the shallowest to the deepest level of the depths, an interval of 0.1 between two consecutive evaluation points. The third order local polynomial estimator (equation 3.22) is minimized.

$$min_{\beta_j} \sum_{i=1}^{n} \Big(Y_i - \sum_{j=o}^{3} \beta_j (X_i - x)^j\Big)^2 w\Big(\frac{X_i - x}{h}\Big) \tag{3.22}$$

Figure 3.17 (top) shows an example of a temperature profile with depth in one of the 3 hourly temperature measurements for the north basin in September 2002. Obviously, the pattern of temperature profile with depth may subjectively be represented by a cubic pattern as the figure is rotated in -90$^o$ about the y-axis.

The appropriate inflection point is chosen subject to the following criterion.

Firstly, such an inflection point must be somewhere within the metalimnion as subjectively exhibited from the exploratory analysis where the temperature changes abruptly with a small change in depth. This zone is known to be a possible position for the development of the thermocline (Victor and Robin, 2005). Since the local regression model is fitted based on a series of local/evaluation points, it is more likely that the second derivative corresponding to the inflection points is around zero and thus, such a feature can be mathematically defined as follows:

$$\left.\frac{d^2Y}{dX^2}\right|_{x_o} \approx 0$$

where $x_o$ is the depth corresponding to the position of thermocline.

Secondly, where there are several estimated inflection points, an appropriate point is chosen subject to a minimum value of the first derivative as such a point may highlight the position of no change in temperature with depths.

Figure 3.17 shows some illustrations of the first (center) and second (bottom) derivatives of a temperature curve with depth (top). Both derivatives suggest the position of a thermocline is likely to be at approximately 20m below the water surface.

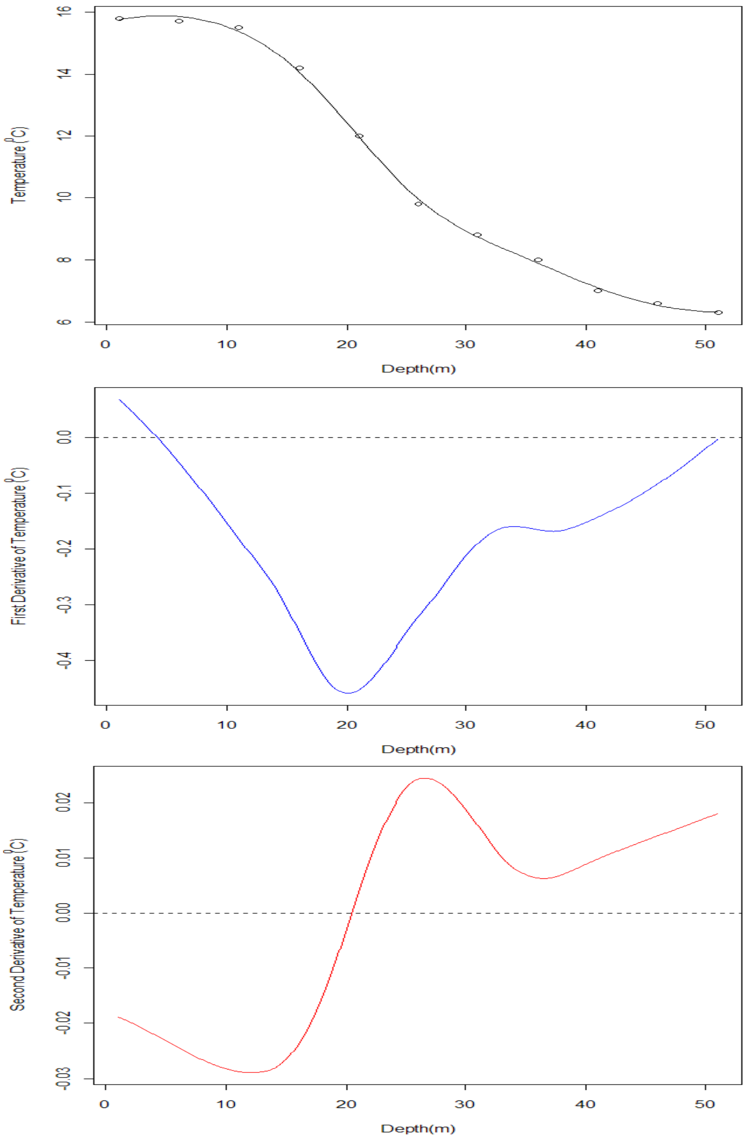

**Figure 3.17.** A smooth curve of temperature with depth (top), estimate of the first (center) and second (bottom) derivatives at one of the time points in summer month, in the north basin

## 3.5 Results

The fitted model for temperature over the year for each depth in the north, mid and south basins are defined and its features are explained in this section. Additionally, the estimates of the thermocline in the deeper water (north and mid basins) from different approaches are presented. The details of the results are as follows.

### 3.5.1 Temperature Variability Over The Year, With Depths

The results of the fixed effects from the mixed-effects Models 3.9 and 3.10 are tabulated in Tables 3.13, 3.4 and 3.5, showing the evidence of each fixed effect in the models for the north, mid and south basins, respectively. Similar estimates for the fixed-effects of the two models with different error structures at a given site are presented. Since the temperature measurements for the north and south basins are recorded in the same hydrological year, their fixed effects can be reasonably compared. However, there are only few fixed effects in the model for the mid basin than can directly be compared to the north and south basins due to the different predictors in the models and distinct time points of the first recorded data.

The estimate of the initial temperature at 12am on 1 September 2002 shows that the fixed effect $\beta_0$ in the south basin (Table 3.5) is relatively larger than that in the north basin (Table 3.13) and is likely due to greater solar radiation in the south basin. The fixed effect $\beta_1$ denotes the mean of changes in temperature over the time period and therefore, can be used for comparing the changes from different basins. The temperature declines in the north and south basins from

September 2002 - August 2003. Conversely, in the mid basin, $\beta_1$ (Table 3.4) indicates the rise of temperature from April 2008 - May 2009. $\beta_1$ for the north and south basins shows the decrease in temperature over the time period and is likely due to similar ecological process over the same hydrological year. The increase of temperature in the mid basin does not tailor to the changes in the other two basins and such a feature could be attributed by the distinct ecological conditions in different years. The fixed effect $\beta_2$ for the north and south basins may highlight the degree of curvilinear pattern of temperature over the year where similar degrees are shown for both basins.

The standard error for the fixed effect $\beta_0$ from the model in the north is slightly larger than the mid and south basins, indicating that the largest variability of the temperature at the first time point of the hydrological year could be observed in the north basin. The low variability of the $\beta_0$ in the mid basin is likely due to a similar measurements in the first time point at each of the depths. The standard errors for the fixed effect $\beta_1$ and $\beta_2$ of the model for the north are larger than the south basin, indicating a greater variability in the mean change and the degree of curvilinear pattern observed in the north basin. The above features are likely due to the deeper water body (north) that contribute to a larger variation in $\beta_1$ and $\beta_2$ than the shallower water body (south).

The fixed effects $\beta_2$ for the north and south basins are very small and are due to the way of defining the time in the model. The fixed effects $\beta_1$ and $\beta_2$ in the mid basin are relatively larger than that in the north and south basins whilst the estimates of fixed effects $\beta_3$ for the mid basin are very small.

| Term | No Correlation | | | AR(1) | | |
|---|---|---|---|---|---|---|
| | Estimate | s.e. | P-value | Estimate | s.e. | P-value |
| $\beta_o$ | 11.04 | 1.33 | $< 0.001$ | 11.03 | 1.31 | $< 0.001$ |
| $\beta_1$ | -0.22 | 0.05 | $< 0.001$ | -0.24 | 0.04 | $< 0.001$ |
| $\beta_2$ | $2.0 \times 10^{-3}$ | $5.0 \times 10^{-4}$ | $< 0.001$ | $2.0 \times 10^{-3}$ | $4.0 \times 10^{-4}$ | $< 0.001$ |

**Table 3.3.** The fixed-effects from the mixed-effects models with and without AR(1) error structure for the north basin.

| Term | AR(1) | | | AR(2) | | |
|---|---|---|---|---|---|---|
| | Estimate | s.e. | P-value | Estimate | s.e. | P-value |
| $\beta_o$ | 6.66 | 0.10 | $<0.001$ | 6.67 | 0.06 | $<0.001$ |
| $\beta_1$ | 0.52 | 0.10 | $<0.001$ | 0.51 | 0.09 | $<0.001$ |
| $\beta_2$ | -0.01 | $2.1 \times 10^{-3}$ | $<0.001$ | -0.01 | $2.0 \times 10^{-3}$ | $<0.001$ |
| $\beta_3$ | $6.7 \times 10^{-5}$ | $1.4 \times 10^{-5}$ | $<0.001$ | $7.0 \times 10^{-5}$ | $1.3 \times 10^{-5}$ | $<0.001$ |

**Table 3.4.** The fixed-effects from the mixed-effects models with AR(1) and AR(2) error structures for the mid basin.

| Term | AR(1) | | | AR(2) | | |
|---|---|---|---|---|---|---|
| | Estimate | s.e. | P-value | Estimate | s.e. | P-value |
| $\beta_o$ | 16.28 | 0.24 | $< 0.001$ | 16.25 | 0.17 | $< 0.001$ |
| $\beta_1$ | -0.49 | $3.4 \times 10^{-3}$ | $< 0.001$ | -0.50 | $3.3 \times 10^{-3}$ | $< 0.001$ |
| $\beta_2$ | $5.3 \times 10^{-3}$ | $6.0 \times 10^{-5}$ | $< 0.001$ | $5.2 \times 10^{-3}$ | $5.9 \times 10^{-5}$ | $< 0.001$ |

**Table 3.5.** The fixed-effects from the mixed-effects models with AR(1) and AR(2) error structures for the south basin.

The variance of the random effects is used to highlight the variability of the fixed-effect from each of the corresponding random effects at different depths. The variance of the random effects from the models for the north, mid and south basins are tabulated in Tables 3.6, 3.7 and 3.8, respectively. Models with different error structures for each basin shows similar variance for a given random effect. The largest variability of the random effect $b_0$ is noticeable in the north compared to other basins, highlighting greater variability in the mean intercept for the north basin. The variability of the random effect $b_2$ in the north is slightly larger than the south basin, suggesting a greater variability in the curvilinear pattern of temperature over the year between depths in the deeper location of the loch. The above features of the random effects $b_0$ and $b_2$ for the north and mid basins are in broad agreement with the initial exploratory plots in Figures 3.5 and 3.10. The smaller variability of the random effects $b_0$, $b_1$ and $b_2$ for the south basin are apparent compared to the north and mid basins, suggesting the adequacy of the model of temperature without any random effect. Hence, the fixed effect Model 3.7 is fitted to the temperature in the south basin with the error structure as defined by the AR(2) model. The `gls` function in `nlme` library in R (Pinheiro and Bates, 2000) is used for fitting the fixed effect model with correlated errors. The lowest variance of the residuals is shown in the north basin, highlighting the smallest deviation of the errors produced by the model.

| Term | Model | |
|---|---|---|
| | No Corr. | AR(1) |
| $\sigma_o^2$ | 18.75 | 19.53 |
| $\sigma_1^2$ | 0.03 | 0.03 |
| $\sigma_2^2$ | 2.2 $\times 10^{-6}$ | 4.0 $\times 10^{-6}$ |
| $\sigma_e^2$ | 0.46 | 0.22 |

**Table 3.6.** The variance of the random-effects in the mixed-effects models for the north basin.

| Term | Model | |
|---|---|---|
| | AR(1) | AR(2) |
| $\sigma_o^2$ | 0.92 | 0.90 |
| $\sigma_1^2$ | 0.10 | 0.10 |
| $\sigma_2^2$ | 7.57 $\times 10^{-5}$ | 7.40 $\times 10^{-5}$ |
| $\sigma_3^2$ | 3.25 $\times 10^{-9}$ | 3.36 $\times 10^{-9}$ |
| $\sigma_e^2$ | 0.67 | 0.58 |

**Table 3.7.** The variance of the random-effects in the mixed-effects models for the mid basin.

| Term | Model | |
|---|---|---|
| | AR(1) | AR(2) |
| $\sigma_o^2$ | 0.01 | 0.01 |
| $\sigma_1^2$ | 1.21 $\times 10^{-4}$ | 1.44 $\times 10^{-4}$ |
| $\sigma_2^2$ | 3.61 $\times 10^{-8}$ | 3.50 $\times 10^{-8}$ |
| $\sigma_e^2$ | 0.66 | 0.35 |

**Table 3.8.** The variance of the random-effects in the mixed-effects models for the south basin.

A comparison between the two mixed models with different error structures for the north and mid basins is carried out with the use of AIC and BIC and followed by the likelihood ratio test. For the south basin, the fixed effect Model 3.7 with AR(2) error structure is compared to the mixed-effects Model 3.9 with AR(1) correlation structure and the significant model is then compared to the mixed-effects Model 3.9 with AR(2) correlation structure. The goodness of fit of each model; AIC and BIC are also presented.

| Model | df | AIC | BIC | Loglik | L. Ratio | p-value |
|---|---|---|---|---|---|---|
| No Correlation | 10 | 12320.44 | 12392.02 | -6150.22 | 6829.53 | |
| AR(1) Correlation | 11 | -6420.34 | -6341.61 | 3221.17 | 18742.79 | $< 0.001$ |

**Table 3.9.** Model comparison via AIC, BIC and Likelihood Ratio Test for the north basin.

| Model | df | AIC | BIC | Loglik | L. Ratio | p-value |
|---|---|---|---|---|---|---|
| AR(1) Correlation | 16 | -9398.48 | -9273.89 | 4715.24 | 190.58 | |
| AR(2) Correlation | 17 | -9428.39 | -9296.00 | 4731.19 | 31.90 | $<0.001$ |

**Table 3.10.** Model comparison via AIC, BIC and Likelihood Ratio Test for the mid basin.

| Model | df | AIC | BIC | Loglik | L. Ratio | p-value |
|---|---|---|---|---|---|---|
| Simple | 6 | -25987.56 | -25953.91 | 12997.78 | | |
| AR(1) Correlation | 11 | -27154.33 | -27075.60 | 13588.17 | 1180.78 | $<0.001$ |
| AR(2) Correlation | 12 | -27429.33 | -27343.60 | 13726.17 | 276.79 | $<0.001$ |

**Table 3.11.** Model comparison via AIC, BIC and Likelihood Ratio Test for the south basin.

The results of the model testing for the north, mid and south basins are shown in Tables 3.9, 3.10 and 3.12, respectively. The lowest AIC and BIC are shown for the mixed-effects model with the most complex error structure for each basin, indicating that such a model is a considerable model improvement to the model with the less complex error structure. The p-values ($< 0.001$) corresponding to the likelihood ratio strongly suggests that the models of temperature over the year with depth taken to be a random effect provide a substantially better fit to the data following the incorporation of the complex correlation structure of the deseasonalised residuals.

Therefore, the mixed-effects models of temperature with a quadratic pattern is adequate for the north and south basins, whilst the cubic pattern is more appropriate for the mid basin. The mixed-effects models of temperature over the year for the north, mid and south basins incorporate the fixed-effects models 3.23, 3.24 and 3.25, respectively,

$$\{Temp\}_{ij} = 11.03 - 0.24t_{ij} + (2.00 \times 10^{-3})t_{ij}^2 + \epsilon_{ij} \tag{3.23}$$

$$\{Temp\}_{ij} \quad = \quad 6.67 + 0.51t_{ij} - 0.01t_{ij}^2 + (7.0 \times 10^{-5})t_{ij}^3 + \epsilon_{ij} \tag{3.24}$$

$$\{Temp\}_{ij} = 16.25 - 0.50t_{ij} + (5.2 \times 10^{-3})t_{ij}^2 + \epsilon_{ij} \tag{3.25}$$

with the random effects corresponding to the fixed effects are incorporated in each of the model and the error structure within depth in the north basin is defined by the AR(1) model whilst the AR(2) model represents the correlation structures of $\epsilon_{ij}$ for the mid and south basins.

The use of depth as a random effect in the above linear mixed-effects model on temperature over the time period with depths, however, could be criticized since it could also become a fixed effect. The inability of the mixed effects model to provide an estimate of temperature at a particular time point and depth could be overcome by the use of a fixed effects model with the interaction terms as follows.

Two fixed effects models of temperature over the time period for the north basin are fitted and they are compared for the identification of any similar significant predictors. The details of the modelling are as follows.

A regression model is used to model the temperature using time and depth as predictors as well as the interaction between depth and time. Models of temperature, defined in equations 3.26 and 3.27, are initially fitted and a comparison between them is carried out via the F-test. The full model 3.27 incorporates the interaction between depth and each of time and time$^2$. The reduced model 3.26, on the other hand, only incorporates the interaction between depth and time. The depth factors from the surface down to the bottom of the water body are denoted by 1 until 11.

$$\begin{aligned} temp_i &= \beta_0 + \beta_1\{time\}_i + \beta_2\{time\}_i^2 + \beta_3\{depth\}_i + \\ &\quad \beta_4\{time\}_i * \{depth\}_i + \epsilon_i \end{aligned} \tag{3.26}$$

$$\begin{aligned} temp_i &= \beta_0 + \beta_1\{time\}_i + \beta_2\{time\}_i^2 + \beta_3\{depth\}_i + \\ & \quad \beta_4\{time\}_i * \{depth\}_i + \beta_5\{time\}_i^2 * \{depth\}_i + \epsilon_i \qquad (3.27) \end{aligned}$$

A comparison between the above model is carried out using an F-test to determine the appropriate model. The F-statistic and its p-value for models of temperature is tabulated in Table 3.12. The results show evidence of all the predictors and its interaction terms, indicating the appropriateness of the full model to explain the variability of temperature in the north basin.

| Model | df | F-value | p-value |
|---|---|---|---|
| Reduced | 4 | | |
| Full | 5 | 24006 | <0.001 |

**Table 3.12.** Model comparison via F-test for the north basin.

Table 3.13 shows the estimates for the fixed effects from the full model. The estimate of the fixed effect $\beta_1$ shows that the mean temperature declines over the time period. The fixed effect $\beta_2$ may highlight the degree of curvilinear pattern of temperature over the year. The standard error for each of the fixed effects are small (close to 0). The fixed effect $\beta_3$ indicates that the temperature decreases as the level of depth goes closer to the bottom of the loch. The fixed effects $\beta_2$, $\beta_4$ and $\beta_5$ are very small (close to 0) due to the way of defining the time in the model. The standard error for each of the fixed effects are also very small.

| Term | Estimate | s.e. | P-value |
|---|---|---|---|
| $\beta_0$ | 19.1 | 3.5 $\times 10^{-2}$ | $< 0.001$ |
| $\beta_1$ | -0.51 | 1.6 $\times 10^{-3}$ | $< 0.001$ |
| $\beta_2$ | 5.1 $\times 10^{-3}$ | 1.5 $\times 10^{-5}$ | $< 0.001$ |
| $\beta_3$ | -1.2 | 5.3 $\times 10^{-3}$ | $< 0.001$ |
| $\beta_4$ | 4.9 $\times 10^{-2}$ | 2.4 $\times 10^{-4}$ | $< 0.001$ |
| $\beta_5$ | -5.1 $\times 10^{-4}$ | 2.3 $\times 10^{-6}$ | $< 0.001$ |

**Table 3.13.** The fixed-effects model for the north basin.

Figure 3.18 shows the diagnostic plots from model 3.27. Strong autocorrelation and a long tail distribution are highlighted. The skewness problem could be overcome by using the log transform on the temperature measurements and the incorporation of an appropriate AR model for the error structure is required to deal with the violation of the autocorrelation assumption.

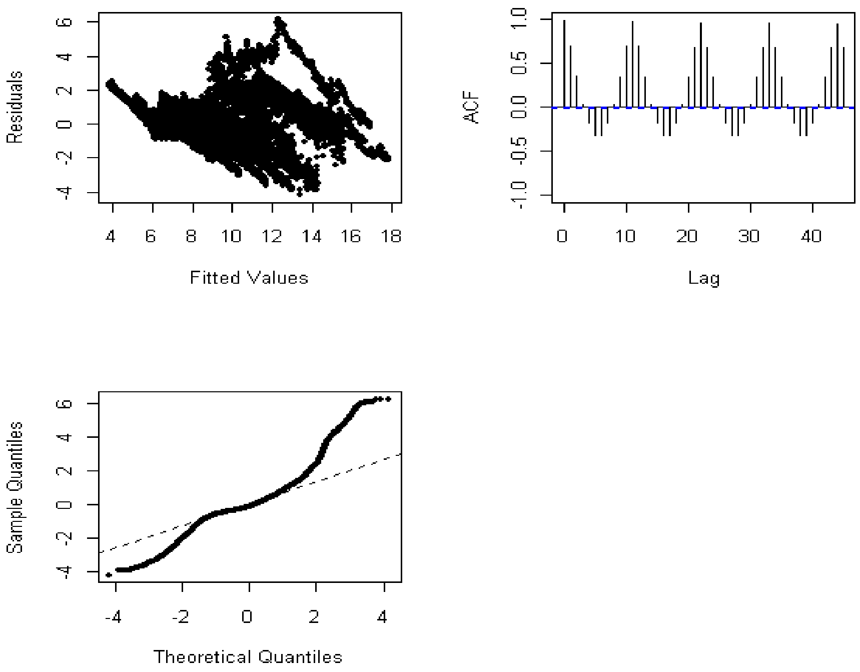

**Figure 3.18.** Diagnostic plots for model 3.27

### 3.5.2 Maximum Relative Rate of Change in Temperature with Depth

Figure 3.19 shows plots of the maximum relative rate of change in temperature with depth in the north (top) and mid (bottom) basins, respectively. The x-axis in both figures are matched in terms of months to highlight differences in terms of the position of the estimated thermocline, between these two basins, although the data are from different years.

For the north basin, the maximum relative rate of change lies between 0.2 and 0.65 between September and November 2002, and June and August 2003, but in January - May 2003, the range is much smaller (0.2 - 0.3). For the mid basin, the maximum changes are mainly between 0.2 and 0.5 with a few closer to 1.0 in April 2008. Such large changes in the mid basin, particularly above 0.8 might be due to a dominant change within three consecutive depths near to the surface in April 2008 when the loch is warmed up.

A question arises about what is the appropriate cut-off point for the maximum relative rate of change approach to reasonably estimate the position of the thermocline.

Hence, several cut-off points between 0.2 and 0.6 at a range of 0.05 were tested and the results for the north and mid basins are depicted in Figures 3.20 and 3.21, respectively. In addition, the depth corresponding to each of the cut-off points for the north and mid basins are presented in Figures 3.22 and 3.23, respectively.

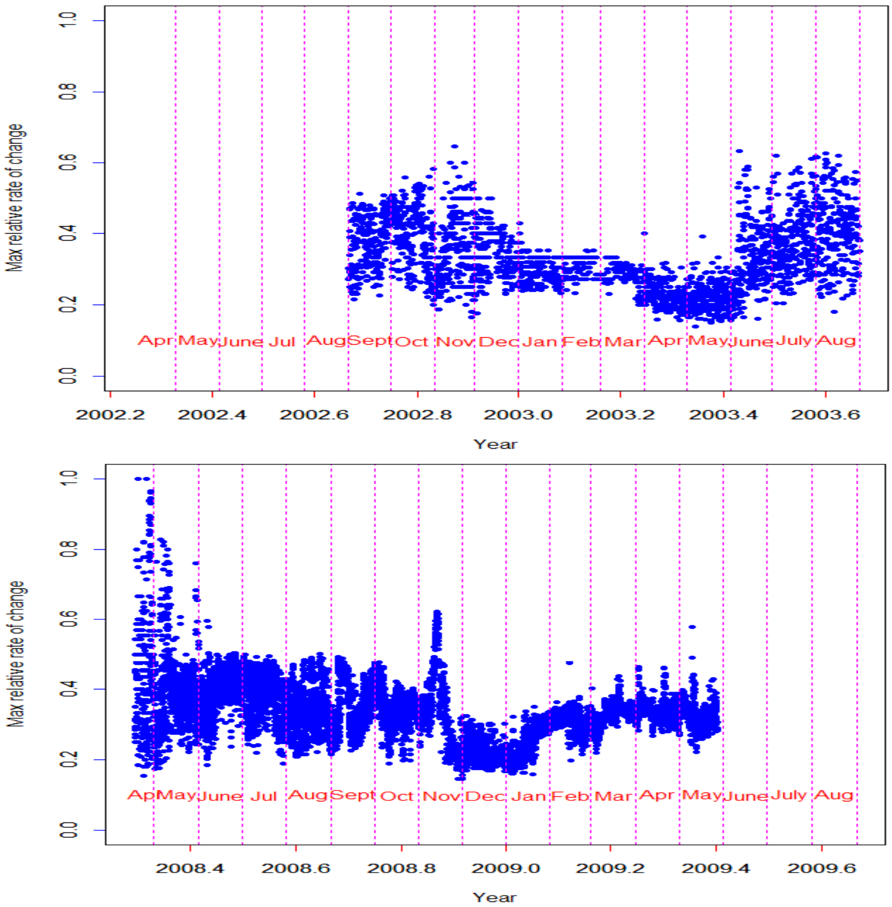

**Figure 3.19.** A series of maximum relative rate of change over time for the north (top) and mid (bottom) basin.

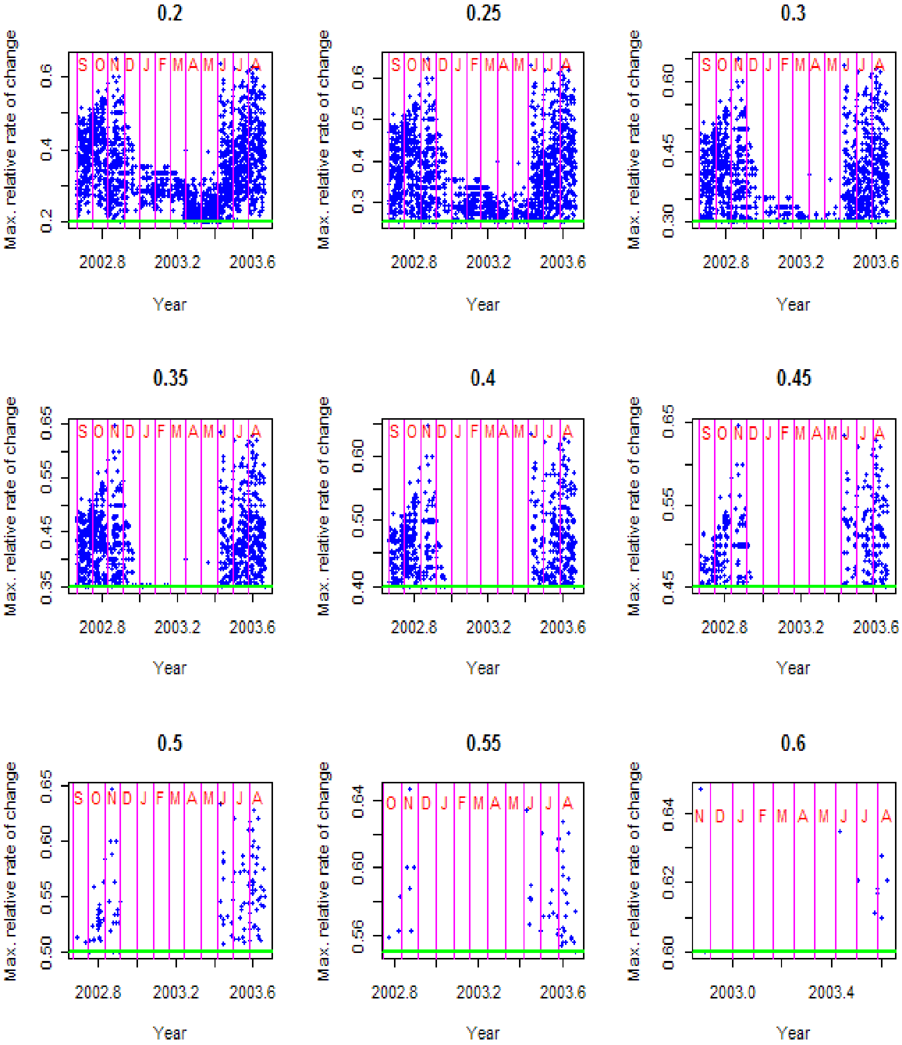

**Figure 3.20.** The maximum relative rate of change over the years in the north basin with cut-off points, between 0.2 and 0.6, ranged 0.05

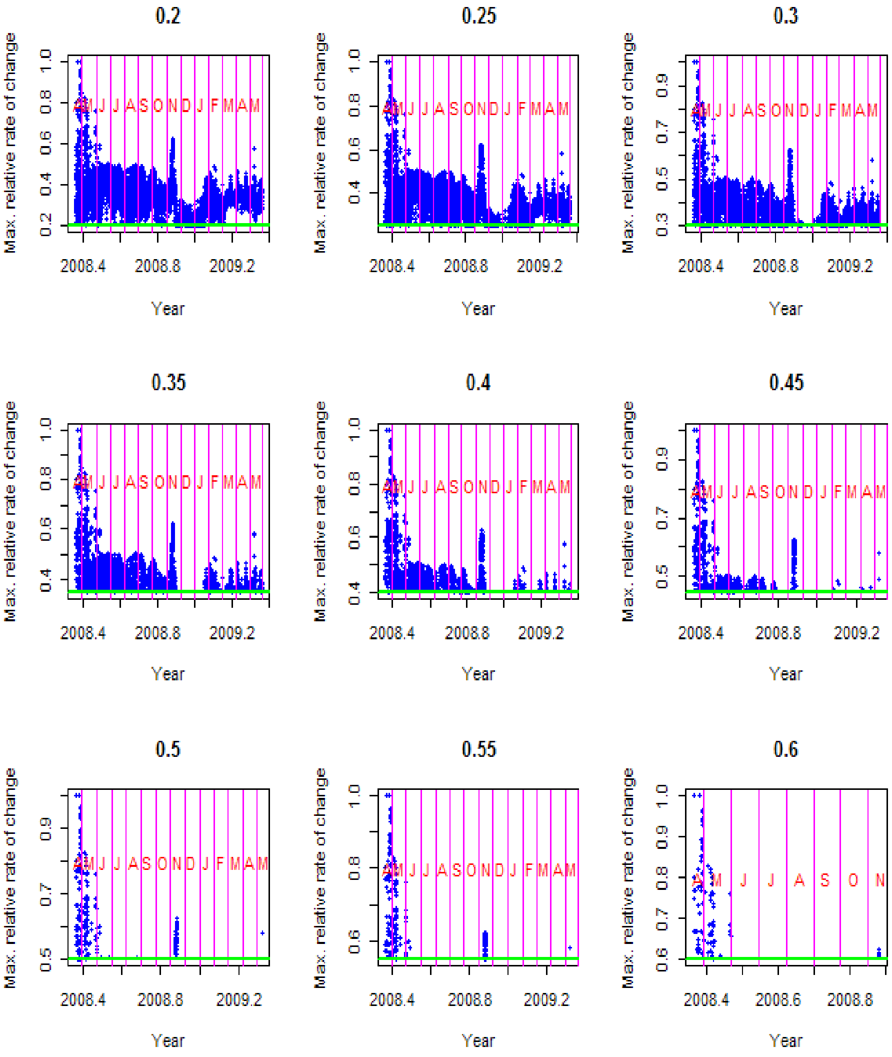

**Figure 3.21.** The maximum relative rate of change over the years in the mid basin with cut-off points, between 0.2 and 0.6, ranged 0.05

For the north and mid basins; at cut-off points between 0.2 and 0.45, the plots do not represent the occurrence of the thermocline very well as there is a large number of points appearing in the winter months and this finding contradicts the feature of the temperature profiles with depth in the exploratory analysis. Conversely, for cut-off points above 0.45, there is no indication that the thermocline occurs in colder months but the drawback is, the number of points becomes extremely small in the warmer months as the cut-off point increases.

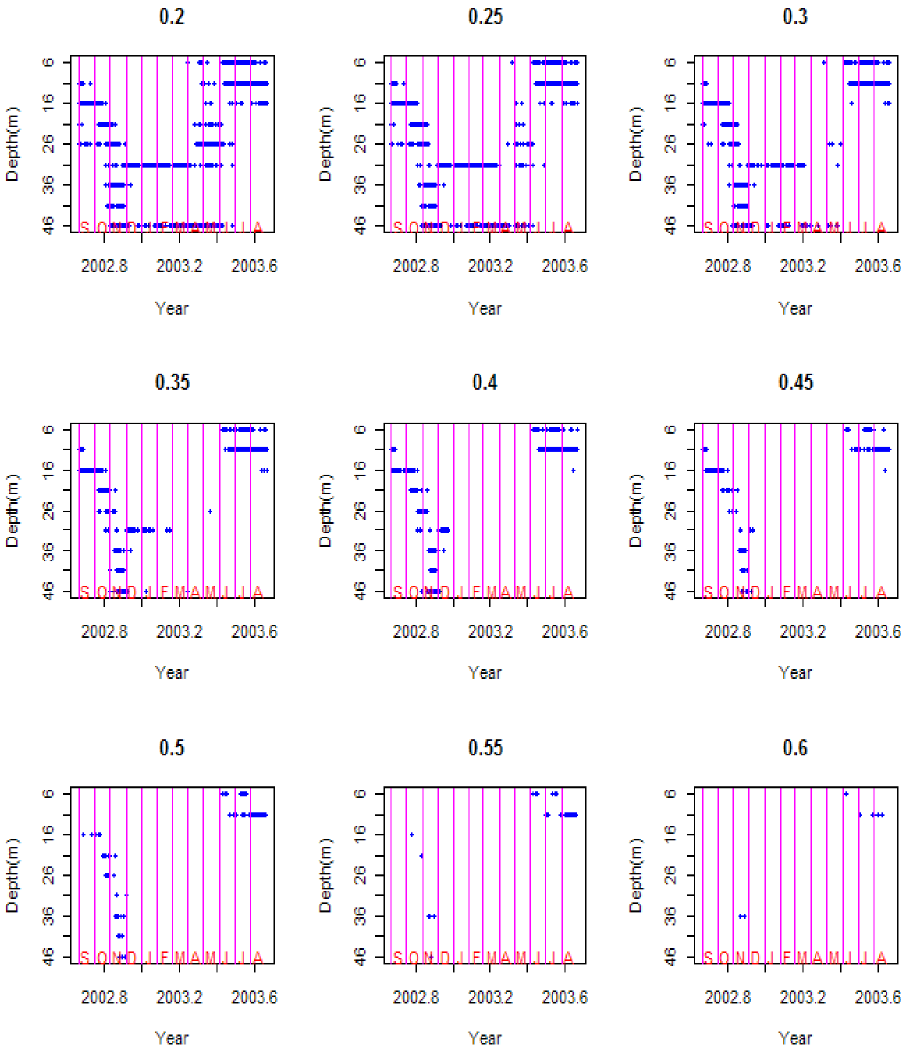

**Figure 3.22.** The depth position over the years in the north basin, with cut-off points between, 0.2 and 0.6, in steps of 0.05

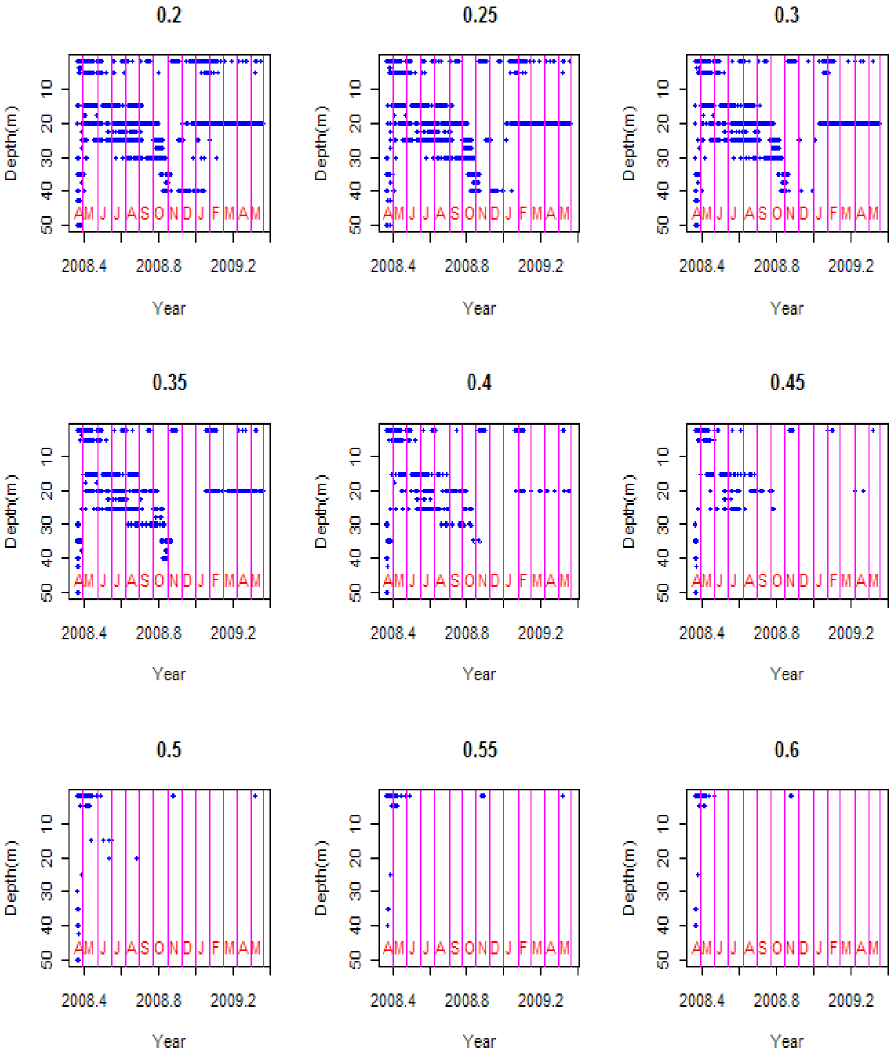

**Figure 3.23.** The depth position over the years in the mid basin, with cut-off points between, 0.2 and 0.6, in steps of 0.05

A possible cut-off point of 0.45, which is subject to the following criteria; a reasonable number of depth positions throughout the year with a small number of depth positions in the winter months, is therefore chosen for both basins. Figure 3.24 displays the results from Figures 3.23 and 3.22 for the north and mid basins at a cut-off of 0.45.

For the north basin; Figure 3.24 (top) suggests that the thermocline initially appears at 11 metres in September and in the depth profile at 26 metres from the surface in November 2002, however, the points determined in December 2002 are more likely to be unidentified features in the lake since the thermocline is not expected to develop in the colder months. There is no evidence of the thermocline development between January and May 2003 as the temperature in the water column is almost constant, as shown by the fairly linear temperature profiles with depth from the exploratory plot. Since the exploratory plot highlights the potential of the thermocline development between April and May 2003, the unobserved points within this period are likely due to the small changes in temperature between depths at a given time point. In mid 2003, the thermocline appears to develop at 11 and 16 metres in June near to the water surface and remains until August.

For the mid basin; Figure 3.24 (bottom) highlights the possible position of the thermocline in several warmer months in 2008. The thermocline appears to develop near to the surface, with no clear temporal pattern, from April to October 2008. The points in the colder months between November 2008 and March 2009, are not taken into consideration as they are not believed to be evidence of the thermocline. This is due to the fact that no thermocline is developed in such periods and this is also highlighted from the feature of the temperature profiles with depth in the exploratory analysis.

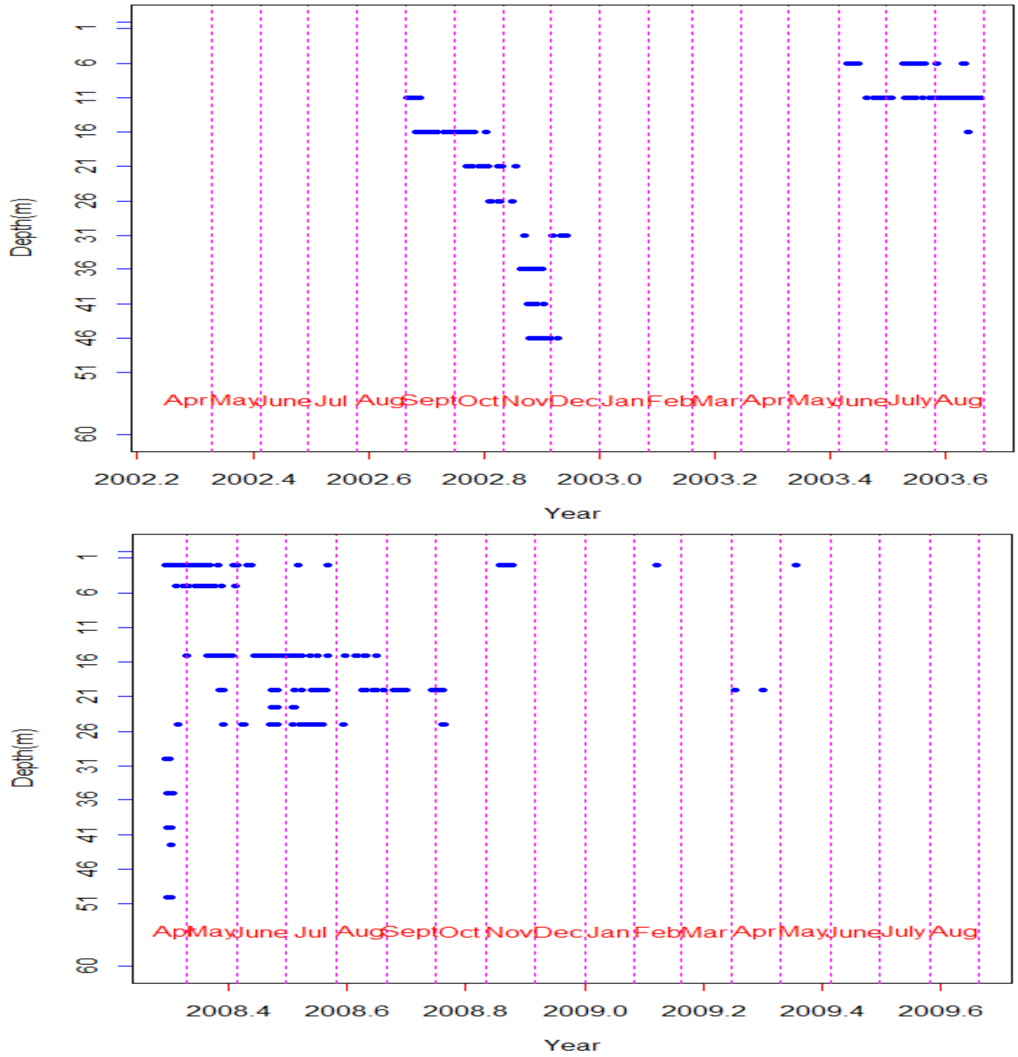

**Figure 3.24.** The depth series through years, corresponding to the cut-off point of 0.45, for the north (top) and mid (bottom) basins

### 3.5.3 Changepoint Regression of Temperature Over Depth

The above 'exploratory' approach is not satisfactory and highly subjective. Therefore, a statistical approach is more appropriate and hence, the investigation is extended using a statistical modelling approach for estimating the thermocline depth. The changepoint regression method as suggested by Julious (2001), is therefore, used to determine the approximate thermocline depth.

Figure 3.25 exhibits time series plots of the estimates of changepoint using this approach for the north (top) and mid (bottom) basins from 1 September 2002 to 31 August 2003 and 17 April 2008 to 27 May 2009, respectively.

Figure 3.25 (top) highlights a cluster of changepoints identified at about 5 - 48 metres to the surface from the end of October to November 2002 and between 5 - 25 metres close to the surface from April to August 2003 in the north basin. There are no changepoints identified in September, in most of October 2002 and between January and March 2003. The results in September and October 2002 contradict the characteristics of the temperature profiles in the exploratory analysis, which suggest evidence of the thermocline formation in these warmer months. This is a result of the fact that the changepoint regression method used here cannot detect the multiple changepoints evident in these months. The changepoints identified in November and December 2002 are more likely to be other features in the temperature profile since the thermocline can only occur in the warmer months. In 2003, there is evidence of the thermocline, appearing at about 25 metres to the surface in April and moving upwards through the water column (although there is no clear temporal pattern), between June and August. These results are consistent with the previous approach.

Figure 3.25 (bottom) shows a cluster of changepoints at about 1 - 48 metres to the surface from April to July 2008 as well as more scattered points near to the surface and bottom of the lake in November 2008, but no changepoints are shown in the same summer months of the following year. The results in November 2008 contradict the features, highlighted from the exploratory analysis, which suggest no evidence of the thermocline formation in the colder months. The changepoints identified in such periods are more likely to be other features in the water body. There is evidence of the thermocline in several warmer months, appearing in April near to the surface and moving down to the middle of the water column of about 25 metres to the surface in July 2008, with a moderately clear temporal pattern. However, no changepoints are estimated in the remaining warmer months, specifically between August and October 2008. This is a result of the lack of potential to pick up several changepoints in each of the temperature profiles with depth using this approach.

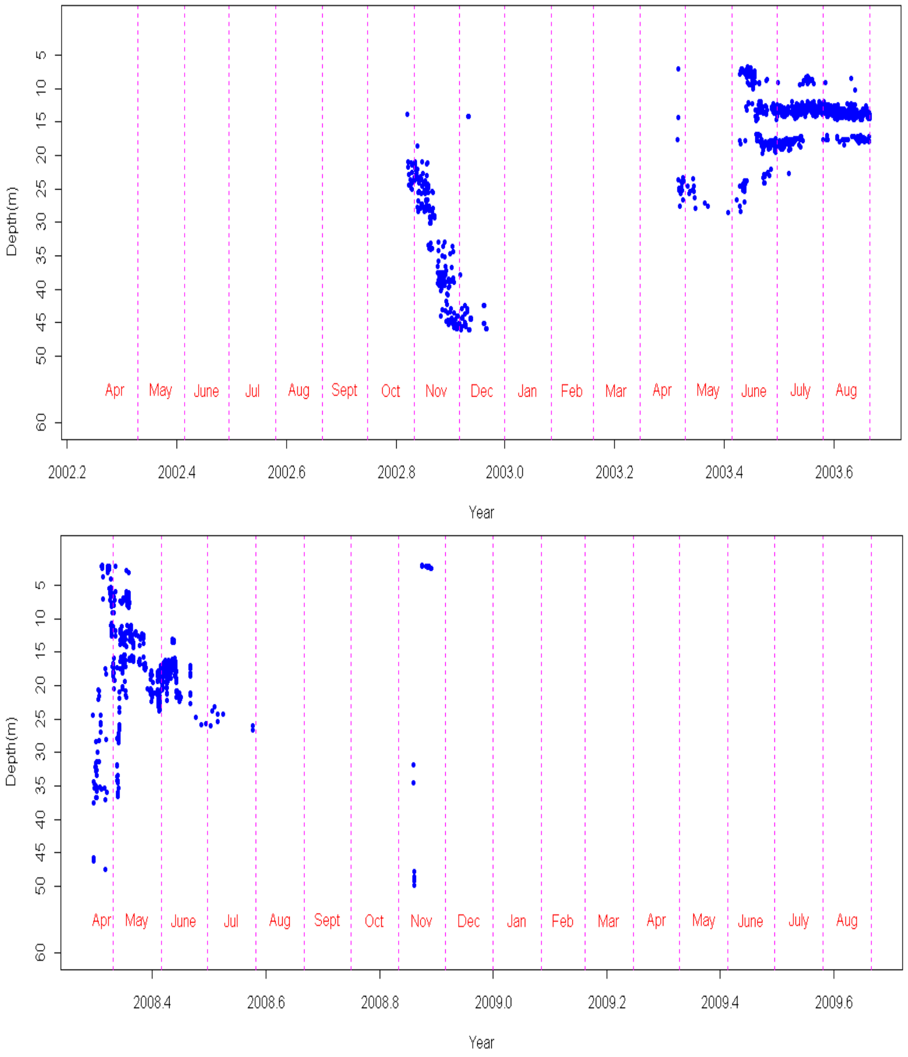

**Figure 3.25.** The estimated change points in the north (top) and mid (bottom) basins.

### 3.5.4 Derivative of A Smooth Curve of Temperature Over Depth

The results from the changepoint regression approach essentially do not show any signs of the thermocline formation in several of the warmer months where it would be expected to occur. The temperature profiles with depth from the exploratory analysis highlight the occurrence of inflection points at particular depths in warmer months. In addition, more than one changepoint is apparent in the temperature profiles with depths, hugely affecting the effectiveness of estimates in the current changepoint regression approach. A final approach will therefore investigate derivatives of smooth curves of the temperature profiles.

Figure 3.26 shows a series of depths throughout the years, corresponding to the inflection points on the smooth curve of the temperature profiles with depth, with respect to the north (top) and mid (bottom) basins. The inflection points associated with the north and mid basins are from September 2002 until August 2003 and April 2008 until May 2009, respectively.

In the north basin, there is evidence of the thermocline, appearing in September at about 18 metres near to the surface and gradually moving down through the water column until 50 metres from the surface in November 2002, with a clear temporal pattern. The thermocline re-appears at about 3-26 metres from April to June 2003. The constant and curvilinear patterns of temperature with depth in the colder and warmer months, respectively, producing inflection points at a position which is far beyond the metalimnion. As a result, some of the inflection points in the north basin, were identified at the bottom and top of the water column between November 2002 and May 2003 and from April to August 2003, respectively.

In the mid basin, there is evidence of the thermocline, developed at approximately 3 metres near to the surface in April and moving down the water column until 15 metres near the surface in May and remains constant in about the same position in the water column until August 2008. Finally, the thermocline moves down the water column until 40 metres from the surface in October 2009. The results from November 2008 to March 2009 do not match the characteristics of the temperature profiles in the exploratory analysis, showing no proof of the thermocline formation. The points observed within this period are due to the almost zero curvature of the temperature pattern down the depth profile at a given time point. The thermocline re-appears in April 2009 at about 10-15 metres and remains constant until May 2009. The inflection points marked at about 10m from the surface, between November 2008 and March 2009 are not given much attention since no evidence of the thermocline in the colder months is displayed in the exploratory plot. In fact, the inflection points, appearing at these positions may represent other features in the water column.

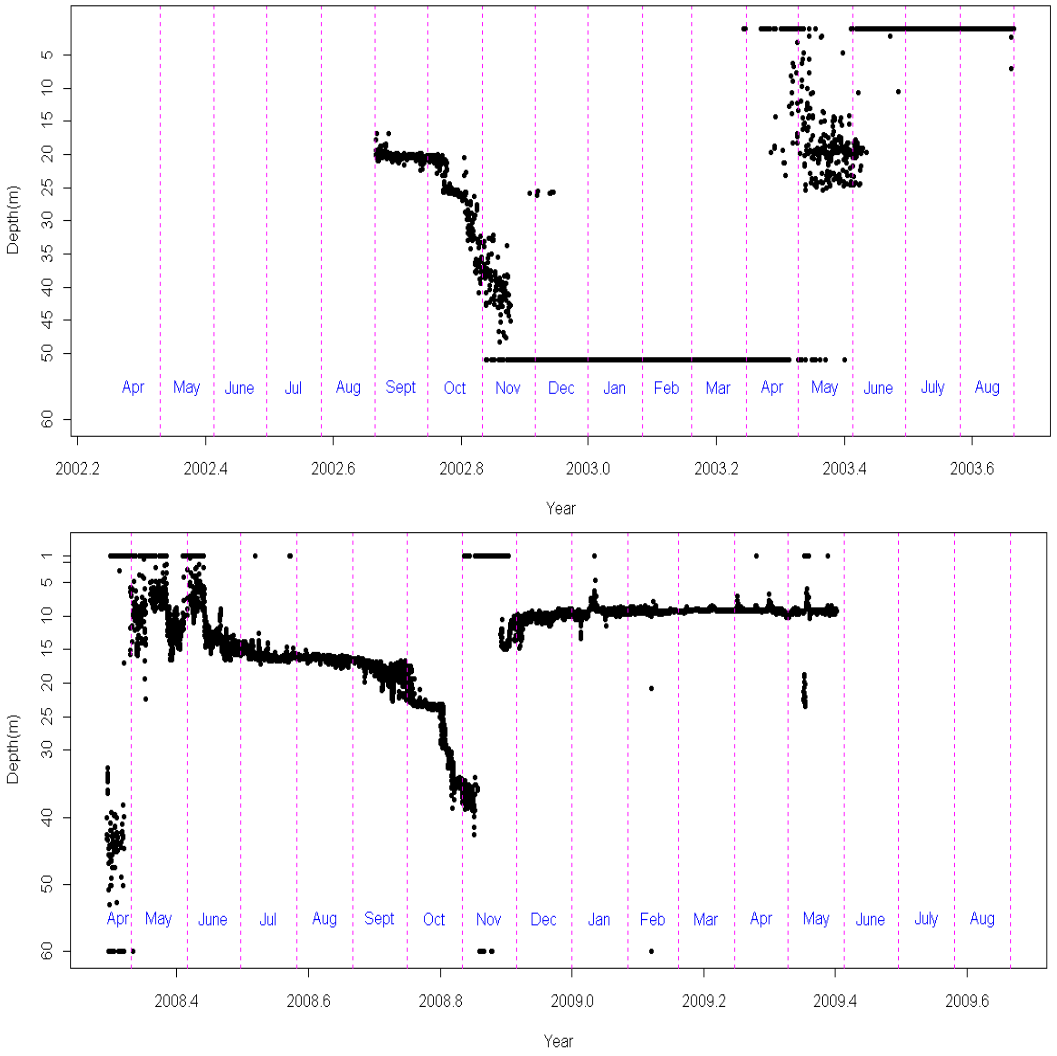

**Figure 3.26.** The estimated inflection points throughout years in the north (top) and mid (bottom) basins.

### 3.5.5 The Plausible Month of the Thermocline

Tables 3.14 and 3.15 show the months of the observed thermocline resulting from different approaches; the maximum relative rate of change, changepoint regression and derivative of a smooth curve of temperature denoted by 1, 2 and 3, respectively, for the north and mid basins. The warmer months are of interest here due to the fact that the thermocline develops in summer.

The maximum relative rate of change shows no thermocline is observed in April and May 2003 for the north and October 2009 for the mid basin. The unobserved points within these warmer months are likely due to the small changes in temperature between depths at a given time point. The changepoint regression highlights the unobserved thermocline in September 2002 for the north and September 2008 and April-May 2009 for the mid basin and are likely due to the fact that the changepoint regression used here cannot detect the multiple changepoints evident in these months. Despite the previous results from the derivative of a smooth curve for the north basin, showing the observed thermocline as discrete values close to the surface of the loch from June-August 2003, these positions are not taken into attention since the results are unlikely to occur. The curvilinear pattern of temperature with depth within this period leads to the estimates of the infection points at a position which is far beyond the metalimnion.

| Month | Year | Approaches | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| Sept | 2002 | / | – | / |
| Oct | 2002 | / | / | / |
| Nov | 2002 | / | / | / |
| Dec | 2002 | – | – | – |
| Jan | 2003 | – | – | – |
| Feb | 2003 | – | – | – |
| Mar | 2003 | – | – | – |
| Apr | 2003 | – | / | / |
| May | 2003 | – | / | / |
| June | 2003 | / | / | – |
| Jul | 2003 | / | / | – |
| Aug | 2003 | / | / | – |

**Table 3.14.** The observed thermocline for each month over the year from different approaches; the maximum relative rate of change (1), changepoint regression (2) and derivative of a smooth curve of temperature (3), for the north basin.

| Month | Year | Approaches | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| Apr | 2008 | / | / | / |
| May | 2008 | / | / | / |
| June | 2008 | / | / | / |
| Jul | 2008 | / | / | / |
| Aug | 2008 | / | / | / |
| Sept | 2008 | / | – | / |
| Oct | 2008 | – | – | / |
| Nov | 2008 | – | – | – |
| Dec | 2008 | – | – | – |
| Jan | 2009 | – | – | – |
| Feb | 2009 | – | – | – |
| Mar | 2009 | – | – | – |
| Apr | 2009 | / | – | / |
| May | 2009 | / | – | / |

**Table 3.15.** The observed thermocline for each month over the year from different approaches; the maximum relative rate of change (1), changepoint regression (2) and derivative of a smooth curve of temperature (3), for the mid basin.

## 3.6 Discussion

The exploratory plots show some important features in the pattern of temperature over the year, with depths, in the north, mid and south basins. In particular, there is apparent variability of temperature over warmer months at a given depth in the north and mid basins. A similar temperature pattern is depicted for each depth in the south basin over warmer months in 2002 but it begins to deviate in the warmer months of the following year (2003). The temperature over the colder months for each basin are similar at a given depth.

Such a feature in the shallower part (south basin) of the loch over the warmer months in 2002 is expected since the rate of the heat conduction, attributed by solar radiation on the water surface down the depth profile is similar and so, the entire water body in the south basin is affected by a similar amount of heat transfer. The deviation of the temperature over the warmer months in 2003 is likely due to the influence of the wind on the water surface. The north and mid basins on the other hand, are dominantly influenced by solar radiation. In addition, higher wind speed on the surface of the loch may contribute to the large variability in the temperature pattern over the summer period, with depth. Dissimilarity of temperature profile over the summer, with depths, may suggest different ecological process in the water body. In winter, similar temperature pattern for each depth may indicate the similarity of the ecological process in the water column.

The evidence of correlation in the time series of deseasonalised residuals for each depth in the north, mid and south basins as highlighted by the autocorrelation functions, is anticipated as the dependency of the errors is often observed in environmental time series data. The AR(1) model for the north and AR(2) model for the error structures in the mid and south basins were appropriate. The north and mid basins are predominantly deeper water and so, similar error structure is expected in the north and mid basins.

Several diagnostic tools are used for assessing the adequacy of the linear mixed-effects models such as AIC and BIC and likelihood ratio tests of two nested models.

The lowest AIC and BIC in measuring the relative goodness of fit of the models shows the adequacy of AR(1) for the north and AR(2) for the mid and south basins error structures. Despite the lowest variability of the random effects corresponding to the fixed effects in the model of temperature over the year, with a quadratic pattern, the likelihood ratio test shows evidence of the mixed model incorporating the most complex error structure for each basin. Hence, the linear mixed-effects models, which accomodate depth-specific variation in its random effects and ties together all levels of depth by fixed-effect and variance-covariance matrix are the appropriate model for each basin.

The mixed-effects model provides information on changes in temperature over the year for each basin. The incorporation of the random effect $b_{1j}$ may highlight the mean of changes of temperature at a given depth and so, the variability of the random effect $b_{1j}$ can also be highlighted. The above feature shows that the mixed-effects model is an efficient way to highlight the variability in temperature measurements at different depths compared to the classical approach that required model fitting for each depth. Additionally, the mixed-effects model allows the estimates of the variability for the fixed effects via its random effect and, more informative features can be gained.

While the exploratory plots highlight the variability of temperature pattern over the year for a given depth in each basin, the homogeneity of the temperature profile with depth at a given time point in the north and mid basins are also explored prior to estimating the position of the thermocline. In the colder months, a constant temperature with depth is highlighted whilst approximately quadratic and cubic patterns of temperature profiles with depth appear in the warmer months. However, the odd patterns in the mid basin from March - May 2009 might be the result of the presence of unrecognized physical disturbance in the water column as the temperature with depth is reasonably constant over the colder months. The contour plots of temperature across depth and year highlight the occurrence of the thermocline and so, few approaches are used to identify the position of this natural feature in the loch.

The maximum relative rate of change, changepoint regression and derivative of a smooth curve methods have proven to be useful in determining the formation of the thermocline in the north and mid basins in warmer months. Ecologically there appear to be some differences between the two basins and the two time-periods considered. In particular, deeper water in the north than the mid basin provides distinct lower and upper positions of the thermocline. The different periods of temperature measurements between the two basins may results in different estimates of the positions of the thermocline in the loch. Such a result is likely due to the difference in ecological processes in the warmer months. Since the period of temperature measurements can be distinguished between the two basins, the ecological processes that characterize the variability of temperature with depths at a given month may not always be similar. For instance, the exploratory plot of temperature with depth in November highlight homogeneous temperature down the depth profile in the mid basin but fairly nonhomogeneous in the south basin. Consequently, the estimates of the thermocline for the mid basin in November 2008 should be given much attention since it is likely that the thermocline is developed in the water column compared to November 2002 in the north basin.

In the north and mid basins of Loch Lomond, the maximum relative rate of change approach with a cut-off point of 0.45 shows that the number of depths corresponding to maximum changes in the temperature gradient with depth in the warmer months are large.

Despite providing the positions of the thermocline subject to proposed criteria, several drawbacks in the maximum relative rate of change approach should be taken into account. The depth corresponding to the maximum relative rate of change simply depends on three measurements from each of the temperature profiles with depth, indicating that only a small number of observations contribute to the determination of the identified position of the thermocline. In addition, this approach is constrained by the value of the cut-off point, which is subjective. Furthermore, the approximate thermocline depths presented by this method, which are treated as discrete instead of continuous values, do not adequately tailor to the natural ecological perspective. The problems of insufficient statistical property and unnatural positions of the thermocline provided by this approach is therefore, overcome by the use of a changepoint regression approach.

The changepoint regression has proved to be a useful and effective way to estimate the changepoints corresponding to the thermocline depths in a few warmer months because of its capability to highlight the positions of a sudden change in temperature with depth in both basins. Despite providing appropriate positions of the thermocline in the warmer months, this method is unable to highlight the occurrence of the thermocline when there is more than one changepoint.

The derivative of a smooth curve approach is essentially an effective way to investigate the position of the thermocline since it has provided appropriate estimates of inflection points in several warmer months in both basins. This approach essentially provides estimates at each of the time points and so, the entire results have to be carefully examined to avoid any misleading interpretations. A good knowledge of freshwater ecology is essential and therefore, should be acquired prior to perform this approach. Despite using third order local polynomial regression, the smooth curves of the temperature profiles with depth in the colder months provide estimates which do not make sense as the points lie at the top and bottom of the lake. These point estimates, however, have not received as much attention as in the warmer months as no thermocline develops in the colder period.

A combination of changepoint regression and derivative of a smooth curve are essentially providing appropriate estimates of the thermocline depths in the warmer months and therefore, is suggested here. This is due to the fact that the change-point regression and derivative approaches satisfies the thermocline definition in the perspective of limnologist and mathematician, respectively. Additionally, the effectiveness of each of these approaches could be distinguished in different warmer months and therefore, they are given attention.

Despite several advantages in the proposed approach to estimate appropriate thermocline depths in the warmer months, a number of drawbacks should be taken into consideration for future study. The present approach has inadequate physical basis which results in less information of dynamical interaction on the development of the thermocline. The absence of influence of physical factor such as wind's velocity, which may effect the turbulence of the water and the creation of the thermocline, may result in a lack of ability to estimate the position of such a natural feature in lakes in warmer months. A model of the position of the thermocline over the years is not produced by the current approaches and therefore, any predictions of such an important feature in a lake could not be carried out.

# Chapter 4

# Variability, Coherence and Recovery

## 4.1 Introduction

The previous two chapters have investigated statistical models for low and moderate frequency environmental time series. Moreover, with the introduction of automatic monitoring buoys, semi-continuous measurements are becoming much more common (Kuh et al., 2005), providing greater insight with instantaneous ecological processes within waterbodies.

Three objectives are outlined in this chapter. Firstly, to assess if there are temporal patterns in temperature, pH, conductivity and barometric pressure at short time scales. Secondly, to investigate the nature of the relationship between pH and conductivity at short time scales. Finally, to determine and model the recovery period of pH and conductivity following extreme discharge events.

Continuous 15-minute and 30-minute measurements of temperature, barometric pressure, pH, conductivity and discharge are available for the River Char in Aberdeen and Drumtee stream in Whitelee over three hydrological years. The data from Charr and Drumtee were recorded from October 2004 - September 2007 and October 2007 - September 2010, respectively. The data were supplied by Professor Susan Waldron from the University of Glasgow.

Continuous environmental time series measurements are useful as they can reveal evidence of short-term variability and extremes in river water quality become apparent and they allow the examination of water quality signals over short time scales (Jarvie et al., 2001). Short-term water quality events and patterns may include acid excursion during periods of high rainfall and variation in the diurnal cycle in pH related to biological processes in the rivers (see (Neal et al., 1998) for examples).

Diurnal fluctuation of physical and chemical parameters are typical in many drainage systems (Tobias et al., 2010). Water chemistry, particularly $CO_2$ in rivers may affect the variability of pH (Stumm and Morgan, 1981). Closer examination of pH revealed distinct diurnal variations during base flow in the River Dee, most apparent in the summer when pH was high and its signal showed a moderately large amplitude in the diurnal cycle (Jarvie et al., 2001). Florentina et al. (1999) shows that river pH is also influenced by discharge where organic acids and soil $CO_2$ exported under high flow result in lower pH. Where discharge is not a controlling factor however, the relationship between $CO_2$ and pH is essentially related to photosynthesis. Additionally, Florentina et al. (1999) shows that the daily variations in pH are closely related to changes in solar radiation and water temperature.

The changes in temperature in rivers could be influence by biological and physical processes which occur periodically. For instance, a diurnal variation in water temperature is a result of the occurrence of photosynthesis and respiration. Superimposed on this daily fluctuation are seasonal fluctuations, reflecting the hydrological regime, seasonal climate changes, corresponding seasonality of biogeochemical cycling and additionally anthropogenic impacts (Tobias et al., 2010).

Conductivity can show significant temporal variation as this is a measure of the ionic concentration in the rivers and so is primarily controlled by the influence of rock weathering derived water. However, $CO_2$ may contribute to the changes in cations (negative ions). This under high event flow conductivity, can also change due to influx of water rich in $CO_2$ and low in dissolved ions. Conductivity is also influenced by hydrological events. For instance, Tobias et al. (2010) show that a rapid decrease in conductivity in rivers over a few hours in late spring is a result of a large amount of snow melting. In particular, the increase of $CO_2$ may contribute to changes in cations (negative ions) and so, conductivity may also be affected.

Barometric pressure, also known as air pressure or atmospheric pressure, is the pressure brought down by the weight of air (Phillis, 1997). The rise of barometric pressure may result in increase in water movement in windy days (Stevenson and Van Schaik, 1967) and hence, the river flow could be affected by such air pressure. The following characteristics of barometric pressure are taken from (Phillis, 1997). The amount of air pressure is a combination of the number of molecules of air, the rate of movement of the molecules and the frequency of their collision. In particular, a strong gravity results in a large number of molecules of air. The increase of mass in the air resulting from the large air molecules contribute to the increase in air pressure. The air pressure may change constantly, resulting in the occurrence of wind and it is useful for weather forecasting. In particular, the rise of barometric pressure often coincides with fair weather whilst the decline of pressure is associated with severe storms.

Discharge also known as streamflow, is defined as the volumetric rate of flow water (volume per unit time) including sediment or other dissolved solids in an open channel of water (Turnipseed and Sauer, 2010). Discharge may vary over many time scales and is influenced by hydrological events. In particular, the longer time scales can be interannual or seasonal, whereas shorter time scales can last only several days, such as during a flood (Yucheng et al., 2011). Rajendra and David (2002) show that the increase of discharge at the onset of the winter period is characterized by high rainfall. Theoretically both pH and conductivity may show changes corresponding with discharge. Josep and Anna (1992) show an inverse relationship between pH and discharge for the weekly stream water measurements in Prades and Montseny catchments and a similar relationship between discharge and conductivity is expected. The increase of discharge results in the rise of $CO_2$ (Josep and Anna, 1992) and so, both pH and conductivity are likely to be influenced. Hence, pH and conductivity may behave similarly following particular features in discharge.

Gurnell et al. (1992) and Hodgkins (2001) found that a combination of time domain statistical techniques such as linear regression, cross-correlation, autoregressive integrated moving average (ARIMA) and transfer functions models are acceptable to explain the variability of the process in climatic and hydrological time series data. However, different statistical approaches are required to provide further insight on the relationship between timing and magnitude of the events via a time scale (or frequency method). In principle, the identification of scales of variability in a time series can be achieved by the use of spectral analysis, which is strictly based on the assumption of stationarity in time series. This frequency-based approach decomposes a time series into its frequency components. The time series is broken up into sine waves and hence, the signal can be expressed in terms of the frequency and power of its constituent waves. However, real world processes are often non-stationary, resulting in changes in their statistical correlation properties over time and thus, the assumption required for this approach is violated. It has become a standard practice for time series analysts to consider differencing or transforming their time series to achieve stationary time series (Jin and Yao, 2006), however, these approaches do not always work well. Higher time resolution always appear with a lower frequency resolution and vice versa, causing another problem and so, a compromise between time and frequency uncertainties is therefore required to describe the process. The wavelet method is capable of providing the above compromise and its advantages over the Fourier transform is that the chosen time resolution is proportional to scale and it is designed for both stationary and non-stationary processes over numerous frequencies scale (Daubechies, 1990); (Kumar and Fouroula-Georgiou, 1997) and therefore, misleading interpretation can be avoided. The key to wavelet analysis is partitioning the variations of signals into scale (frequency/period) and time localization and hence, the detailed variation with respect to temporal scale and time locations can be acquired.

An increase in discharge may result in a decline of pH and conductivity following high precipitation and snow melting. However, both pH and conductivity could increase and return to their natural levels if none of these hydrological processes occur. Hence, the period taken by each of pH and conductivity to recover may characterize certain ecological features in rivers. There does not appear to have been previous work on the estimation of such a recovery period and therefore, the terminology of the period of this natural process in rivers is required. The recovery period is defined as the duration of time taken by pH and conductivity to recover following high river flows. Specifically, the recovery period for a particular environmental variable in rivers begins from the time point where the measurements start to decrease in response to an event in discharge, until the values return to the pre-event levels. Such a discharge event is determined by a particular threshold. It is generally considered that a threshold can be found that is high enough to ensure the independence of the occurrences in the hydrological series (Santiago, 2005). The baselines are defined as the measurements of pH and conductivity corresponding to the 2 hours before the extreme discharge. The recovery period for pH and conductivity at both rivers is determined based on its definition. In particular, the recovery period is determined based on considerably higher river flow.

## 4.2 Exploratory Analysis

The initial impressions of temporal variability in temperature, barometric pressure, pH and conductivity, and the relationship between pH and conductivity are presented for the River Charr. The above work is followed by the exploration of the recovery period of pH and conductivity at the same river. However, the exploration of recovery period for pH and conductivity at any other river is essential for comparison since literature on previous studies done on this natural process has not been found and so, the initial analysis is then extended to the River Drumtee in Whitelee.

The features and properties of the time series of temperature, barometric pressure, pH and conductivity measurements from October 2004 - September 2007 at the River Charr are explored graphically. Plots of temperature, barometric pressure, pH and conductivity measurements within each of the days for each month in the study period are then explored to ascertain any daily temporal pattern. Additionally, the relationship between pH and conductivity is explored for any similar patterns throughout the year.

Time series plots for each of the environmental determinands over the time period are displayed to obtain initial impressions of any particular patterns. Plots of temperature, barometric pressure, pH and conductivity from October 2004 - September 2007 are shown in Figure 4.1. There are large amounts of variability for each of the time series measurements. Temperature shows a clear seasonal pattern over the 3 years period following a natural cyclical pattern for each year. Conversely, pH, conductivity and barometric pressure show indistinct annual cyclicals with the lower values more pronounced in winter than summer for each year. The lower values could be as a result of any hydrological events such as precipitation and discharge.

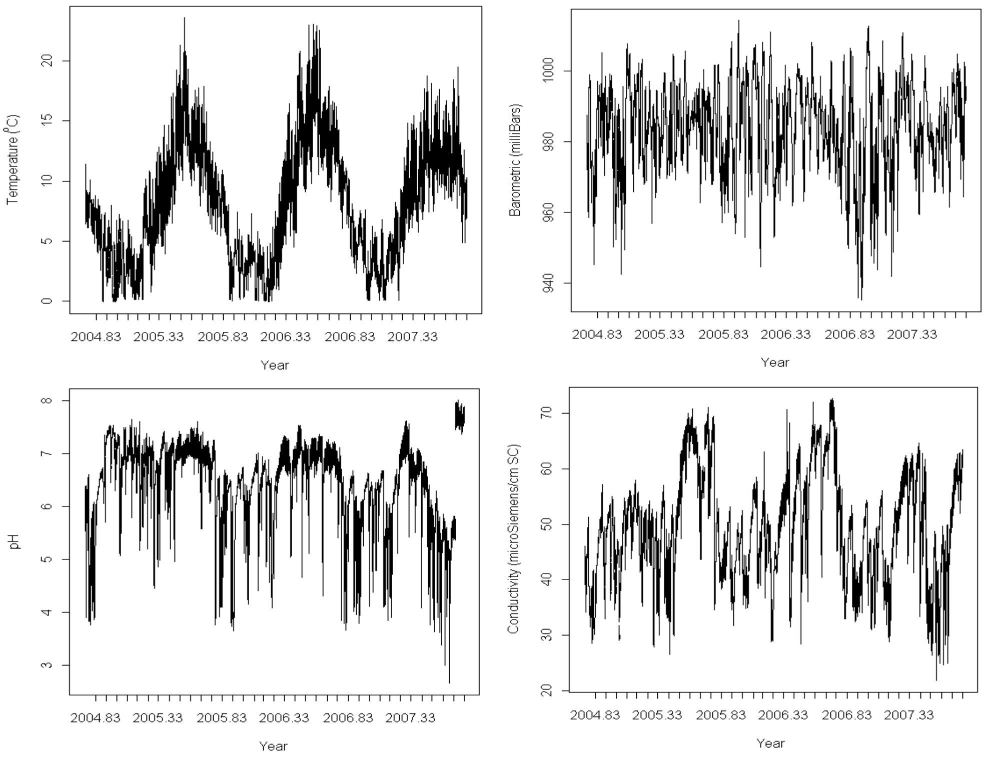

**Figure 4.1.** Plots of 15 minutes temperature (top left), barometric pressure (top right), pH (bottom left) and conductivity (bottom right) measurements from Oct 2004-Sept 2007.

Previous studies show that the observed time series has to often be pre-treated before applying the wavelet power spectrum to remove components of the signal that are not of interest. For instance, Torrence and Compo (1998) removed the seasonal means for the entire record of temperature time series of El Nino-Southern oscillation, Torrence and Webster (1999) treated the rainfall time series by removing the mean of the measurements and Schaefli et al. (2007) deseasonalized the time series of temperature by substracting the interannual mean value from the measurements and followed by division by the interannual variance. They also applied the log-transform and cubic root for discharge and precipitation series, respectively.

Here, a first order differencing is used for each of the observed time series to remove the trend component which is not of interest. As a result, the linear trend could be removed and the large variability of each variable could be reduced, and the resulting data are plotted in Figure 4.2. There are obviously no linear trend exhibited by each of the post-differenced data.

A direct interpretation for each of the post-treated data to the actual measurements is used by the above researchers in the wavelet analysis. Hence, a direct interpretation for the wavelet spectrum which does not take the occurrence of a linear trend into account is used in this work.

The variability of temperature is lower and stable compared to the remaining variables, despite highlighting a seasonal pattern over the year. Barometric pressure shows more variability than pH and conductivity. pH and conductivity highlight clearly unstable variability over the year.

(Jarvie et al., 2001) show the effectiveness of 15-minute measurements in environmental time series measurements - short-term variability is apparent in summer at the River Dee at Mar Lodge in Scotland. Therefore, I first explore the temporal variability in temperature, barometric pressure, pH and conductivity for the River Charr but not for Drumtee as this site is logged less frequently: 30 minutes measurements rather than 15 minute. However, exploration of the recovery period for pH and conductivity is carried out for both rivers and a comparison between them is made.

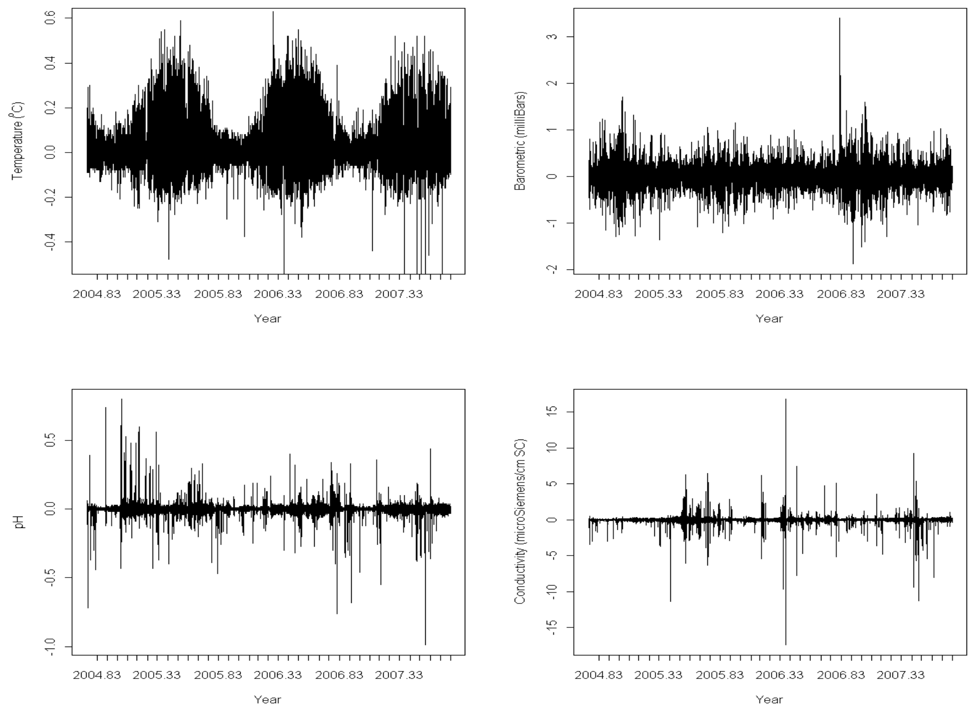

**Figure 4.2.** Plots of first order differencing on 15 minutes temperature (top left), barometric pressure (top right), pH (bottom left) and conductivity (bottom right) measurements from Oct 2004-Sept 2007.

Figures 4.3 to 4.6 display plots with respect to temperature, barometric pressure, conductivity and pH measurements for each day, grouped by month. The black, red and green curves are for the period of 2004-2005, 2005-2006 and 2006-2007, respectively. Figure 4.3 shows a pronounced diurnal pattern in temperature with high variability from April to September for each of the years, indicating that a daily temporal pattern has occurred in the warmer months. The daily pattern is similar within the day for the same month, whilst its variability could be distinguished between months. In contrast, no clear daily pattern in winter (October - March) is apparent since the time series are fairly flat. Figure 4.4 shows no apparent daily pattern in barometric pressure for each month, however, the variability in winter (October-March) is slightly larger than in the summer (April-September) months. Figure 4.5 reveals weak diurnal cycle in conductivity over the summer (April-September) for most of the years. In winter (October-March), however, no clear daily pattern is shown as highlighted by time series which are fairly flat. There are several odd patterns, particularly in March and May, which might be due to measurements errors. Figure 4.6 shows a weak diurnal cycle in pH over the summer (April-September) and winter (Feb-March and October) for most of the years. No clear daily pattern is shown in the remaining months in winter as exhibited by approximately constant curves across the day.

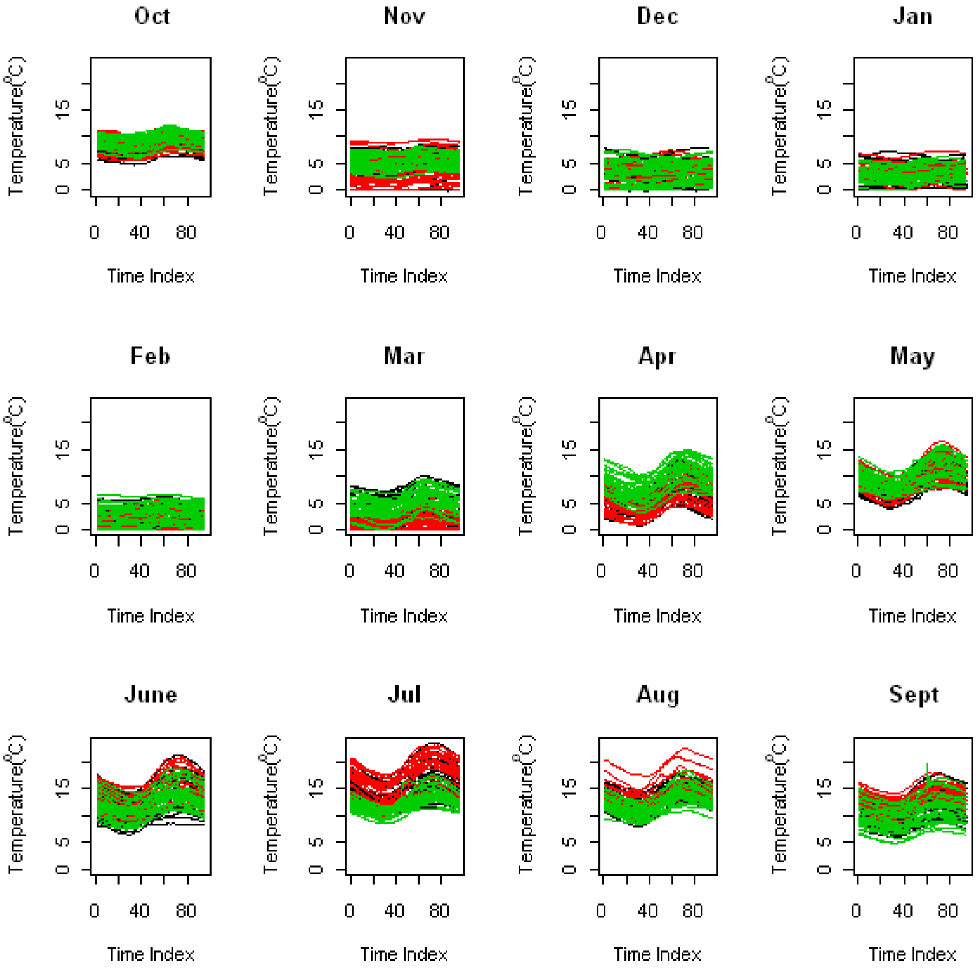

**Figure 4.3.** 15-minute measurements for temperature in a daily period for each of the years, grouped by months, with the black, red and green lines are for the periods of 2004-2005, 2005-2006 and 2006-2007, respectively.

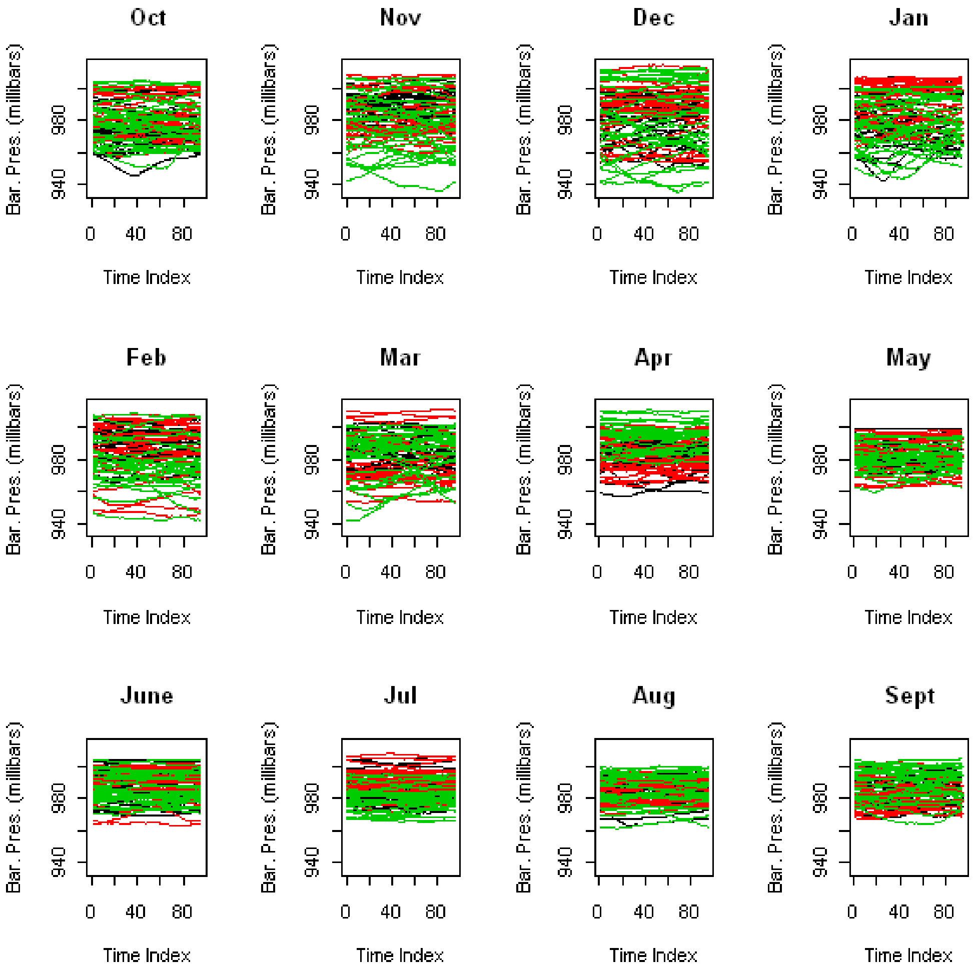

**Figure 4.4.** 15-minute measurements for barometric pressure in a daily period for each of the years, grouped by months, with the black, red and green lines are for the periods of 2004-2005, 2005-2006 and 2006-2007, respectively.

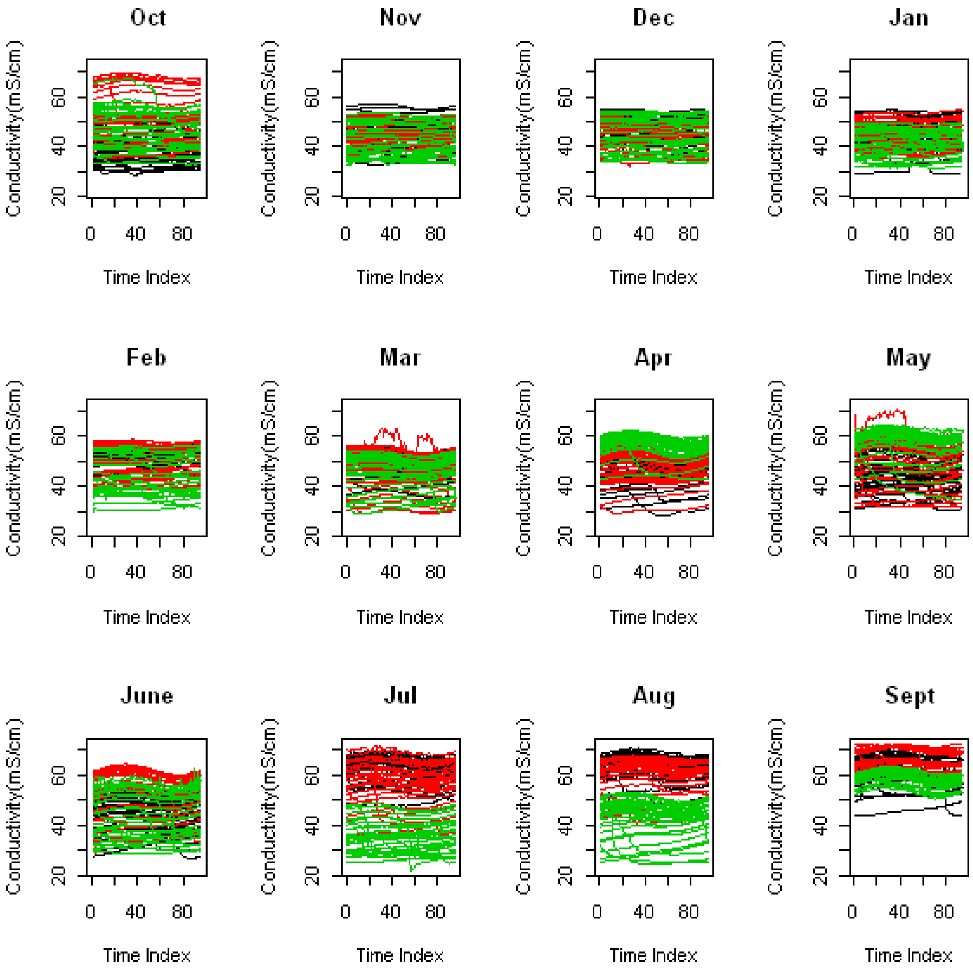

**Figure 4.5.** 15-minute measurements for conductivity in a daily period for each of the years, grouped by months, with the black, red and green lines are for the periods of 2004-2005, 2005-2006 and 2006-2007, respectively.

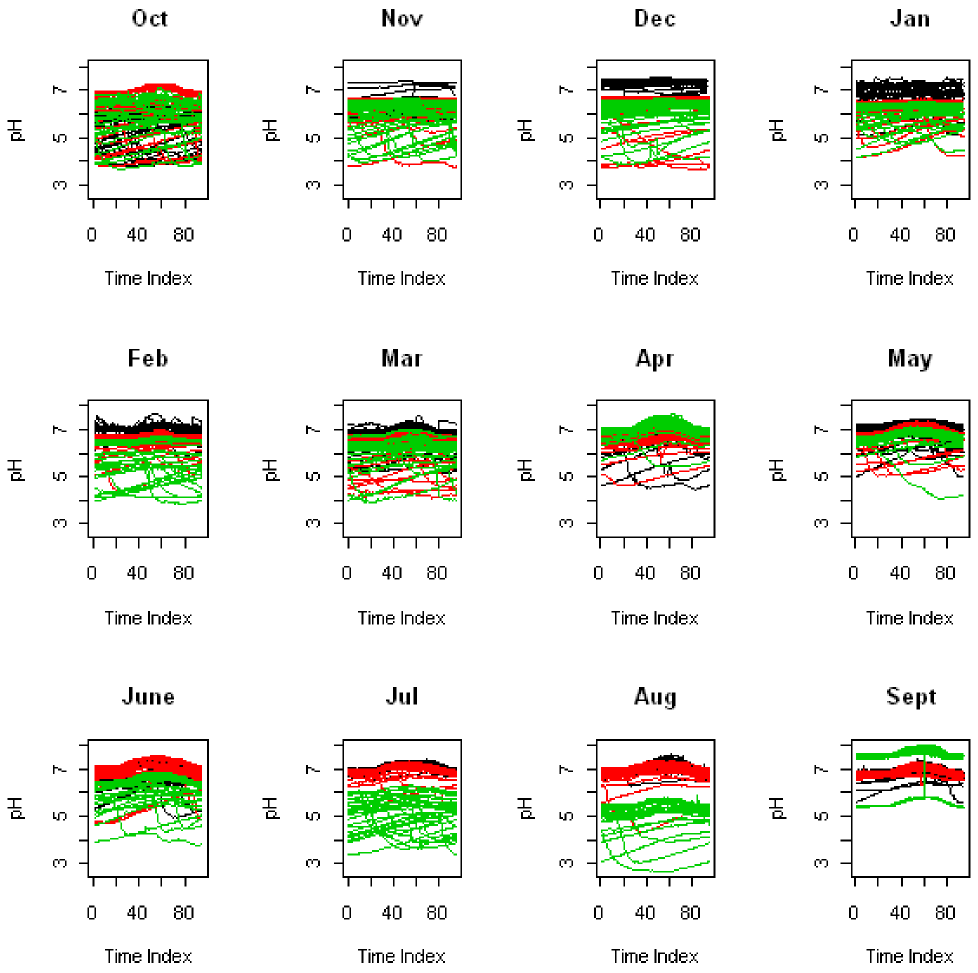

**Figure 4.6.** 15-minute measurements for pH in a daily period for each of the years, grouped by months, with the black, red and green lines are for the periods of 2004-2005, 2005-2006 and 2006-2007, respectively.

It is likely that pH and conductivity at the River Charr may exhibit similar patterns as a result of events in discharge and therefore, their relationships are explored more fully. Figure 4.7 shows the 15-minute discharge (top), conductivity (centre) and pH (bottom) measurements at the River Charr from October 2004 to Sept 2007.

A large number of missing values are apparent in discharge in a few consecutive months at the end of the period, suggesting the failure of the monitoring device to record the data in this period. A big shift in pH measurements is apparent in September 2007, which could be due to the change in the gauge during the calibration process. As a result of these two features, the time series at the River Charr will only be investigated until August 2007. The patterns for pH and conductivity are similar over the years, suggesting a close relationship between them. There is also suggestion of a response in both of these variables to events in discharge. In particular, conductivity and pH have a large number of low values following high discharge. Conversely, low discharge appears to be associated with high pH and conductivity. Higher discharge in winter is often a result of a large amount of precipitation and snow melting and hence, an exploration of such events is carried out next by exploring the data seasonally.

For illustration, Figures 4.8 - 4.9 depict plots of discharge, pH and conductivity with respect to winter and summer months over the hydrological year from October 2004 to September 2005. The magnitude and variability of high values in discharge (top) are greater in winter than summer. A number of low values for pH (centre) and conductivity (bottom) are apparent in winter following high discharge events and this feature is exhibited, for instance, in October 2004.

Moreover, there are fewer extreme events in discharge in summer with higher levels evident for conductivity and pH. Both pH and conductivity appear to behave similarly over both seasons.

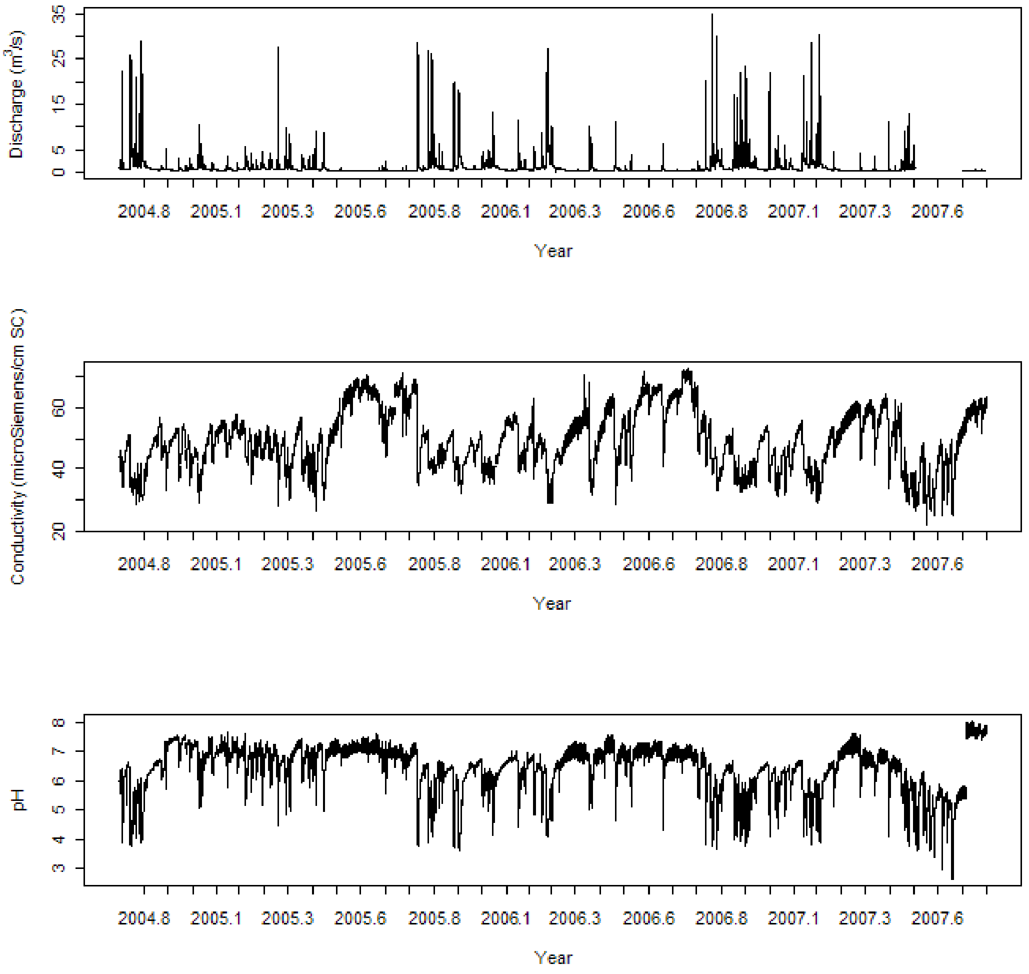

**Figure 4.7.** 15-minute measurements of discharge (top), conductivity (centre) and pH (bottom) at the River Charr from October 2004 to September 2007

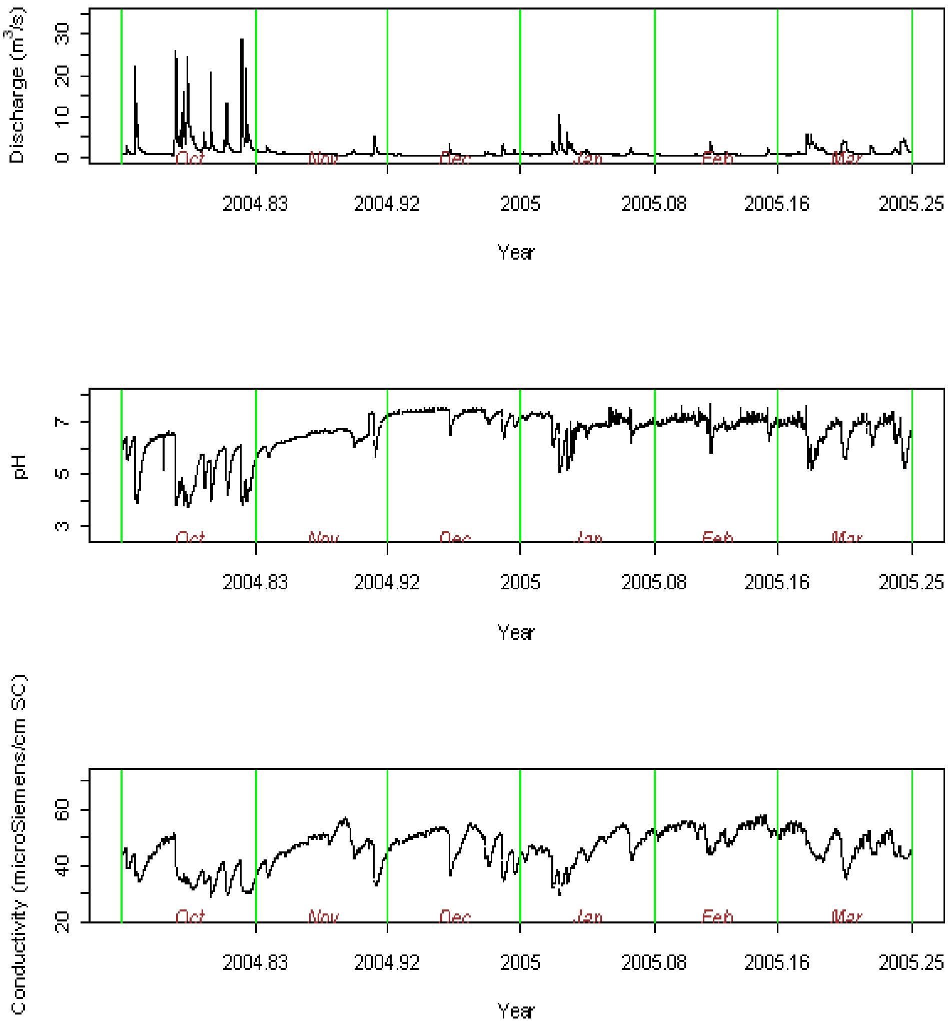

**Figure 4.8.** Plots of discharge (top), pH (centre) and conductivity (bottom) in winter over the hydrological year of October 2004 - March 2005 at the River Charr.

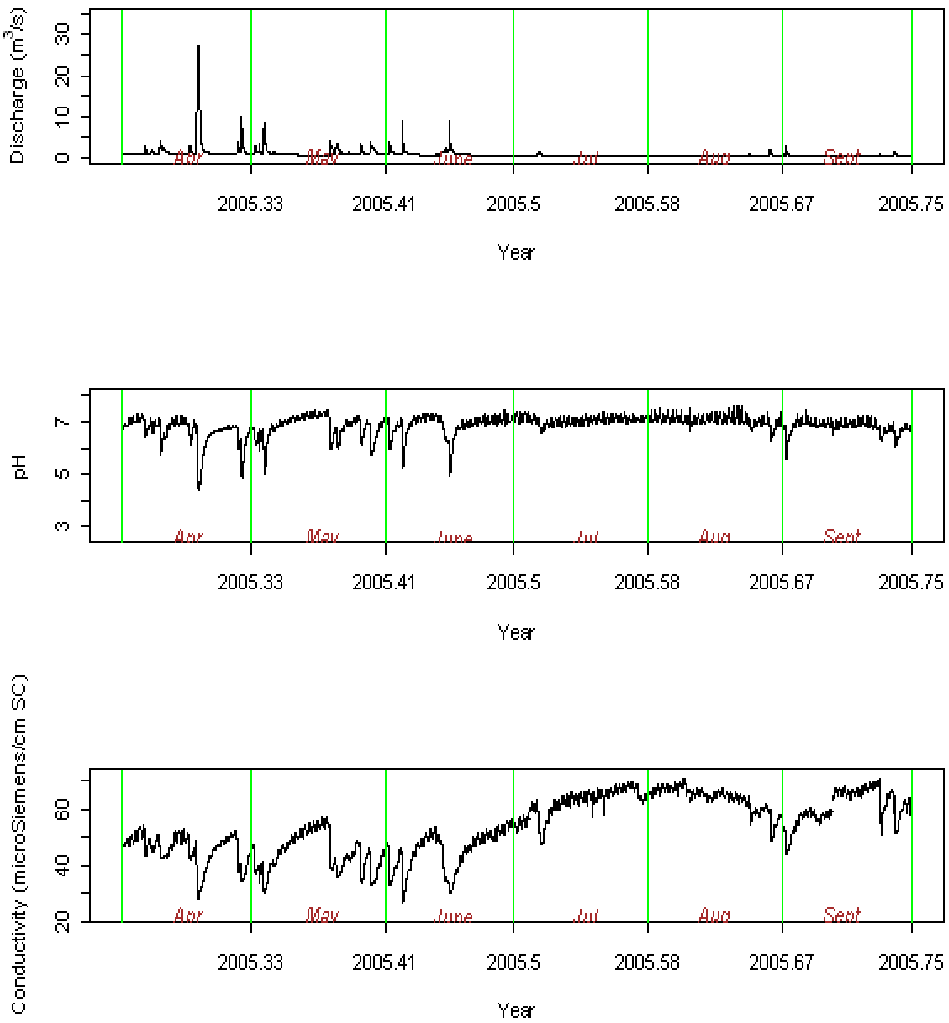

**Figure 4.9.** Plots of discharge (top), pH (centre) and conductivity (bottom) in summer over the hydrological year of April - September 2005 at the River Charr.

## 4.3 Methods

The first and second objectives in this chapter are achieved by using the wavelet technique (Schaefli et al., 2007). In particular, the wavelet power spectrum and wavelet coherence of the environmental time series are investigated. For the third objective, the recovery periods for the environmental variables following the extreme events in discharge are determined, a regression model is fitted and the extremal index is used to check the dependency of clustering in extreme discharge. Details of each of the above approaches are as follows.

### 4.3.1 Wavelet Transform

Generally, there are two classes of wavelet transform; the Continuous Wavelet Transform (CWT) and Discrete Wavelet Transform (DWT). The CWT is suitable for extracting the features of the time series whilst the Discrete Wavelet Transform (DWT) is useful for noise reduction and data compression of a time series (Torrence and Compo, 1998).

The CWT has been widely used in geophysics and meteorology to characterize the temporal variability in storms (Kumar, 1996); (Szilagyi et al., 1996) and to analyze localized variations within geophysical time series including climatic indices (Lucero and Rodriguez, 1999). It has also increasingly been used in the field of hydrology to investigate stream flow features (Smith et al., 1981), to characterize daily stream flow (Saco and Kumar, 2000), to interpret temporal patterns of different basin responses that include rapid process and slow recharges (Anctil and Coulibaly, 2004) and to examine temporal patterns in discharge (Labat and Ronchail, 2005). The CWT can also be used to analyze coherence between two geophysical and meteorological variables, particularly to identify coherent convective storm structures (Kumar, 1996); (Szilagyi et al., 1996). The wavelet coherence is used to determine the statistical persistence and the relationship of the persistence to discharge and rainfall (Shesh et al., 2010).

Firstly, the CWT of a discrete sequence of observations $x_n$ is described according to (Torrence and Compo, 1998). The CWT is defined as the convolution of $x_n$ with a scaled and translated wavelet function $\psi(\eta)$ (equation 4.1),

$$W_n^x(s) = \sum_{n^{'}=1}^{N} x_{n'} \psi^* \left[ \frac{(n^{'} - n)\delta t}{s} \right] \tag{4.1}$$

where $n^{'}$ is local point, $N$ is the number of points in the time series, $\psi(\eta)$ is the wavelet function at scale $s$ and translated in time $n$, $\delta t$ is the time step and asterisk ($*$) indicates the complex conjugate. The above transformation can also be determined using a Fourier transform (equation 4.2) due to faster calculations in Fourier space,

$$W_n^x(s) = \sum_{k=1}^{N} \hat{x}_k \hat{\psi}^*(s\omega_k) e^{i\omega_k n \delta t} \tag{4.2}$$

where $\hat{x}_k$ (equation 4.3) is the discrete Fourier transform of $x_n$ , $k$ is the frequency index $(1, 2, \ldots, N)$ and $\hat{\psi}^*$ is the Fourier transform of the wavelet function at scale $s$ and frequency $\omega_k$ .

$$\hat{x}_k = \frac{1}{N}\sum_{n=1}^{N} x_n e^{-2\pi ik\frac{n}{N}} \tag{4.3}$$

The wavelet transforms 4.1 or 4.2 are normalized as defined in equations 4.4 and 4.5, respectively, allowing a reasonable comparison of two or more time series for each scale $s$,

$$W_n^x(s) = \left(\frac{\delta t}{s}\right)^{\frac{1}{2}} \sum_{n=1}^{N} x_{n'}\psi_0\left[\frac{(n^{'}-n)\delta t}{s}\right] \tag{4.4}$$

$$W_n^x(s) = \left(\frac{2\pi s}{\delta t}\right)^{\frac{1}{2}} \sum_{k=1}^{N} \hat{x}_k\hat{\psi}_0(s\omega_k)e^{i\omega_k n\delta t} \tag{4.5}$$

where $\psi_0$ is a wavelet basis function and $\omega_k$ is the angular frequency, defined as:

$$\omega_k = \begin{cases} \dfrac{2\pi k}{N\delta t}; k \leq \dfrac{N}{2} \\ \\ -\dfrac{2\pi k}{N\delta t}; k > \dfrac{N}{2} \end{cases}$$

A wavelet function $\psi(\eta)$ is a signal that has zero mean and can be localized in both time and frequency space (Farge (1992); Misiti et al. (1996)). Torrence and Compo (1998) stated that the types of chosen wavelet functions rely on the objectives of the analysis and the nature of the time series, orthogonality of the basis and features of width and shape as follows.

Information on the amplitude and phase of the time series data are provided by a complex wavelet function which is suitable to capture the oscillatory characteristics whilst a real wavelet function can distinguish the features of peaks of a time series data (Torrence and Compo, 1998).

The wavelet function with an orthogonal basis can be used with a discrete wavelet transform whilst a nonorthogonal basis is appropriate for either the discrete or continuous wavelet transform (Farge, 1992). In orthogonal wavelet transforms, the number of convolutions at each scale is proportional to the width of the wavelet basis at that scale, leading to discrete blocks of wavelet power which is useful for signal processing (Torrence and Compo, 1998). In contrast, a nonorthogonal wavelet analysis is highly redundant at large scales in which the wavelet spectrum at adjacent times is highly correlated. Such a wavelet transformation is useful if the wavelet spectrum has continuous variations (Torrence and Compo, 1998).

The shape of the wavelet function determines the characteristic of the time series and the width of the function determines whether a good resolution in time or frequency is produced. A broad wavelet function provides a good frequency resolution but poor resolution in time and vice versa (Torrence and Compo, 1998).

The work throughout this chapter uses a complex Morlet function in the CWT. This basis function consists of a plane wave modulated by a Gaussian process as defined in equation 4.6,

$$\psi(\eta) = \pi^{-\frac{1}{4}} e^{i\omega_0 \eta} e^{-\frac{\eta^2}{2}} \tag{4.6}$$

where $\omega_0$ is dimensionless frequency and $\eta$ is dimensionless time, depending on the time and scale of the data (Torrence and Compo, 1998). The Morlet wavelet with $\omega_0 = 6$ provides a good balance between time and frequency localization as it can describe the shape of time series data more clearly than any other wavelet functions such as the Mexican wavelet function and classical sine function (Labat et al., 2000). For a given $\omega_0 = 6$, the Morlet wavelet scale is similar to the Fourier period and so, they produce similar values. This is due to the fact that the Fourier period ($\lambda_{wt}$) is similar to the scale ($\lambda_{wt} = 1.03s$) (Grinsted et al., 2004).

### 4.3.2 Wavelet Power Spectrum

The wavelet power spectrum is used to investigate any temporal patterns for each environmental determinant. The wavelet power spectrum is defined as the square of the amplitude of the wavelet transform $|W_n^x(s)|^2$. Since the Morlet wavelet function is a complex number $z$ of real ($a$) and imaginary ($bi$) parts, where $z = a + bi$, the square of the absolute value of $z$ is defined as the product of $z$ and its complex conjugate $z^*$ as follow,

$$(zz^* = (a + bi)(a - bi) = a^2 + b^2)$$

Hence, the wavelet power spectrum is defined by equation 4.7,

$$W_n^{xx}(s) = W_n^x(s)W_n^{x*}(s) = |W_x^n(s)|^2 \tag{4.7}$$

producing the variance in the time series at a given scale $s$ and time $n$.

The statistical significance of the wavelet power spectrum can be assessed by comparing the spectrum with a background spectrum (noise). The structure of the background spectrum depends on the nature of the time series. For instance, the background spectrum in geophysical processes is often denoted by either white noise (constant variance across all scales or frequency) or red noise (decreasing variance with decreasing scale or increasing frequency), Torrence and Compo (1998). However, Grinsted et al. (2004) shows that an AR(1) model is representative for most of the background spectrum in a geophysical time series.

The mean Fourier power of the background spectrum is defined as :

$$P_k = \frac{1-\alpha^2}{1+\alpha^2-2\alpha\cos(\frac{2\pi k}{N})}$$

where $\alpha$ is the lag-1 autocorrelation for the time series (Torrence and Compo, 1998). The wavelet power spectrum of the time series is significant if it exceeds the above background spectrum at a particular significance level.

Torrence and Compo (1998) introduce a pointwise significance test for the wavelet power spectrum by assuming the background spectrum as a red noise for the Null hypothesis and the significance testing is carried out for each time-scale. However, pointwise testing has several disadvantages that leads to false results (Schaefli et al., 2007).

Firstly, a repetition of the test for $N$ wavelet coefficients provides an average of

$\alpha N$ false positive results (multiple testing effect) for a given significance level of $1 - \alpha$. Secondly, it is highly likely that there is a correlation of the neighbouring wavelet coefficients (intrinsic correlation effect), which appear in a contiguous patch and so, the occurrence of false positive results could be observed since the correlation issue in the wavelet power spectrum is not taken into consideration. The effect of intrinsic correlation appears in every time-frequency analysis and represents a time-scale uncertainty. The above effects leads to an unclear significant patch, since the observed patch might be reflected by the actual physical characteristics or just simply result of a multiple testing effects and intrinsic correlation of the wavelet coefficients. However, the above problems in pointwise significance testing of the wavelet power spectrum are overcome using an area-wise significance testing Maraun et al. (2007) as follows.

For the CWT with a Gaussian white noise $\psi(\eta)$, the intrinsic correlation between two wavelet coefficients at $(n, s)$ and $(n^{'}, s^{'})$ is given by a kernel $K_{\psi,\psi}((n - n^{'})/s^{'}, s/s^{'})$ moved to the time $n^{'}$ and stretched to the scale $s^{'}$ (Schaefli et al., 2007) as defined in equation 4.8,

$$C(n, s; n^{'}, s^{'}) \sim K_{\psi,\psi}\left(\frac{n - n^{'}}{s^{'}}, \frac{s}{s^{'}}\right) \tag{4.8}$$

Since the intrinsic correlations are given by the above kernel function, the patch areas for random fluctuations are also given by such a kernel and so, any patches from the pointwise test that are smaller than the above reproducing kernel are assumed to be similar to the noise (Maraun et al., 2007).

Given a set of patches with pointwise significant values $P_{pw}$, the areawise significant patches Schaefli et al. (2007) are defined as follows.

For each $(s, n)$, a critical area $P_{crit}(s, n)$ is given by the reproducing kernel at $(s, n)$ that exceeds a particular critical levels $K_{crit}$ (equation 4.9),

$$P_{crit}(s,n) = \left\{(s^{'}, n^{'})|(K(s,n;s^{'},n^{'}) > K_{crit})\right\} \tag{4.9}$$

where the critical area represents the size of the reproducing kernel for a given significance level $K_{crit}$.

The subset of several areawise significant wavelet spectral coefficients is given by the incorporation of all critical areas that lie completely inside the patches of pointwise significant values (equation 4.10).

$$P_{aw} = \bigcup_{P_{crit}(s,n) \subset P_{pw}} P_{crit}(s,n) \tag{4.10}$$

The above areawise testing is summarized as follows:

- A pointwise testing is performed on the significance level 1-$\alpha$.
- A significance level for the reproducing kernel 4.8, 1-$\alpha_{aw}$ for the areawise testing and the corresponding critical area $P_{crit}(s, n)$ of the reproducing kernel are chosen.
- The evidence of the wavelet power spectrum for each $(s, n)$ is determined by comparing with the critical area $P_{crit}(s, n)$. A point inside a patch is defined

as areawise significant, if any critical area containing this point totally lies within the patch.

The above areawise significance testing may highlight the significant patches that exceed the noise level, resulting from the occurrence of extreme events Schaefli et al. (2007).

Here, the wavelet power spectrum for each environmental variable is investigated at short time scales with the use of `wsp` function in `sowas` library (Maraun et al., 2007).

### 4.3.3 Wavelet Coherence

In wavelet analysis, the coherence of two signals of two different time series can be determined by the use of wavelet coherence as follows.

Let $x_n$ and $y_n$ be two time series of pH and conductivity, respectively. The similarity of the pattern of the two processes can be identified using cross wavelet transform 4.11, providing the covariance of $x_n$ and $y_n$.

$$W_n^{xy}(s) = W_n^{x}(s) W_n^{y*}(s) = |W_n^{xy}(s)| \tag{4.11}$$

Another useful property of the cross wavelet spectrum is the phase angle (equation 4.12), defined as the local relative phase between $x_n$ and $y_n$,

$$\theta_n^{xy}(s) = Tan^{-1}\left(\frac{\Im(W_n^{xy}(s))}{\Re(W_n^{xy}(s))}\right) \tag{4.12}$$

where $\Im(W_n^{xy}(s))$ and $\Re(W_n^{xy}(s))$ are the imaginary and real parts of the cross wavelet spectrum, respectively.

The two time series $x_n$ and $y_n$ are assumed to have independent power at overlapping time and scale intervals, indicating no covariance between the two power spectra. Hence, the information about one of the processes is not sufficient for predicting the other process. However, the real wavelet cross spectrum can always be different from zero. Any significant peaks can occur not only in the case of covarying power but also if either one or both of the individual wavelet spectra show strong power (Schaefli et al., 2007) and so, the wavelet coherence which produce a normalized measure is more appropriate.

The wavelet coherence (equation 4.13) is defined as the square of the cross wavelet spectrum normalized by the individual power spectrum from each of the time series $x_n$ and $y_n$.

$$\frac{|W_n^{xy}(s)|^2}{|W_n^x(s)|^2|W_n^y(s)|^2} \tag{4.13}$$

Wavelet coherence can highlight high common power of two time series at particular scales and time points. In particular, a common oscillation between the two signals which have a rather stable phase difference (Maraun et al., 2007) can be revealed. The statistical significance of the wavelet coherence can be assessed by using the pointwise and areawise significance testing, however, the areawise testing is preferred due to the previous advantages in the wavelet power spectrum.

Significant coherence is identified based on a 90% confidence level of areawise testing. The cross phase angle of the wavelet coherence is a useful property to highlight any phase distinctions between peaks of two signals of two processes at particular time points and scales. Positive and negative differences are often referred to as in and out of phase, respectively. The steps of areawise testing is exactly the same as in the previous wavelet power spectrum but the only different is the critical patch-size $P_{crit}$ that need to be re-estimated in the areawise testing (Maraun et al., 2007).

Here, the wavelet coherence between pH and conductivity is investigated at each time point and short time scale using `wco` function in `sowas` library (Maraun et al., 2007).

### 4.3.4 The Determination of Recovery Period

The recovery period is the duration of time taken by pH and conductivity to recover following high river flows. Specifically, the recovery period for pH and conductivity in rivers begins from the time point where the measurements start to decrease in response to an event in discharge, until the values return to the pre-event levels. This period is determined using the following approach.

1. Discharge measurements are sorted from highest to lowest and their cumulative distribution function is determined corresponding to a particular percentage of extreme discharge. For instance, the threshold for discharge corresponding to 2% of the complement for cumulative probability distribution is defined as $1 - p(X \leq x_i) = 0.02$.

2. All of the times which are at 2 hours before the extreme discharge events $i$ are determined where $i = 1, 2, \ldots, p$. Let $t_i = t_1, t_2, \ldots, t_p$ be the times corresponding to 2 hours before the extreme events and $X_{t_i}$ are conductivity and pH measurements corresponding to time $t_i$.

3. The first baseline $(X_{t_1})$ for conductivity and pH is chosen corresponding to $t_1$ and the time point at the end of the recovery period corresponds to the time at which the level of conductivity and pH returns to the baseline level $X_{t_1}$.

4. Each of the conductivity and pH measurements after time $t_1$ is subtracted from $X_{t_1}$ and the sign of conductivity and pH measurements (+ or -) highlight the positions of the measurements, above or below the baseline, respectively. Negative values are of interest here as they highlight the occurrence of measurements below the baseline. The first positive value would indicate the time point of recovery and end of the process. The recovery period is determined by the difference between the end and initial time points. The unit of the recovery period is standardized by converting into a daily scale.

5. The subsequent baselines are determined by repeating steps 3-4 at the remaining times $t_2, t_3, \ldots, t_p$ and the corresponding recovery periods are estimated.

The same approach is then applied on pH and conductivity at the River Drumtee to check for similarity of the recovery in summer and winter. Since the baselines of the recovery period for pH and conductivity rely on the threshold for extreme discharge, three levels of threshold for extreme discharge for both rivers are chosen. The first threshold is chosen at considerably high discharge during storm events as defined in the earlier part of this chapter. The second threshold is chosen at moderately high discharge and finally, the third threshold is determined at lower discharge. The first, second and third thresholds are 2%, 3% and 10% of the extreme discharge, respectively.

### 4.3.5 Regression Model

The recovery period for conductivity and pH could be influenced by particular features in their own time series as well as in discharge and so, six potential explanatory variables have been identified. They are baseline of the conductivity and pH, maximum value of discharge within recovery period, minimum conductivity and pH within recovery period, rate of change (slope) of the conductivity and pH from the baseline time point to the minimum value within the recovery period, number of extreme event in discharge within recovery period and the area encompassed by extreme discharge within the recovery period.

The recovery period, area of extreme discharge, rate of change of pH and conductivity and number of extreme discharge are log transformed to stabilize the variance. For illustration, Figure 4.10 shows the relationships of log recovery for pH with each of the previously identified predictors at the River Charr by season, subject to a 2% threshold for discharge. The circle and cross symbols represent winter and summer, respectively. No clear differences are shown between season in terms of the relationships of log recovery with each of the explanatory variables. A strong and positive linear relationship of each log number extreme discharge and log area of the extreme discharge with log recovery is highlighted whilst a weak linear relationship is shown between log recovery and log rate of change for pH. Conversely, minimum pH has a negative linear relationship with log recovery. Baseline seems to have a weak linear relationship with log recovery for pH. The relationship between log recovery for pH and maximum discharge appears to possibly follow a curvilinear pattern.

The same potential explanatory variables are used for each river. Models of recovery period for pH and conductivity for the River Charr and River Drumtee are fitted and they are compared for the identification of any similar significant predictors. The details of the modelling are as follows,

A regression model is used to model log recovery using all previously identified predictors as well as the interaction between season and these predictors. Models of log recovery at the River Charr, defined in equations 4.14 and 4.15, are initially fitted and a comparison between them is carried out via the F-test. The full model 4.14 incorporates a quadratic term in log(max extreme discharge) as a curvilinear pattern is highlighted by the previous plots of relationships, whilst the remaining predictors are assumed to follow linear relationships. The reduced model 4.15, on the other hand, assumes linearity in all the predictors. The seasonal factors are winter and summer, denoted by 0 and 1, respectively.

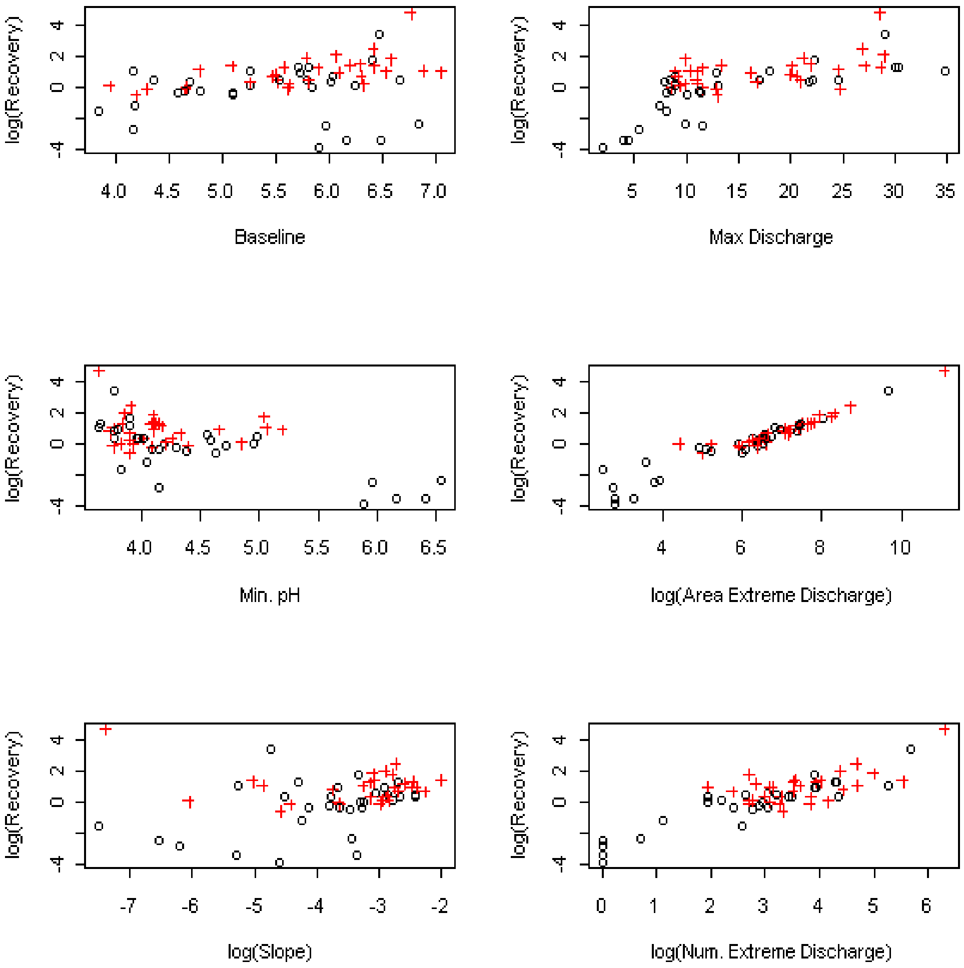

**Figure 4.10.** Plots of relationship between log recovery for pH with each of the predictors at the River Charr, subject to 2% threshold for discharge.

$$\begin{aligned} log(recovery)_i &= \beta_0 + \beta_1\{baseline\}_i + \beta_2\{maxdis\}_i + \beta_3\{maxdis\}_i^2 \\ &+ \beta_4\{min\}_i + \beta_5\{log(slope)\}_i + \beta_6\{log(areadis)\}_i \\ &+ \beta_7\{log(numextremedis)\}_i + \beta_8\{season\}_i \\ &+ \beta_9\{baseline\}_i * \{season\}_i + \beta_{10}\{maxdis\}_i * \{season\}_i \\ &+ \beta_{11}\{min\}_i * \{season\}_i + \beta_{12}\{log(slope)\}_i * \{season\}_i \\ &+ \beta_{13}\{log(areadis)\}_i * \{season\}_i \\ &+ \beta_{14}\{log(numextremedis)\}_i * \{season\}_i + \epsilon_i \qquad (4.14) \end{aligned}$$

$$\begin{aligned}
log(recovery)_i &= \beta_0 + \beta_1\{baseline\}_i + \beta_2\{maxdis\}_i + \beta_3\{min\}_i \\
&+\beta_4\{log(slope)\}_i + \beta_5\{log(areadis)\}_i \\
&+\beta_6\{log(numextremedis)\}_i + \beta_7\{season\}_i \\
&+\beta_8\{baseline\}_i * \{season\}_i + \beta_9\{maxdis\}_i * \{season\}_i \\
&+\beta_{10}\{min\}_i * \{season\}_i + \beta_{11}\{log(slope)\}_i * \{season\}_i \\
&+\beta_{12}\{log(areadis)\}_i * \{season\}_i \\
&+\beta_{13}\{log(numextremedis)\}_i * \{season\}_i + \epsilon_i \qquad (4.15)
\end{aligned}$$

Diagnostic plots on the residuals are checked to determine that the residuals $\epsilon_i$ are independent and normally distributed with zero mean and constant variance.

The above approach is repeated by using different levels of threshold for discharge at 3% and 10% for both conductivity and pH.

Additionally, models of recovery for pH and conductivity measurements at the River Drumtee are fitted, subject to the same threshold levels of discharge as at the River Charr. The two fitted models from the two locations for conductivity/pH are then compared to determine any similarities or differences in the significant predictors.

### 4.3.6 Model Validation

The models for pH and conductivity for the River Charr are validated by using the available measurements from January - October 2008. Model validation is carried out to see how well the model performs on future data. The recovery periods for both pH and conductivity from January - September 2008 are used and the recent fitted models are validated using mean square error, i.e. the sum of the difference between the estimates from the fitted model ($\hat{y}_i$) and actual recovery period ($y_i$) relative to the number of recovery period ($n$) as $\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}$.

### 4.3.7 Extremal Index

It is of interest to determine the features of extreme discharge. In particular, the independence extreme discharge is investigated and the corresponding thresholds could be identified since clustering of the extreme events in discharge may vary corresponding to different thresholds. Extreme events in discharge are often clustered together and so, the independence of the exceedences is likely to be violated (Ferro, 2003).

Previous work carried out by several researchers show some development of the approaches used for dealing with the issue of independence of extreme values as follows. Hsing (1987) considered the use of clustering and shows that clusters of extreme values could be independent within a particular limit. Davison and Smith (1990) characterized cluster maximum where a peaks-over-threshold approach was used to determine the independence of the extreme values prior to fitting a model of extreme events. Smith and Weissman (1994) and Weissman and Novaks (1998) developed the estimation of the extremal index in order to determine the declustering of exceedences.

The extreme discharge is therefore investigated for evidence of any signs of clusters throughout the years. The evidence of clustered extreme discharge for each of the thresholds is determined by the use of an extremal index in order to measure the level of clustering (Ferro, 2003). However, the thresholds corresponding to the declustering of extreme discharge here are not necessarily the same thresholds used in the previous model fitting on log recovery for pH and conductivity.

Ferro (2003) defines the extremal index as follows. Let $\xi_1, \xi_2, \ldots, \xi_n$ be stationary sequence random variables that have a marginal distribution $F$, a right end point $\omega = sup\{F(x) < 1\}$ and a tail function $\bar{F} = 1 - F$. If $M_{k,l}$ is defined as $max\{\xi_i : i = k+1, 2, \ldots, l\}$ for integers $0 \leq k \leq l$, then $\xi_1, \xi_2, \ldots, \xi_n$ has extremal index $\theta$ that lies between 0 and 1 if for every $\tau > 0$, a sequence of $u_1, u_2, \ldots, u_n$ exists such that as $n \to \infty$

- $n\bar{F}(u_n) \to \tau$
- $P(M_{0,n} \leq u_n) \to e^{-\theta\tau}$

Leadbetter et al. (1983) shows that there is no evidence of clustering in extreme data if the extremal index $\theta = 1$ while the exceedences above threshold tend to cluster as $\theta < 1$. $F$ is essentially hard to determine for the real extreme events and alternatively, the following approach is used to obtain the index. The estimation of $\theta$ involves choosing a threshold $u$ as in the Pareto distribution and defining $N$ as:

$$N = \sum_{i=1}^{n} I(\xi_i > u)$$

where $\xi_1, \xi_2, \ldots, \xi_n$ is the extreme discharge above threshold $u$. N is the number of exceedences of $u$ and $1 \leq S_1 < \ldots < S_N \leq n$ are the exceedence times. The interexceedence times $T_i$ are defined as $S_{i+1} - S_i$ where $i = 1, 2, \ldots, N-1$.

Ferro and Segers (2002) shows that two estimates of extremal index $\hat{\theta}$ can be produced, subject to the conditions of whether the interexceedence time $T_i$ is above or below than 2 (equation 4.16).

$$\tilde{\theta}_n(u) = \begin{cases} 1 \wedge \hat{\theta}_n(u); max\,\{T_i : 1 \leq i \leq N-1\} \leq 2, \\ 1 \wedge \hat{\theta}^*_n(u); max\,\{T_i : 1 \leq i \leq N-1\} > 2, \end{cases} \tag{4.16}$$

where the estimates of the extremal index $\hat{\theta}_n(u)$ and $\hat{\theta}^*_n(u)$ are defined in equations 4.17 and 4.18, respectively.

$$\hat{\theta}_n(u) = \frac{2\left(\sum_{i=1}^{N-1} T_i\right)^2}{(N-1)\sum_{i=1}^{N-1} T_i^2} \tag{4.17}$$

$$\hat{\theta}^*_n(u) = \frac{2\left(\sum_{i=1}^{N-1} (T_i - 1)\right)^2}{(N-1)\sum_{i=1}^{N-1} (T_i - 1)(T_i - 2)} \tag{4.18}$$

The confidence interval for the extremal index is then determined by using a bootstrap sampling method where the quartiles of 2.5% and 97.5% of the 1000 estimates of an extremal index represent the lower and upper of 95% confidence limits. The extremal index corresponding to a series of thresholds from sufficiently low to high discharge measurements are computed and the confidence interval corresponding to each of extremal index is determined. The `exi` function in the `evd` library in R (Ferro, 2003), is used to estimate the extremal index.

## 4.4 Results

### 4.4.1 Temporal Variability in Temperature, Barometric Pressure, pH and Conductivity

The scalograms of wavelet power spectrum for the differenced temperature, barometric pressure, pH and conductivity values from October 2004 - September 2007 are portrayed in Figure 4.11. The dubious contiguous patches from lower up to considerably higher time scales are more likely to be attributed to the occurrence of extreme events.

The temporal pattern for temperature at a 1 day scale is clearly exhibited in the warmer months over the year, showing an evidence of a diurnal cycle in summer. The conductivity and pH show similar temporal patterns as for temperature over warmer months. However, the occurrence of noise denoted by several continuous dubious patches from the lower up to higher scales at a given time point restricts a clear presentation of the significant diurnal cycle. The small patches in barometric pressure are constantly highlighted at 1-2 day scales over the year, suggesting that the air pressure is not as seasonally dependant as for temperature, pH and conductivity.

A large number of spurious patches within the three years results in an unclear power spectrum and so, the evidence of the individual temporal patterns is investigated over a smaller time period. Here, the power spectrum for the above variables are investigated by year.

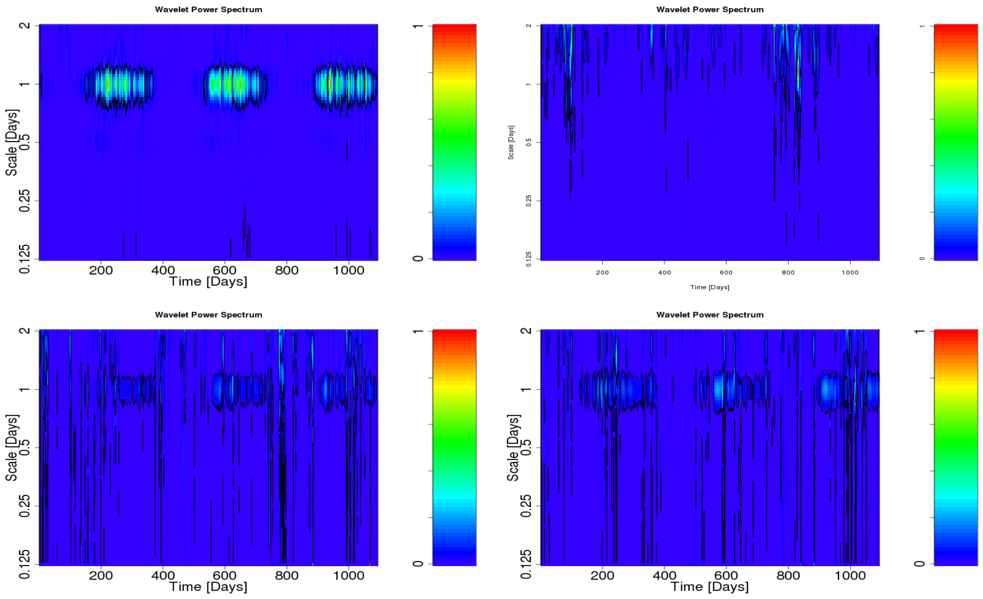

**Figure 4.11.** Wavelet Power Spectrum of temperature (top left), barometric pressure (top right), pH (bottom left) and conductivity (bottom right) from October 2004 - September 2007, with patches to identify areawise significance.

The scalograms of wavelet power spectrum for temperature, barometric pressure, pH and conductivity from October 2004 - September 2005, October 2005 - September 2006 and October 2006 - September 2007 are shown from Figures 4.12 - 4.14, respectively. There are several small significant patches in pH and conductivity that are connected by thin bridges, lying from lower to higher scales. Schaefli et al. (2007) explain about these type of patches and have defined it as spurious power spectrum and so, they are not studied further here.

The temporal patterns highlighted from the wavelet power spectrum for temperature, pH, conductivity and barometric pressures over each of the years are more apparent than the previous 3 years. There is evidence of a daily temporal pattern in temperature, pH and conductivity in warmer months as exhibited by large significant patches at a 1 day scale over the year. Barometric pressure shows evidence of the temporal patterns at a 1-2 days scale over each of the hydrological years, with a strong temporal pattern shown clearly in each of the years. In particular, strong evidence of a temporal pattern for barometric pressure is constantly occurring from October 2004 - September 2005, however in the later years, the strong temporal patterns are more apparent in colder months.

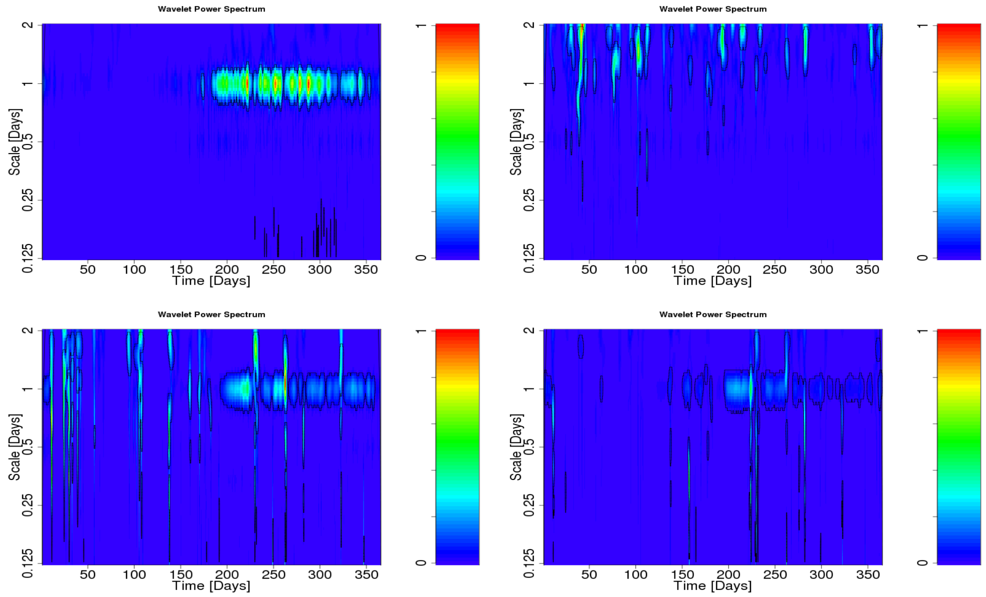

**Figure 4.12.** Wavelet Power Spectrum of temperature (top left), barometric pressure (top right), pH (bottom left) and conductivity (bottom right) from October 2004 - September 2005, with patches to identify areawise significance.

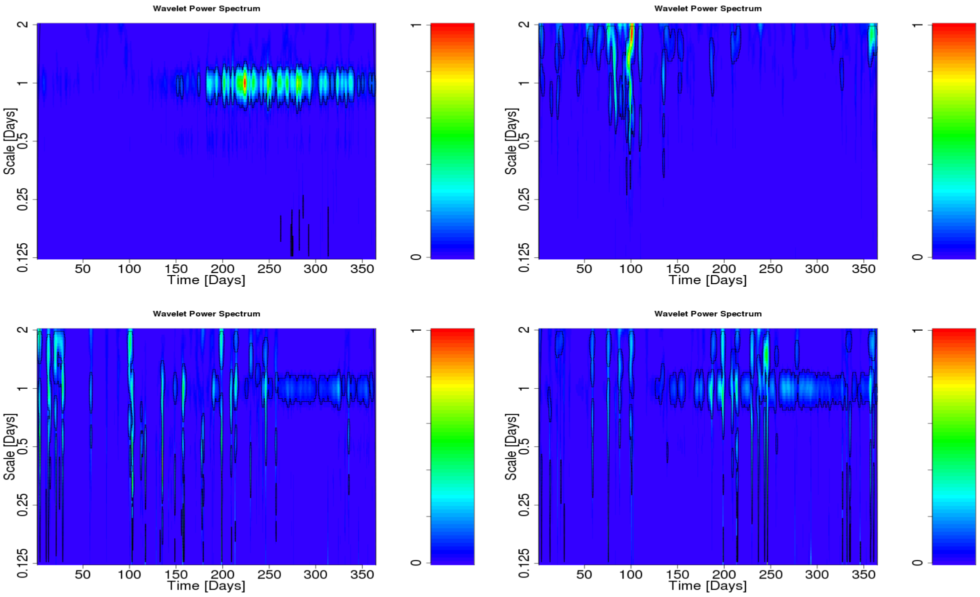

**Figure 4.13.** Wavelet Power Spectrum of temperature (top left), barometric pressure (top right), pH (bottom left) and conductivity (bottom right) from October 2005 - September 2006, with patches to identify areawise significance.

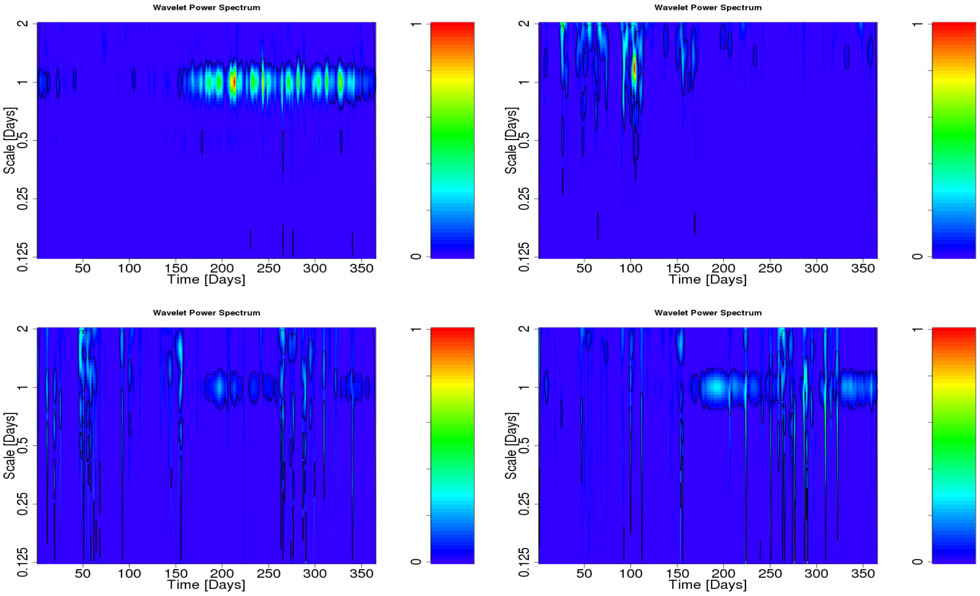

**Figure 4.14.** Wavelet Power Spectrum of temperature (top left), barometric pressure (top right), pH (bottom left) and conductivity (bottom right) from October 2006 - September 2007, with patches to identify areawise significance.

### 4.4.2 Relationship between pH and Conductivity

Similar features in pH and conductivity are noticeable from the previous temporal patterns across the year. High pH may occur along with high conductivity and vice versa, suggesting a relationship between them. Here, the post-differencing pH and conductivity measurements are used and the similarity of their oscillations is further investigated via the wavelet coherence.

Similar results are highlighted in each of the hydrological years and therefore, only one of the hydrological years is chosen and presented here. Figure 4.15 shows the wavelet coherence (top) and cross phase angle (bottom) between pH and conductivity from October 2004 - September 2005. Patches, representing the significant coherence between pH and conductivity are apparent at particular time points and scales.

There are numerous small patches at a scale of less than 24 hours in both seasons. According to Schaefli et al. (2007), these small areas of significant coherence at lower scales is a result of the occurrence of two extreme events of the two processes, coinciding and exceeding the noise level (assumed to be Gaussian). These patches probably do not reflect the actual coherency of pH and conductivity but tend to be coincidence of extreme events of both processes. In fact, the occurrence of many extreme events (low values) for both pH and conductivity as depicted from the exploratory plots (Figure 4.1) are apparent in both seasons of the year, suggesting that significant coherence at a scale of less than one day is contributed by such events.

At a daily scale, the significant coherence is clearly shown by a mixture of patches over the year. The significant coherence denoted by large patches constantly appear in summer, particularly from June to August 2005 (from about 273 to 335 days). The significant coherence is likely due to the fact that the variability of pH and conductivity are approximately constant as they are not significantly influenced by the hydrological events and so, the signals of the two processes are similar.

The significant coherence of the two processes at a higher scale of 2-3 days is apparent and appears to be constant over the year. There are few small patches in both seasons while a relatively large patch is apparent in December 2004 (approximately from 70 - 92 days) and May - June 2005 (about 220 - 243 days).

Out of phase signals of about $90^o$ is apparent during a 24-hour period in summer months but no clear phase difference is exhibited at scales of 2-3 days, indicating that the peaks of both signals are at similar time points.

For illustration, Figure 4.16 shows plots of pH (top left) and conductivity (top right) in July 2005. The snapshots of pH (bottom left) and conductivity (bottom right) on $25th$ July ($298th$ day) are displayed. The displayed feature indicates that peaks of pH lead conductivity by about 6 hours.

The significant coherence at a 1 and 2 days scale over the summer and both seasons, respectively, can be distinguished by a particular feature of the signals of pH and conductivity. The cross phase angle presents the differences of such a characteristic of the two signals at different scales. In particular, the evidence of coherence at a 1 day scale is resulting from the similarity of the two signals. The variable pH responds more quickly than conductivity at this scale but at a 2 day scale, both pH and conductivity respond similarly.

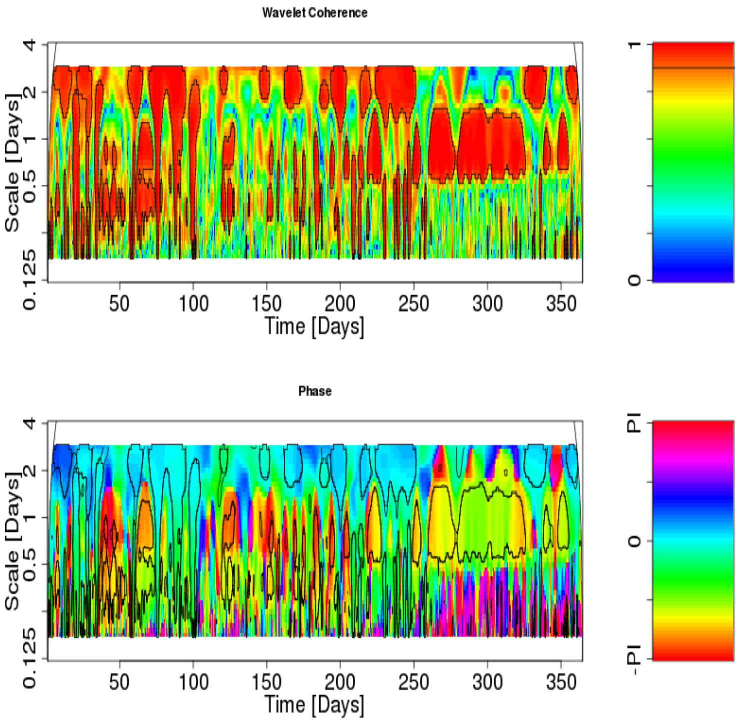

**Figure 4.15.** Wavelet coherence (top) and cross phase angle (bottom) between the 15-minute measurements of pH and conductivity from Oct 2004 - Sept 2005, with patches to identify the areawise significance

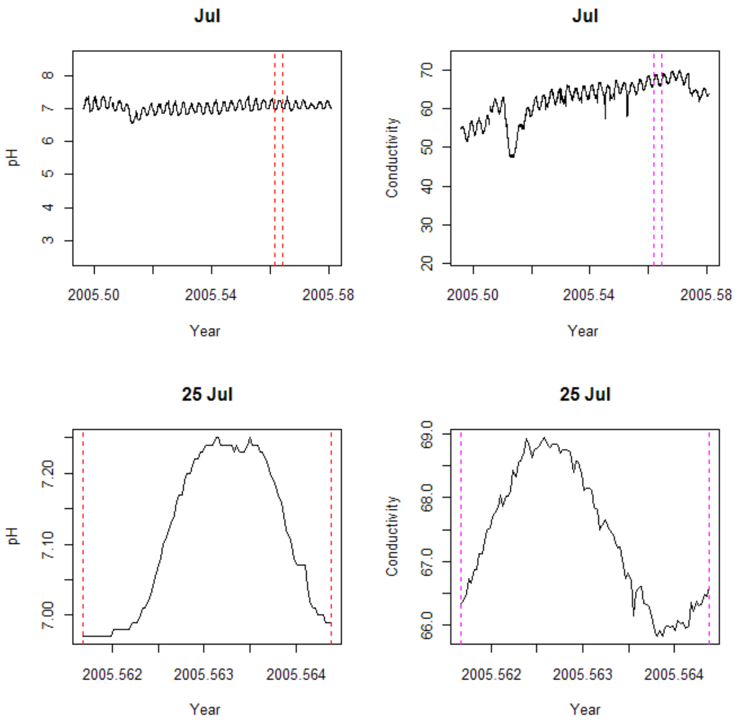

**Figure 4.16.** Plots of pH (top left) and conductivity (top right) in July 2005; pH (bottom left) and conductivity (bottom right) on 25 July 2005, with patches to identify areawise significance.

### 4.4.3 Recovery Periods for pH and Conductivity

Table 4.1 shows the values of discharge($m^3s^{-1}$) corresponding to 2%, 3% and 10% of the complement of the cumulative probability distribution for both rivers. The values of extreme discharge corresponding to each of the above complement are as follows.

| Threshold(%) | River Charr | River Drumtee |
|---|---|---|
| 2 | 7.20 | 1.62 |
| 3 | 5.78 | 1.37 |
| 10 | 3.00 | 0.63 |

**Table 4.1.** Discharge ($m^3s^{-1}$) corresponding to 2%, 3% and 10% thereshold levels for Charr and Drumtee.

For illustration, the above thresholds for discharge at the River Charr and River Drumtee are highlighted in Figure 4.17. The blue, green and red horizontal lines across the year are 2%, 3% and 10% threshold levels.

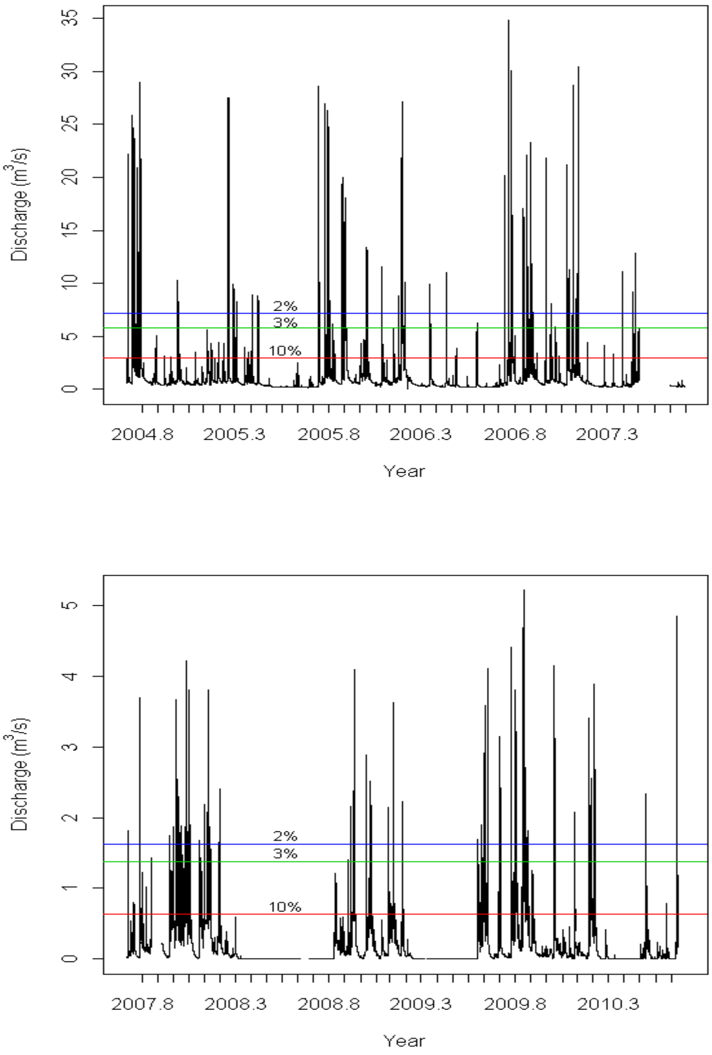

**Figure 4.17.** Time series plots of 15 minutes discharge measurements at Charr (top) and Drumtee (bottom), with the blue, green and red horizontal lines denote 2%, 3% and 10% of the threshold levels, respectively.

The recovery period for pH and conductivity for both rivers corresponding to each of the 3 levels of threshold is determined. For illustration, the recovery period for pH and conductivity from Charr and Drumtee over one hydrological year are presented. Plots of discharge, pH and conductivity at the River Charr in winter and summer from October 2004 - September 2005 are shown in Figures 4.18 and 4.19, respectively. Blue and red vertical dashed lines denote the initial and end time points of each of the recovery periods. The recovery periods corresponding to a threshold of $7.2m^3s^{-1}$ (2%) for discharge are marked by blue horizontal lines.

Figures 4.20 and 4.21 depict plots of discharge, pH and conductivity measurements at the River Drumtee over two different seasons from October 2007 - September 2008. The recovery period corresponding to a threshold of $1.62m^3s^{-1}$ (2%) for discharge is highlighted. A continuous period of missing values from mid November until December 2007 is clearly exhibited in the three determinands.

The number of recovery periods in winter is larger than in summer for both rivers. This feature is expected as extreme events in discharge (high values) are more apparent in winter than summer and as a result, more extreme events in pH and conductivity (lower values) are observed in colder months. High discharge and low pH and conductivity between the two rivers are comparable in winter. The distinction of the levels is due to the fact that the water flow at the River Charr is likely to be a result of snow melting and water runoff from a nearby mountainous area and so, higher discharge is clearly observed compared to the River Drumtee. Lower river flow in Whitelee, however, is due to less influence of the above two factors, resulting in slightly lower conductivity and pH at the River Charr compared to the River Drumtee.

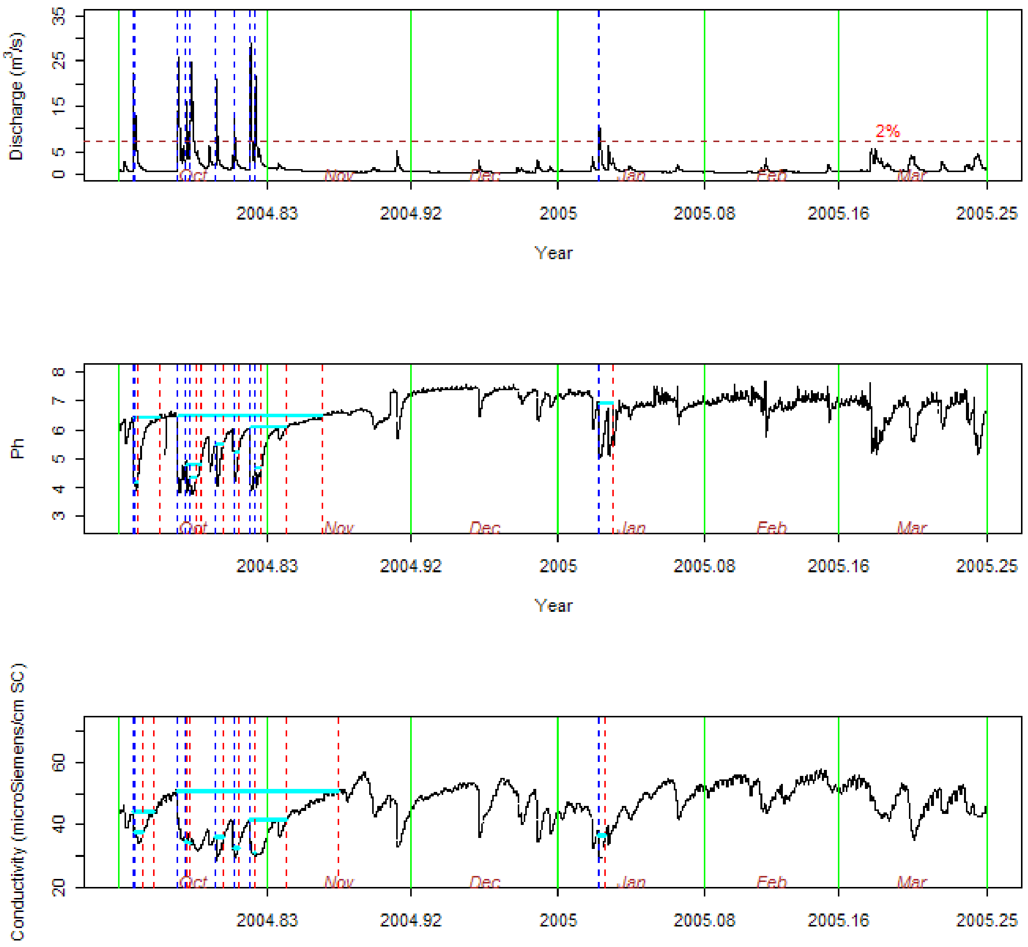

**Figure 4.18.** Time series plots of 15 minutes discharge (top) and conductivity (bottom) measurements at the River Charr in winter from Oct 2004-March 2005, with the blue and red vertical dashed lines denote the initial and end time points of each of the recovery period.

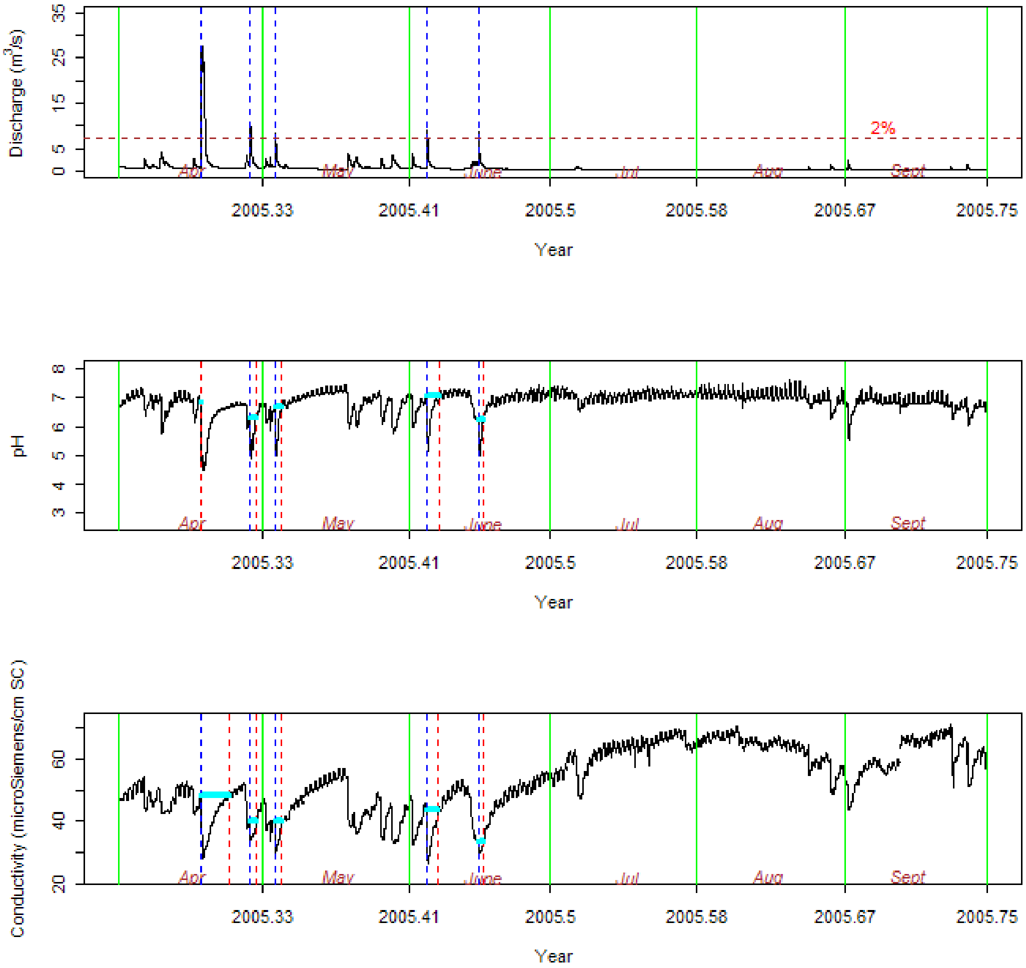

**Figure 4.19.** Time series plots of 15 minutes discharge (top) and conductivity (bottom) measurements at the River Charr in summer from Apr - Sept 2005, with the blue and red vertical dashed lines denote the initial and end time points of each of the recovery period.

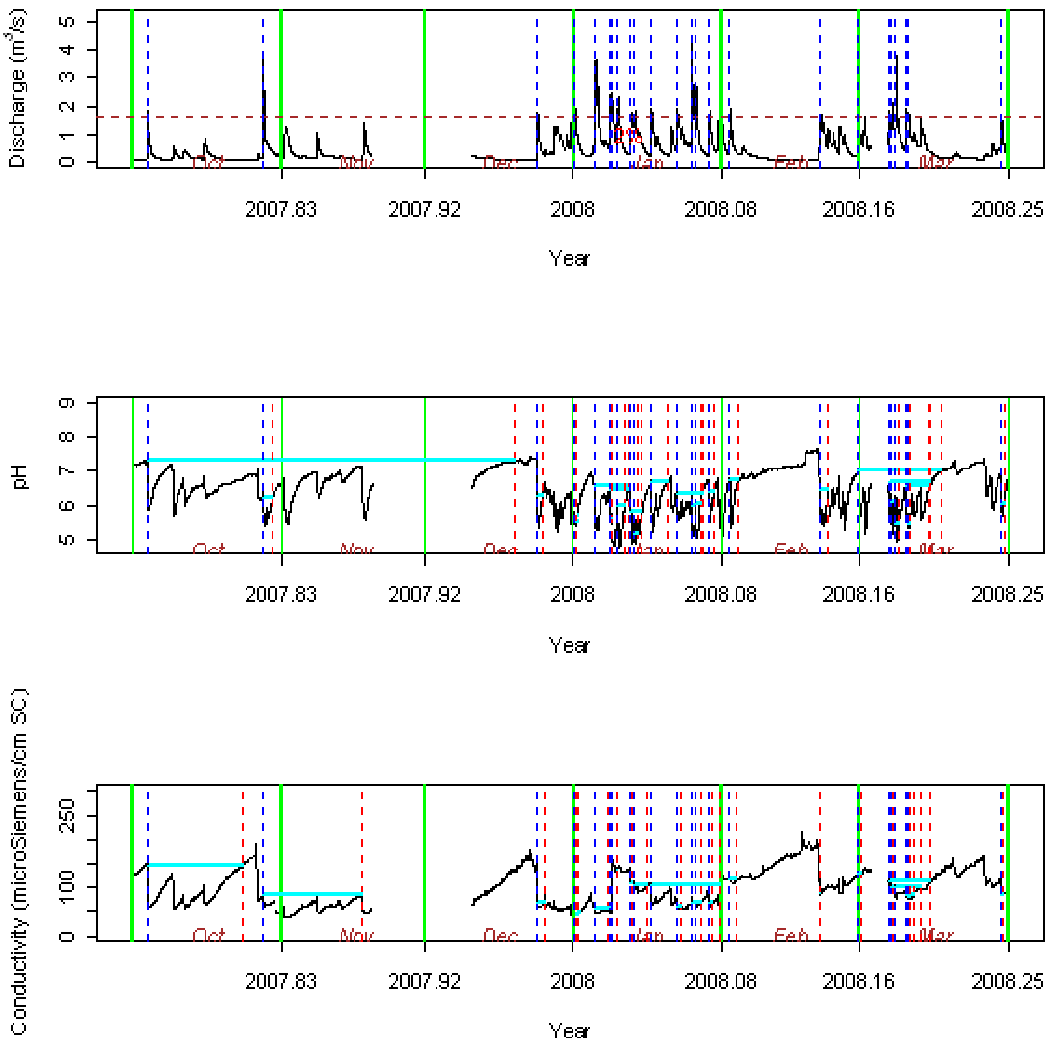

**Figure 4.20.** Time series plots for 30 minutes discharge (top) and conductivity (bottom) measurements at the River Drumtee over winter from Oct 2007-March 2008, with the blue and red vertical dashed lines denote the initial and end time points of each of the recovery period.

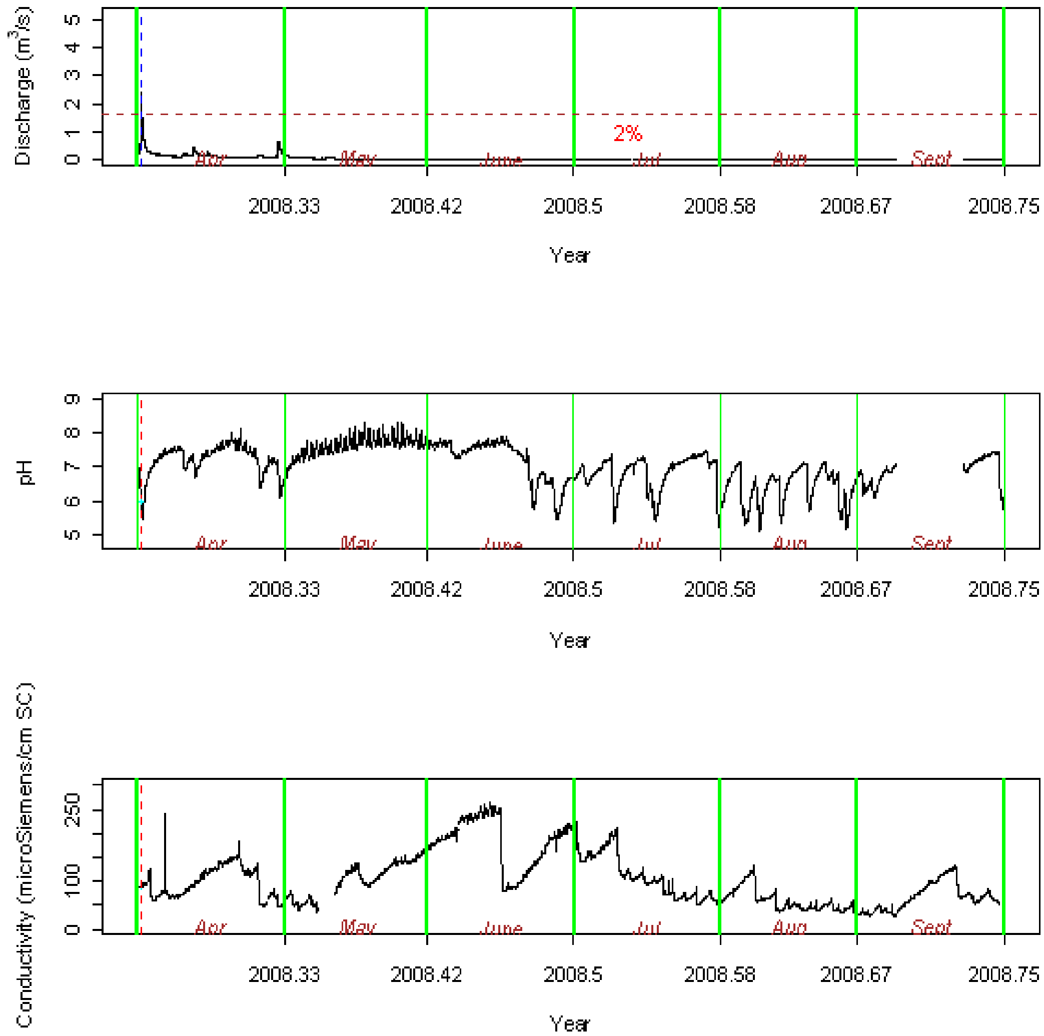

**Figure 4.21.** Time series plots for 30 minutes discharge (top) and conductivity (bottom) measurements at the River Drumtee over summer from Apr - Sept 2008, with the blue and red vertical dashed lines denote the initial and end time points of each of the recovery period.

### 4.4.4 Models of Log Recovery Period for pH and Conductivity

The final regression models for log recovery for pH at the River Charr corresponding to 2%, 3% and 10% threshold levels of extreme discharge are defined from equations 4.19 - 4.21, respectively. These models were determined using F-tests for model comparison as described earlier.

Table 4.2 summarizes the significant predictors for each model for pH and conductivity corresponding to threshold levels of 2%, 3% and 10% for the River Charr. The details of the model are as follow.

| Predictor | | pH | | | Conductivity | |
|---|---|---|---|---|---|---|
| | 2% | 3% | 10% | 2% | 3% | 10% |
| baseline | | / | / | | | / |
| max dis | / | / | / | | | / |
| max dis$^2$ | | / | / | | | / |
| min(pH/cond) | / | / | / | | / | / |
| log(slope) | / | / | / | | | |
| log(area dis) | / | / | / | / | / | / |
| log(num extreme dis) | / | / | | | / | / |
| season | / | / | / | | | |
| baseline*season | | | | | | |
| max dis*season | | | / | | | |
| min*season | / | / | / | | | |
| log(slope)*season | | | | | | |
| log(area dis)*season | / | / | / | | | |
| log(extreme dis)*season | | | | | | |

**Table 4.2.** Significant predictors at $\alpha$=0.05 for model of log recovery for conductivity corresponding to 2%, 3% and 10% threshold levels at the River Charr.

The fitted model 4.19 corresponding to 2% of extreme discharge shows linearity in all of the predictor terms, however, the quadratic term of maximum discharge is observed as the threshold decreases. Each of the fitted models shows evidence of maximum discharge, minimum pH, log of rate of change for pH, log of area above extreme discharge, the seasonal factor and an interaction between season and minimum pH as predictors in explaining the variability in log recovery for pH at all levels of threshold, indicating that these predictors are not affected by the levels of threshold. The seasonal factor indicates that the log recovery in winter corresponding to thresold levels of 2%, 3% and 10% is larger than summer by about 5.24, 3.67 and 6.06 units, respectively.

$$\begin{aligned} log(\widehat{recovery}) &= -2.43 - 0.03\{maxdis\}_i - 0.55\{minpH\} \\ &\quad -0.06\{log(slopepH)\} + 0.76\{log(areadis)\} \\ &\quad +0.17\{log(numextremedis)\} - 5.24\{season\} \\ &\quad +0.71\{minpH\} * \{season\} \\ &\quad +0.26\{log(areadis)\} * \{season\} \end{aligned} \tag{4.19}$$

$$\begin{aligned} log(\widehat{recovery}) &= -2.18 + 0.31\{baseline\} - 0.07\{maxdis\} \\ &\quad +0.001\{maxdis\}^2 - 0.67\{minpH\} - 0.07\{log(slopepH)\} \\ &\quad +0.59\{log(areadis)\} + 0.20\{log(numextremedis)\} \\ &\quad -3.67\{season\} + 0.38\{minpH\} * \{season\} \\ &\quad +0.25\{log(areadis)\} * \{season\} \end{aligned} \tag{4.20}$$

$$\begin{aligned} log(\widehat{recovery}) &= -0.27 + 0.66\{baseline\} - 0.20\{maxdis\} \\ &\quad +0.004\{maxdis\}^2 - 1.24\{minpH\} - 0.18\{log(slopepH)\} \\ &\quad +0.64\{log(areadis)\} - 6.06\{season\} \\ &\quad +0.19\{maxdis\} * \{season\} \\ &\quad +0.85\{minpH\} * \{season\} \qquad (4.21) \end{aligned}$$

Additionally, equations 4.22 - 4.24 display the final models of log recovery for conductivity at the River Charr corresponding to thresholds of 2%, 3% and 10% for extreme discharge. The strong effect of log of area above extreme discharge is highlighted at each threshold levels. The recovery is not influenced by seasonal factors, indicating that the length of the recovery periods following the extreme discharge in both winter and summer is similar.

Models of log recovery for pH and conductivity become more complex as the threshold levels increase. Different seasons may result in different length of days of recovery in pH, however, similar recovery period for conductivity is given over the year. This distinction suggests that pH takes slightly longer than conductivity to recover in the presence of extreme discharge over the winter season.

Models for pH are slightly more complex than conductivity for the same levels of threshold.

$$log(\widehat{recovery}) = -4.83 + 0.83\{log(areadis)\} \qquad (4.22)$$

$$\begin{aligned} log(\widehat{recovery}) &= -3.68 - 0.02\{mincond\} + 0.68\{log(areadis)\} \\ &\quad +0.20\{log(numextremedis)\} \end{aligned} \tag{4.23}$$

$$\begin{aligned} log(\widehat{recovery}) &= -2.98 + 0.02\{baseline\} - 0.10\{maxdis\} \\ &\quad +0.002\{maxdis\}^2 - 0.04\{mincond\} \\ &\quad +0.17\{log(numextremedis)\} \\ &\quad +0.72\{log(areadis)\} \end{aligned} \tag{4.24}$$

Table 4.3 summarizes the significant predictors for model of log recovery for pH and conductivity corresponding to threshold levels of 2%, 3% and 10% for the River Drumtee. The fitted models for pH and conductivity are shown in equations 4.25 - 4.30.

For the River Drumtee, the significant models of log recovery for pH corresponding to threshold levels of 2%, 3% and 10% for extreme discharge are defined from equations 4.25 - 4.27, respectively. All the fitted models show significant minimum pH, log of area above extreme discharge and log of number of extreme discharge. No evidence of a seasonal factor is highlighted in each of the models, indicating that the recoveries for both determinand are comparable between winter and summer.

| Predictor | pH | | | Conductivity | | |
|---|---|---|---|---|---|---|
| | 2% | 3% | 10% | 2% | 3% | 10% |
| baseline | | / | / | / | / | / |
| max dis | / | / | | | | / |
| max dis$^2$ | | | | | | |
| min(pH/cond) | / | / | / | / | / | |
| log(slope) | / | / | | | / | |
| log(area dis) | / | / | / | | | |
| log(num extreme dis) | / | / | / | / | / | / |
| season | | | | | | / |
| baseline*season | | | | | | |
| max dis*season | | | | | | / |
| min*season | | | | | | |
| log(slope)*season | | | | | | |
| log(area dis)*season | | | | | | |
| log(extreme dis)*season | | | | | | |

**Table 4.3.** Significant predictors at $\alpha$=0.05 for model of log recovery for conductivity corresponding to 2%, 3% and 10% threshold levels at the River Drumtee.

$$\begin{aligned} log(\widehat{recovery}) &= -1.53 - 0.33\{minpH\} - 0.13\{log(slopepH)\} \\ &\quad -0.23\{maxdis\} + 0.66\{log(areadis)\} \\ &\quad +0.26\{log(numextremedis)\} \end{aligned} \tag{4.25}$$

$$\begin{aligned} log(\widehat{recovery}) &= -2.36 + 0.58\{baseline\} - 0.60\{minpH\} \\ &\quad -0.21\{maxdis\} - 0.11\{log(slopepH)\} \\ &\quad +0.50\{log(areadis)\} \\ &\quad +0.24\{log(numextremedis)\} \end{aligned} \tag{4.26}$$

$$\begin{aligned} log(\widehat{recovery}) &= -9.64 + 0.55\{baseline\} \\ &\quad -0.57\{minpH\} + 0.38\{log(areadis)\} \\ &\quad +0.51\{log(numextremedis)\} \end{aligned} \tag{4.27}$$

Equations 4.28 - 4.30 represent the best fitted models for log recovery for conductivity at Drumtee, corresponding to 2%, 3% and 10% of extreme discharge. Each of the fitted models is significantly influenced by baseline for conductivity, minimum conductivity, log of rate of change for conductivity and log of extreme discharge corresponding to each of the thresholds. Season appears to be not significant in models 4.28 - 4.30.

$$\begin{aligned} log(\widehat{recovery}) &= -5.00 - 0.004\{baseline\} \\ &\quad +0.004\{mincond\} + 1.09\{log(numextremedis)\} \end{aligned} \tag{4.28}$$

$$\begin{aligned} log(\widehat{recovery}) &= -4.19 - 0.003\{baseline\} + 0.003\{mincond\} \\ &\quad +0.01\{log(slopecond)\} + 1.08\{log(numextremedis)\} \end{aligned} \tag{4.29}$$

$$\begin{aligned} log(\widehat{recovery}) &= -7.65 - 0.01\{baseline\} + 0.59\{maxdis\} \\ &\quad +1.04\{log(numextremedis)\} + 2.70\{season\} \\ &\quad -0.38\{log(maxdis)\} * \{season\} \qquad (4.30) \end{aligned}$$

The above models for both rivers highlight that the number of significant predictors that contribute to the variability in pH are relatively larger than that in conductivity for each threshold. Additionally, this suggests that pH is obviously influenced by certain features of discharge such as the area under the curve of extreme discharge and the number of extreme discharge. In particular, the rise of the number of extreme discharge and area under the extreme discharge results in the increase of recovery period for pH and conductivity. The adjusted coefficients of determination for each of the above fitted models are above 90%, suggesting that the recovery periods for pH and conductivity are reasonably well explained at each of three threshold levels for the two rivers. The diagnostic checking is carried out for the above fitted models of log recovery for pH and conductivity for both rivers.

### Diagnostic Plots

For illustration, Figure 4.22 presents the diagnostic plots for the models of log recovery for pH, subject to a threshold of 10% for the extreme discharge at the rivers Charr (left). The plot of residuals against the fitted values (top left) show the residuals are approximately scattered around zero mean. However, there are particular patterns in the residuals, highlighting heteroscedasticity of the residuals. Such patterns of the residuals around zero mean suggest a non-linear fitted model could be more appropriate for the log recovery for pH, however, no prior information on the previous models of log recovery for pH restrict the fitting of any nonlinear models. The normality plots (top right) show a small number of points at both tails of each of the normal distribution. However, most of the points are scattered around the dashed line and so, the normality assumption is reasonably satisfied. The acf (bottom left) and pacf (bottom right) shows no evidence of correlation structure for residuals. The current models for the log recovery for pH could be considerably accepted since most of the assumptions of the linear model are satisfied despite a small violation of the constancy of the residuals.

Diagnostic checks for the log recovery for pH corresponding to the thresholds of 2% and 3% of the extreme discharge provide similar results as above. Similarly, the model assumptions are satisfied for the model of log recovery for conductivity corresponding to each of the thresholds for both rivers.

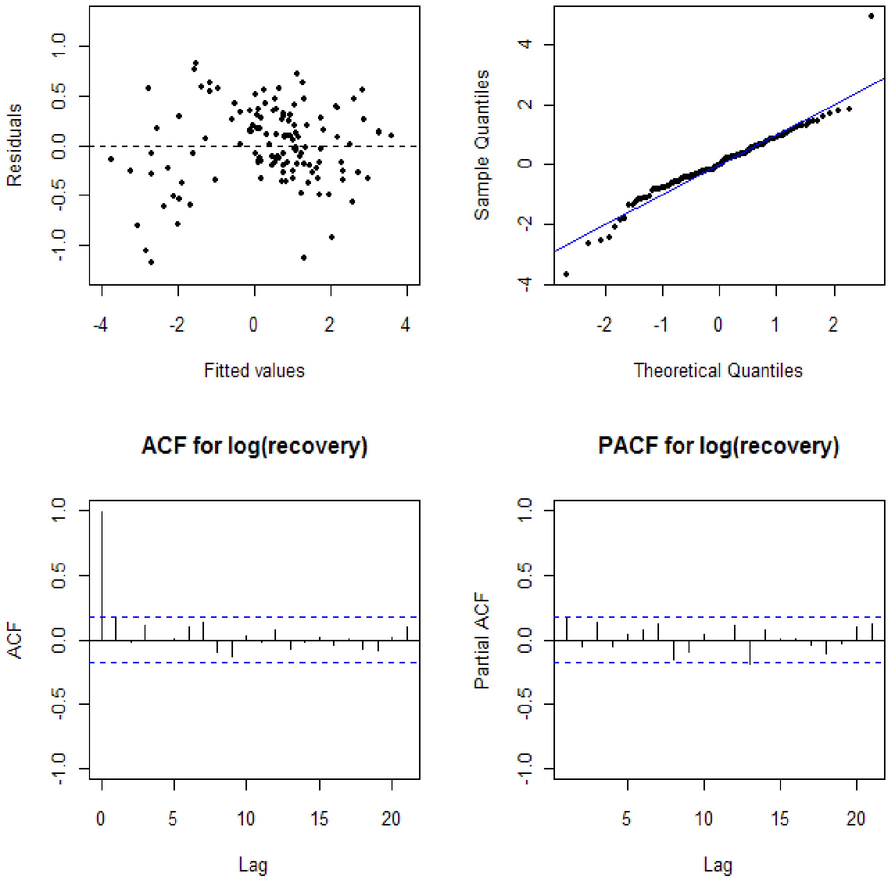

**Figure 4.22.** Diagnostic plots for model of log recovery for pH corresponding to a threshold of 10% for the extreme discharge at the River Charr.

### 4.4.5 Model Validation for River Charr

The previous fitted models for log recovery for pH and conductivity at the River Charr corresponding to each of the chosen thresholds are used to predict recovery periods in the period January - October 2008.

The mean square error (MSE) of the log recovery for pH and conductivity from the fitted models corresponding to the three levels of thresholds for extreme discharge are shown in Table 4.4. The MSE for models of log recovery for conductivity corresponding to threshold levels of 2% and 3% are fairly small. Similarly, low MSEs are observed for log recovery for pH at a threshold of 2% and this small value indicates that the current fitted models allow reasonable prediction. The MSE increases as the number of extreme discharge decreases, suggesting that better models are produced at higher thresholds.

| Threshold(%) | pH | Conductivity |
|---|---|---|
| 2 | 0.22 | 0.25 |
| 3 | 15.78 | 0.58 |
| 10 | 23.14 | 15.94 |

**Table 4.4.** Mean square error of models of log recovery for pH and conductivity

For illustration, Figure 4.23 shows the plots of fitted values against actual values of log recovery for pH (left) and conductivity (right) corresponding to threshold of 2% for extreme discharge. The line of equality is denoted by a blue line for each plot to highlight the similarity between the actual and fitted values of log recovery period. The similarity between the predicted and actual values of log recovery are more apparent in conductivity since the points are equally scattered above and below the line of equality compared to pH. Despite lower MSE for pH of 0.22 (Table 4.4), the actual log recovery for pH are greater than the corresponding predicted values, suggesting that the features of log recovery in 2008 are slighty different than in the previous 3 years and fitted values for log recovery for pH underestimates the actual values systematically.

### 4.4.6 Extremal Index for Discharge

Figure 4.24 illustrates the interval estimates of the extremal index (black line) over a series of thresholds for discharge at the River Charr (left) and the River

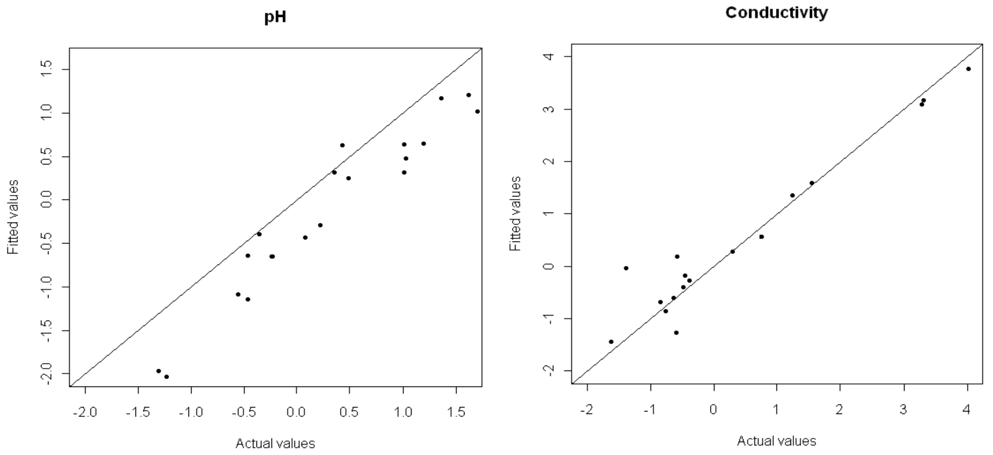

**Figure 4.23.** Plots of fitted against actual values of log recovery for pH (left) and conductivity (right), subject to threshold level of 2% for extreme discharge.

Drumtee (right). The bootstrap confidence limits (dashed colour lines) are highlighted around the extremal indexes and a horizontal line at $\theta$=1 (black dashed line) is marked on each plot to distinguish the feature of extreme discharge. Clusters of extreme discharge can be observed if the upper confidence limit for the extremal index lies below 1. Conversely, declustering of extreme discharge is highlighted if the upper confidence limits for the extremal index lies beyond 1.

Similar features of the extremal index for extreme discharge are highlighted for both rivers. However, a different characteristic of the upper confidence limit for the extremal index is observed at sufficiently large thresholds between Charr and Drumtee. The two plots show that the extremal index rises as the threshold goes up, indicating that the degree of clustering becomes smaller as the levels of threshold increase. The width of the confidence limits increases as the thresholds increases, indicating that higher variability could be observed in the extreme discharge as the chosen threshold increases. In particular, the independence of extreme discharge for River Charr and River Drumtee are observed at thresholds of $29m^3s^{-1}$ and $4.7m^3s^{-1}$, respectively.

The above results show the dependency of the extreme discharge at both rivers in a series of low up to sufficiently high thresholds. This is highly likely due to the fact that the extreme discharge occurs in a series of time points over a particular period. Hydrologically, such a feature is expected since a high discharge is observed during storm events and snow melting in the colder months of the year and so, their appearance can be highlighted in the winter.

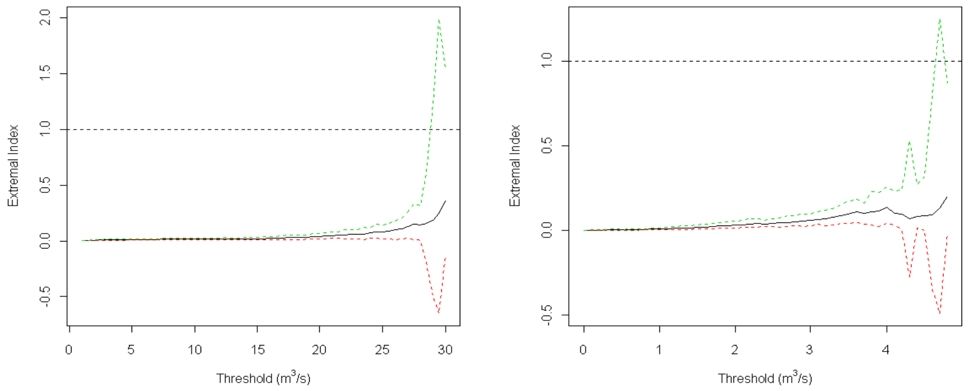

**Figure 4.24.** Plots of extremal indexes over a series of thresholds for discharge at Charr (left) and Drumtee (right).

## 4.5 Discussion

This study demonstrates important findings from analyzing a semi-continuous series of 15-minute and 30-minute measurements of environmental variables at Charr and Drumtee in Scotland, respectively. It also highlights what can be learned from approaches such as wavelets. The main advantage of this statistical approach is the ability to quantify the individual temporal patterns as well as the relationship of two processes at a number of shorter time scales at the River Charr, leading to better understanding of the dynamics of short-term variability in rivers. In addition, the recovery periods for the environmental variables at Charr and Drumtee following the extreme events in discharge have been demonstrated in the latter part of this chapter.

The estimated wavelet power spectrum provides precious information after first order differencing of the temperature, barometric pressure, pH and conductivity measurements at particular time points and scales. The integration of this wavelet spectrum and areawise significance testing is a powerful tool to assist the viewing and identification of dominant environmental processes. The use of first order differencing to remove the linear trend and further stabilise the variance of each time series provides apparent temporal pattern for temperature over summer. The wavelet power spectrum has shown some evidence of constant periodicities in temperature, pH and conductivity in warmer months whilst there is no apparent variation in barometric pressure over the year. Water temperature shows periodic features, particularly within a 24 hour period in summer and it is essentially attributed to physical factors in rivers. A large amount of solar radiation in summer results in increases in water temperature in the broad daylight, however, a significant heat transfer, released from the water into the air at night may cause a big drop in temperature in rivers and this natural diurnal cycle continues over the warmer months. Similar daily patterns in the pH and conductivity in warmer months could be a result of changes in carbonic acid in river in response to diurnal temperature variation. Barometric pressure is an independent meteorological determinant in the sense that there are no specific biological and chemical determinants in the water system that may reflect its temporal variation, however, it is controlled by the movement of air masses. The constant temporal change at 1 and 2 day scales are more likely due to the temporal variation in air masses that reflect high and low air pressures in several days of the colder months.

The wavelet coherence of pH and conductivity exposes regions with high common power and further provides valuable information on the phase relationship between the two signals. Wavelet coherence has proven useful to reveal the evidence of common oscillations between the two time series at particular scales and time points, suggesting a strong relationship between them from 1 to 3 days scale over each year. In particular, a significant relationship is shown at 1 day and 2-3 days bands in summer and both seasons, respectively. Seasonal changes in pH and conductivity could be a result of certain features of discharge. High discharge in winter typically occurs in response to high precipitation and snow melting and the excess partial pressure of carbon dioxide associated with this discharge level may result in low pH. The influence of solar radiation in a shorter day time during colder season is smaller and therefore, daily periodicity is not likely to be observed. In addition, the rise of a carbonic acid from carbon dioxide in stream water is likely to contribute to the drop of conductivity. While reducing carbon dioxide, the acidity of the stream may decrease, resulting in an increase of pH. However, the release of carbon dioxide in the water via respiratory processes during the night time may result in decline of pH.

The cross phase angle between pH and conductivity highlights the time shift between the two signals and so, the sequence of peaks of these signals has been identified and the corresponding ecological processes is as follows. The lead of peaks of pH over conductivity by about 6 hours in summer might be due to the prompt response to changes in carbon concentration in rivers following a diurnal cycle in water temperature. Similar time location of peaks from both signals at scales of 2-3 days over the year could be a result of small influence of temperature on the amount of carbon dioxide in the stream.

Further analysis on pH and conductivity contributes to the estimates of recovery period of each determinant following the extreme discharge at Charr and Drumtee. There are apparent differences in the recovery period for pH between the two rivers which might be explained by certain features of the areas. A seasonal factor appears to be significantly influential at the River Charr and is likely due to a large amount of snowfall. Furthermore, the River Charr is fairly close to the ocean and so, it is likely that frequency of precipitation is higher in Aberdeen than in Whitelee. The large amount of acid rain in the north east of Scotland over the winter season may result in seasonal dependence of recovery for pH compared to the east of Scotland which is much drier.

The number of predictors in the model of log recovery for pH at the River Charr is larger than the River Drumtee for the same level of threshold. This characteristic is expected since Aberdeenshire, which is mountainously area and often blanketed by snow fall in winter results in more variability in recovery as the snow melting occurs. This natural phenomenon which is often identified at the end of winter may require more predictors to explain the variability of recovery for pH.

However, a similarity of the predictors in the models of log recovery for pH between the two areas is observed. There is evidence of log area above extreme discharge and minimum pH in the models of log recovery for pH at both rivers, suggesting the occurrence of primary contributors of recovery, which is not restricted by threshold levels as well as the features of the locations.

Dissimilarity of the predictors between the two rivers is shown from the fitted models of log recovery for conductivity corresponding to each threshold, suggesting that the recovery process of this environmental variable from the two different areas could be influenced by different contributors. Distinct complexity of the models of recovery period for conductivity at the River Charr are observed as the thresholds of extreme discharge increase. A simple model of log recovery for conductivity corresponding to a 2% threshold for extreme discharge is more appropriate at the River Charr than the River Drumtree Burn, suggesting a unique contributor of the recovery period. However, the models of recovery for conductivity at the River Charr become more complex as thresholds decrease, suggesting that the independence feature of the extreme discharge plays an important role in producing a simple model of log recovery. Such a feature at the River Charr is highlighted as the fitted models are validated, indicating that better models of recovery for conductivity are acquired.

Despite a clear difference of the range of discharge measurements between Charr and Drumtee, they exhibit a similar characteristic of extreme discharge. In particular, the declustering tends to occur as the thresholds increase. Such a feature is noticeable from the plots of extremal index on a series of thresholds. The clustering of extreme discharge for each river is apparent and occurs over a wide range of thresholds. The evidence of declustering of extreme discharge is highlighted for each river as the thresholds reach up to sufficiently large values. However, the declustering as a result of high thresholds does not affect the previous fitted model of recovery period for pH and conductivity despite smaller thresholds of 2%, 3% and 10% chosen for the extreme discharge. This is due to the fact that the diagnostic check is satisfied, highlighting the adequacy of the models of log recovery for pH and conductivity.

# Chapter 5

# Conclusions and Future Work

## 5.1 Conclusions

The key concept of this thesis is to present a variety of different statistical approaches that are appropriate to model time series data recorded at low, moderate and high temporal resolutions. This is important to answer questions of interest regarding the environmental and ecological issues of freshwaters as frequently raised by limnologists, biogeochemists and regulators. In particular, the statistical analysis has been applied to the environmental data with different temporal resolutions from Loch Lomond, the rivers Charr and Drumtee Burn in Scotland.

The first part of this work is based on analysis of monthly temperature and chlorophyll measurements from 1987-2005 in Loch Lomond. It was of interest to investigate trend and seasonal pattern in temperature and chlorophyll. The non-constancy of the variability of the chlorophyll is addressed prior to model fitting, by applying a natural log transform to stabilize the variance. A large number of missing values for temperature and chlorophyll do not occur at random since they mostly appear in winter and these have been imputed using a variety of statistical approaches. Moreover, the extended period of missing data for chlorophyll was not included in the analysis.

A variety of models were explored for the low temporal data for surface temperature and chlorophyll for the north and south basins of Loch Lomond. Local linear regression is used for the models which involve estimation of a smooth function for the trend and the flexibility of the model is controlled by the bandwidth, determined by the degrees of freedom. Moreover, a spline smoothing method with the smoothness of the function controlled by the number of knots, is used in the additive model since a cyclical smoother is required for the seasonality term.

The evidence of smooth trend and seasonal pattern for monthly temperature in the north and south basins of Loch Lomond highlighted by the additive model indicates that the pattern changes smoothly over time. In general, the temperature increases from 1987-2005. For log chlorophyll, the significant linear trend and constant seasonal pattern in both basins, resulting from the appropriateness of the harmonic model provide an indicator for the mean changes in the deeper and shallower basins. The same trends for log chlorophyll in the north and south basins may suggest similar influences of physical, chemical and biological determinants in the water body. Further investigation of the contribution of phosphate, nitrate and temperature to the variation of log chlorophyll over the year provides further insight on the ecological process in lake.

Phosphate and nitrate are not significant predictors in the model of log chlorophyll for the north and south basins, suggesting that higher resolution data might be required to explain their significant relationship. There is evidence of a temperature effect in the model of log chlorophyll for the south basin, indicating that the blooms of phytoplankton in shallower water of the loch is highly likely influenced by the temperature. However, no significant influence of temperature and nutrients on log chlorophyll in the north basin and it is more likely that the low frequency data did not pick up the significant relationship.

The second part of the work is the extension of the previous work done on temperature to higher frequency of 1 and 3 hourly data from thermistor chains. Data were recorded at 11 different depths in the north, mid and south basins. The changes of temperature over the year by considering depth as a random effect, at different locations of the loch is determined from the mixed-effects model. In addition, the estimates of the the position of the thermocline due to its importance of partitioning the water column into two strata with different biological and chemical features that reflect the ecological process in lakes, is determined from the deeper water (north and mid basins).

The determination of the correlation structure of the deseasonalised residuals using the autocorrelation function indicates an AR(1) model for the north and AR(2) model for the mid and south basins may provide a plausible description of the autocorrelation structure in the residuals of the temperature. Ecologically, the use of an AR model for defining the correlation structure of the residuals of temperature for each depth in the loch tailors to the natural relationship in the real environmental time series. Both fixed and random effects models were investigated but random effects are most appropriate for all basins.

There is evidence of a decrease of temperature in the north and south basins but for the mid basin, the significant increase of temperature is highlighted over the year. Such a difference is likely due to the different ecological process in different years despite the temperature measurements being collected in the same loch. The changes of temperature at different locations of the loch are fairly similar as marked by the small values of the fixed effects $\beta_1$. Nevertheless, the incorporation of a random effect $b_1$ in the mixed-effects model that represent the mean changes of temperature at different depths may lead to different changes of temperature at different depths.

Further investigation on the moderate temperature measurements with depths at the deeper locations of the loch provides estimates of the position of the thermocline. Different approaches corresponding to different mathematical and limnological terminologies of the thermocline, are used as a basis for estimating such a natural feature in lakes in warmer months. The maximum relative rate of change and changepoint regression approaches, which rely on the mathematical definition, produce potential estimates of the thermocline depths. In addition, the derivative of a smooth curve approach following the limnological definition produce similar results to the previous approaches.

For the maximum relative rate of change approach, the lack of statistical properties, unnatural estimates and the restriction to the cut-off point are drawbacks to the method. Therefore, a changepoint regression, which is able to capture the position of the thermocline at a depth corresponding to a rapid change of temperature with depth at a given time point, was investigated, following the limnological definition of identifying the point of abrupt change.

While the changepoint approach indicates the occurrence of maximum changes of temperature with depth at a particular time point, the approach used here cannot deal with the situation of an inflection point. Therefore, derivatives of smooth curves at each time point were investigated to follow the mathematical definition of detecting inflection points.

The different variability of the estimates of the thermocline depth over the time period produced by the changepoint regression and derivative of a smooth curve in the north basin agree to the limnological and mathematical terminologies of the thermocline but for the mid basin, similar variability is displayed from both statistical approaches. It is likely due to the fact that different ecological processes occur in the water column over different years and so, the position of the thermocline may vary at a given time point in warmer months.

The final part of this work deals with environmental data of high temporal resolution. The temporal patterns for temperature, pH, conductivity and barometric pressure, and the relationship between pH and conductivity at short temporal scales are investigated at the river Charr, and the models of log recovery periods are fitted to pH and conductivity following extreme discharge at the rivers Charr and Drumtree Burn.

Using wavelets, there is strong evidence of a temporal pattern for temperature exhibited in the warmer months but weak evidence of temporal patterns for conductivity and pH in the warmer months are presented by the wavelet spectrum. The evidence of temporal patterns at a scale of 1-2 days are exhibited by the barometric pressure over the year. The areawise significant test does not remove the dubious significant patches for each pH and conductivity. The determination of the relationship between pH and conductivity via wavelet coherence show evidence of the similarity of the signals of the two time series at a 1 day scale in summer and a 2 day scale in summer and winter.

The wavelet power spectrum provides strong evidence of a diurnal pattern for temperature over summer throughout the years as expected whereas, pH and conductivity show a moderate variability over summer, despite the dubious patches. In barometric pressure, the wavelet power spectrum shows temporal patterns at scales of 1-2 days. Further investigation on the coherence between pH and conductivity provide evidence of the similarity of the two signals at a 1 day scale in summer but a more consistent relationship at a 2 day scale in both seasons of the year.

The recovery period for each of pH and conductivity, which begins from the time point where the measurements start to decrease in response to extreme discharge, until the values return to the pre-event levels, is determined. Linear regression models of log recovery period for pH and conductivity are fitted on the baseline of pH and conductivity, maximum values of discharge within the recovery period, minimum pH and conductivity within the recovery period, the area bounded the thresholds and curve of extreme discharge, seasonal factor and the interaction of each of the above predictor with the seasonal factor. The evidence of significant predictors is determined for model of log recovery for pH and conductivity corresponding to three thresholds of extreme discharge determined from its cumulative distribution function.

The fitted models of log recovery for pH and conductivity show the significant influence of a particular feature of the extreme discharge. In particular, the model of log recovery for pH and conductivity corresponding to each of the thresholds for extreme discharge show evidence of the log area of discharge to explain the variability of the log recovery for the river Charr. For the river Drumtee Burn, log number of extreme discharge are significant in the models of log recovery for pH and conductivity corresponding to each of the thresholds for extreme discharge. The area bounded by the the horizontal line of threshold and the curve of extreme discharge contributes significantly to the recovery period for pH and conductivity at the River Charr but for the River Drumtee Burn, the number of extreme discharge appears to be more influential to the recovery period for both pH and conductivity at any of the three levels of threshold for extreme discharge. Moreover, the investigation of cluster of extreme discharge at the rivers Charr and Drumtee Burn provides a sufficiently large thresholds that characterize the independence of the cluster of extreme discharge. In comparison, the threshold for extreme discharge for the river Charr is comparably larger than the river Drumtee Burn.

## 5.2 Future Work

The extension of current works on a low, moderate and high temporal resolutions of environmental data could possibly be carried out in future.

For environmental measurements in lakes with low temporal resolution, the investigation of potential biological determinants on the variability of log chlorophyll in the north and south basins could be carried out to determine the evidence of biological determinants on the bloom of phytoplankton. The effect of biological determinants to the increase of chlorophyll should be given much attention in accordance to the drop of water quality. The number of local aquatic animals for each basin of the lake may also be incorporated in the model of chlorophyll since it may affect the changes of the bloom of phytoplankton.

Additional moderate frequency of temperature measurements in later years may lead to the investigation of the trend and seasonality term for each depth. The linear mixed model with a trend and seasonal term over several years may not only enable the estimation of temperature within the time period of measurements at a particular depth but also allow prediction of the temperature in the later years.

Meanwhile, in the issue of the thermocline, physical factors should be incorporated in determining the position of this natural phenomenon in lakes during the summer since this might increase the precision of the estimates of its position as well as providing more informative results with regard to other significant natural effects. The time series of temperature, physical factors and depth could be modelled together over the year to allow prediction of the position of the thermocline in the later years. Consequently, the estimates could be improved and more understanding on the prominent natural features in lakes could be gained.

For the environmental determinant in rivers with a high temporal resolution, different ways of treating the actual measurements of pH and conductivity could be tested to allow a clear presentation of a significant power spectrum without the disturbance of dubious patches. For an example, the extreme events in pH and conductivity could be removed prior to wavelet analysis. The wavelet coherence shows evidence of a common signal of pH and conductivity at particular short scales. However, no investigation is carried out on the relationship between each of pH and conductivity, and discharge. Such an investigation is also important since it may provide sufficient information on the relationship between environmental and hydrological variables in rivers. The temporal pattern at short time scales can be used as a basis in modelling the corresponding variables. For instance, a clear diurnal cycle for temperature over summer may lead to fitting a model of temperature which incorporate the diurnal term for summer but for winter, this daily cyclical term is not defined.

Despite a simple regression model used to fit the recovery period for pH and conductivity, this model has shown significant relationships between the recovery period and several potential predictors following the extreme discharge corresponding to particular thresholds. This can be extended by incorporating other potential predictors in the model.

The semi-continuous to continuous data monitoring will become more common practice in the future and increasingly sophisticated and complex models will be required to extract features of interest from very noisy, nonstationary data

# Bibliography

Abell, R. (2002). Conservation biology for the biodiversity crisis: a freshwater follow up. *Conserve. Bio.*, 16:1435–1437.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Proc. 2nd Int. Symp. Information Theory*, pages 267–281.

Alain, F. Z., Elena, N. I., Neil, J. W., Anatoly, A. S., and Graham, M. S. (2009). *Mixed Effects Models and Extensions in Ecology in R.* Springer.

Allison, P. D. (2002). *Missing Data.* Sage.

Anctil, F. and Coulibaly, P. (2004). Wavelet analysis of the inter-annual variability in southern Quebec stream flow. *Journal of Climate*, 17:163–173.

Austin, J. A. and Colman, S. M. (2007). Lake superior summer water temperatures are increasing more rapidly than regional air temperatures: A positive ice-albedo feedback. *Geophyical Research Letters*, 34, L06604. doi:10.1029/2006GL029021.

Bowerman, B. L. and O'Connell, R. T. (1993). *Forecasting and time series.* Duxbury.

Bowman, A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with Splus Illustrations.* Oxford statistical sciences series: 18. Oxford: Clarendon Press; New York: Oxford University Press.

Bowman, A. W. and Azzalini, A. (2003). Computational aspects of nonparametric smoothing with illustrations from the sm library. *Computational Statistics and Data Analysis*, 42:545–560.

Cai, Z., Fan, J., and Li, R. (2000a). Effect estimation and inferences for varying-coefficients models. *Journal of the American Statistical Association*, 95(451):888–902.

Cai, Z., Fan, J., and Yao, Q. (2000b). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association*.

Cari, G. K., Valerie, V., and Robert, E. K. (2005). Spline-based non-parametric regression for periodic functions and its application to directional tuning of neutrons. *Statistics in Medicine*, 24:2255–2265.

Carvalho, L., Solimini, A., Phillips, G., van den berg, M., Pietilainen, O.-P., Solhein, A. L., Poikabe, S., and Mischke, U. (2008). Chlorophyll reference conditions for European lake types used for intercalibration of ecological status. *Aquat Ecol*, 42:203–211.

Chatfield, C. (1996). *The Analysis of Time Series*. Chapman and Hall.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83:596–610.

Clifford, M. H., Jeffrey, S., and Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *J. R. Stats. Soc. B*, 60:271–293.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31:377–403.

CRE Energy (2002). *Whitelee windfarm environmental statement.* Chapter 4 and 5.

Daubechies, I. (1990). The wavelet transform time-frequency localization and signal analysis. *IEEE Transactions or Information Theory*, 36:961–1004.

Davison, A. C. and Smith, R. L. (1990). Models for exceedences over high thresholds. *J. R. Stats. B*, 52:393–442.

de Boor, C. (1978). *A practical guide to splines.* Springer, New York.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society*, 39:1–22.

Diane, L. L., Amy, L., and Stanley, L. (2010). Techniques for handling missing data in secondary analyses of large surveys. *Academic Pediatrics*, 10:205–210.

Eubank, R. (1988). *Spline Smoothing and Nonparametric Regression.* Marcel Dekker, New York.

Fan, J. (1993). Local linear regression smoothers and their minimax. *Journal of the American Statistical Association*, 21:196–216.

Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Stats.*, 20:2008–2036.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications*. London: Chapman and Hall.

Fan, J. and Zhang, J. (2000). Functional linear models for longitudinal data. *J. R. Stats Soc, Ser. B*, 62:303–322.

Farge, M. (1992). Wavelet transformations and their application to turbulence. *Annual Reviews of Fluid Mechanics*, 24:395–457.

Ferguson, C. (2007). *Univariate and Multivariate Statistical Methodologies for Lake Ecosystem Modeling*. PhD thesis, Department of Statistics, University of Glasgow.

Ferro, C. A. T. (2003). Inference for clusters of extreme values. *J. R. Stats. Soc. B*, 65:545–556.

Ferro, C. A. T. and Segers, J. (2002). *Automatic declustering of extreme values via an estimator for the extremal index*. Technical Report: EURANDOM, Eindhoven.

Florentina, M., Francoise, F., and Alain, P. (1999). Ph modelling by neural networks. application of control and validation data series in the middle Loire river. *Ecological Modelling*, 120:141–156.

Gillet, C. and Quetin, P. (2006). Effect of temperature changes on the reproductive cycle of roach in Lake Geneva from 1983 to 2001. *Journal of fish biology*, 69:518–534.

Green, P. J. (1987). Penalized likelihood for general semiparametric regression models. *International Statistical Review*, 55:245–259.

Grinsted, A. J., Moore, J. C., and Jevrejeva, S. (2004). Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Process geophys*, 11:561–566.

Gurnell, A. M., Clar, M. J., and Hill, C. T. (1992). Analysis and interpretation of patterns within and between hydroclimatological time series in an alpine glasier basin. *Earth Surface Processes and Landforms*, 17:821–839.

Habib, O. A., Tippett, R., and Murphy, K. J. (1997). Seasonal changes in phytoplankton community structure in relation to physico-chemical factors in Loch Lomond, Scotland. *Hydrobiologia*, 350:63–79.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Sciences*, 1:297–318.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additve Models*, volume 43. Chapman and Hall, first edition.

Helmut, Z. B. and Thomas, P. (2008). The role of temperature, cellular quota and nutrient concentrations for photosynthesis, growth and light-dark acclimation in phytoplankton. *Limnologica*, 38:313–326.

Hodgkins, R. (2001). Seasonal evolution of meltwater generation, storage and discharge at a non-temperate glacier in Svalbard. *Hydrological Processes*, 15(3):441–460.

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4):809–822.

Horne, A. J. and Glodman, C. R. (1994). *Limnology*. McGraw-Hill.

Hsing, T. (1987). On the characterization of certain point processes. *Stochast. Process. Applic.*, 26:297–316.

Hu, W. P., Zhai, S. J., Zhu, Z. C., and Han, H. J. (2008). Impacts of the Yangtze River water transfer on the restoration of lake Taihu. *Ecological Engineering*, 34:30–49.

Hudson, D. J. (1966). Fitting segmented curves whose join points have to be estimated. *J. Am. Statist. Ass.*, 61:1097–1129.

Hurvich, C. M., Simonoff, J. S., and Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *J. Roy. Statist. Soc. B.*, 60:271–293.

Hutchinson, G. E. (1937). *Limnological studies in Connecticut. IV. The mechanisms of intermediary metabolism in stratified lakes.*

Hutchinson, G. E. (1957). *A threat on Limnology. I. Geography, physics and chemistry.*

Imberger, J. and Patterson, J. C. (1990). Physical limnology. *Adv. Appl. Mechanics.*, 27:303–475.

Jarvie, H. P., Neal, C., Smart, C., Owen, R., Fraser, D., Forbes, I., and Wade, A. (2001). Use of continuous water quality records for hydrograph separation and to assess short-term variability and extremes in acidity and dissolved carbon dioxide for the River Dee, Scotland. *The Sci. of Total Env.*, 265:85–98.

Jin, X. and Yao, J. (2006). Empirical study of ARFIMA model based on fractional differencing. *Physical*, 377:138–154.

John, P. S. and Michael, J. S. (2011). Comparative assessment of indices of freshwater habitat conditions using different invertebrate taxon sets. *Ecological Indicators*, 11:370–378.

Josep, P. and Anna, A. (1992). Streamwater pH, alkalinity, $pCO_2$ and discharge relationships in some forested Mediterranean catchments. *Journal of Hydrology*, 131:205–225.

Julious, S. A. (2001). Inference and estimation in a changepoint regression problem. *J. of Royal Stats. Soc. Series D*, 50(1):51–61.

Kendall, M. G., Stuart, A., and Ord, J. K. (1983). *The Advanced Theory of Statistics*. London: Griffin.

Krisnaiah, P. K. and Miao, B. Q. (1988). *Review about estimation of change-points. In Handbook of Statistics*. Amsterdam: North-Holland.

Krokowski, J. (2007). Changes in the trophite state and phytoplankton composition and abundance in Loch Lomond, Scotland, UK. *Int. Jou. Of Oceanography and Hydrobiology*, XXXVI:18–33.

Kuh, L., Yun, B. K., Jong, J. P., SungHyun, N., Kyung, A. P., and Kyung, I. C. (2005). Long-term and real monitoring system of the east Japan sea. *Ocean Science Journal*, 40:25–44.

Kumar, P. (1996). Role of coherent structure in the stochastic dynamic variability of precipitation. *J. Geophys. Res.*, 101:393–404.

Kumar, P. and Fouroula-Georgiou, E. (1997). Wavelet analysis for geophysical application. *Review of Geophysics*, 35:385–412.

Labat, D., Ababou, R., and Mangin, A. (2000). Rainfall-runoff relations for karstic springs: Part ii. continuous wavelet and discrete orthogonal multiresolution analyses. *J. Hydrol.*, 238:149—178.

Labat, D. and Ronchail, J. L. (2005). Recent advances in wavelet analysis:part 2. Amazon, Parana, Orinoco and Congo discharge time scale variability. *Journal of Hydrology*, 314:289–311.

Lai, H. L. and Helser, T. (2004). Linear mixed-effects models for weight-length relationships. *Fisheries Research*, (70):377–387.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38:963–974.

Leadbetter, M. R., Lindgren, G., and Rootzen, H. (1983). *Extremes and Related Properties of Random Sequences and Processes.* Springer, New York.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses.* Wiley, New York.

Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated measures data. *Journal of the American Statistical Association*, 83:1014–1022.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with missing data.* hoboken, NJ: Wiley.

Loader, C. (1999). *Local regression and likelihood.* Springer.

Loch Lomond and Trossachs National Park Authority (2005). State of the Park report 2005. Technical report, Loch Lomond and Trossachs National Park Authority, Balloch, Scotland.

Longford, N. T. (1993). *Random Coefficients Models.* Oxford University Press, New York.

Lucero, O. A. and Rodriguez, N. C. (1999). Relationships between interdecadal fluctuations in annual rainfall amount and annual rainfall trend in a southern mid-latitudes regions of Argentina. *Atmos. Res.*, 52:177–193.

Maitland, J. (1981). *The ecology of Scotland's largest lochs, Lomond, Awe, Ness, Morar and Shiel.* Dr. W. Junk.

Malmqvist, B. and Rundle, S. (2002). Threats to the sunning water ecosystems of the world. *Env. Conserve.*, 29:134–153.

Maraun, D., Kurths, J., and Holschneider, M. (2007). Nonstationary Gaussian processes in wavelet domain: Synthesis, estimation and significance testing. *Physical Review*, E75, 016707.

Meng, Q., Cieszewski, C. J., Madden, M., and Borders, B. (2007). A linear mixed-effects model of biomass and volume of trees using Landsat ETM-images. *Forest Ecology and Management*, (244):93–101.

Misiti, M., Misiti, Y., Oppenheim, G., and Poggi, J.-M. (1996). *Wavelet toolbox User's Guide (for use with Matlab).* The Math Works Inc.: Natwick, MA.

Moyeed, R. A. and Diggle, P. J. (1994). Rates of convergence in semi-parametric modeling of longitudinal data. *Aust. J. Statistics*, 36:75–93.

Murphy, K. J., Hudson, K. D., and Mitchell, J. (1994). Freshwater and wetland plant communities of Loch Lomond. *Hydrobiologia*, 290:63–74.

Nadaraya, E. A. (1964). On estimating regression. *Theory Probab. App.*, 10:186–190.

Neal, C., Harrow, M., and William, R. J. (1998). Dissolved carbon dioxide in the River Thames:spring - summer 1997. *The Sci. of Total Env.*, 210:205–217.

Phillis, E. (1997). *The complete weather resource*. U.X.L.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-Plus*. Springer.

Quandt, R. E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *J. A. Stats. Ass.*, 53:873–880.

Rajendra, G. K. and David, P. H. (2002). Flushing of dense, hypoxic water from a cavity of the Swan river estuary, Western Australia. *Estuaries*, 25:908–915.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, Ser. B*, 53:233–243.

Ricker, W. E. (1937). Physical and chemical characteristics of Cultus Lake, British Colombia. *J. Bio. Board Canada*, 3(4):363–402.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge.

Saco, P. and Kumar, P. (2000). Coherent modes in multiscale variability of streamflow over the United States. *Water Resour. Res.*, 36:1049–1067.

Santiago, B. (2005). Uncertainties in partial duration series modelling of extremes related to the choice of the threshold value. *Journal of Hydrology*, 303:215–230.

Schaefli, B., Maraun, D., and Holschneider, M. (2007). What drives high flow events in the Swiss Alps? recent developments in wavelet spectral analysis and their application to hydrology. *Advances in Water Resources*, 30:2511–2525.

Shesh, R. K., Randall, W. G., Patrick, J. M., Edmund, P., and John, S. S. (2010). Time and frequency domain analyses of high frequency hydrologic and chloride data in an east Tennessee watershed. *Journal of Hydrology*, doi:10.1016.

Shuijing, Z., Weiping, H., and Zecong, Z. (2010). Ecological impacts of water transfers on lake Taihu from the Yangtze River China. *Ecological Engineering*, 36:406–420.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *J. R. Statist. Soc. B*, 47:1–52.

Simonoff, J. S. (1996). *Smoothing methods in statistics.* New York: Springer.

Slack, H. D. (1957). Studies on Loch Lomond. Report, Glasgow, Glasgow.

Smith, B. D., Lyle, A. A., and Rosie, A. J. (1981). *Comparative physical limnology. The ecology of Scotland's largest lochs: Lomond, Awe, Ness and Shiel.* Dr. W. Junk Publishers, The Hague.

Smith, R. L. and Weissman, I. (1994). Estimating the extremal index. *J. R. Stats. Soc. B.*, 56:515–528.

Smith, V. H., Tilman, G. D., and Nekola, J. C. (1999). Eutrophication: impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems. *Environ. Pollution*, 100:179–196.

Spears, B. M., Carvalho, L., and Perkins, R. (2008). Effects of light on sediment nutrient flux and water column nutrient stoichiometry in a shallow lake. *Water Research*, 42:977–986.

Speckman, P. (1988). Kernel smoothing in partially linear models. *J. R. Statist. Soc. B*, 50:413–436.

Spyros, G. M., Steven, C. W., and Rob, J. H. (1997). *Forecasting:Methods and Application.* John Wiley.

Stevenson, D. S. and Van Schaik, J. C. (1967). Some relations between changing barometric pressure and water movement into lysimeters having controlled water tables. *Jour. of Hyd.*, 5:187–196.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Ann. Stats.*, 13:689–705.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. B.*, 36:111–147.

Stumm, W. and Morgan, J. (1981). *Aquatic Chemistry.* Wiley Interscience, New York.

Szilagyi, J., Parlange, M. B., Katul, G. G., and Albertson, J. D. (1996). An objective method for determining principal time scales of coherent eddy structures using orthogonal wavelets. *Adv. Water Resour.*, (6).

Thisted, R. A. (1988). *Elements of Statistical Computing.* Chapman and Hall, London.

Thompson, D. B., Gordon, J. E., and Horsfield, D. (2001). *Montane landscapes in Scotland: are these natural artefacts or complex relicts? : Earth Science and the Natural Heritage.* Stationary Office, London.

Tobias, V., Eduard, H., Philipp, S., Anja, F., Mario, S., and Olaf, A. C. (2010). Fluctuations of electrical conductivity as a natural tracer for bank filtration in a losing stream. *Advances in Water Resources*, 33:1296–1308.

Torrence, C. and Compo, G. P. (1998). A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79(1):67–78.

Torrence, C. and Webster, P. J. (1999). Interdecadal changes in the ENSO-Monsoon system. *Journal of Climate*, 12:2679–2690.

Turnipseed, D. P. and Sauer, V. B. (2010). *Discharge measurements at gaging stations*. U.S. Geological, Reston, Virginia.

Venables, W. N., Smith, D. M., and the R Development Core Team (2011). *An Introduction to R: A programming environment for data analysis and graphics*, volume 2.13.2.

Victor, C. and Robin, A. W. (2005). Using the generalized F distribution to model limnetic temperature profile and estimate thermocline depth. *Ecological Modelling*, 188:374–385.

Wahba, G. (1990). Spline models for observational data. In *CBMF-NSF Regional Conference Series in Applied Mathematics*. SIAM: Philadelphia.

Waldron, S., Flowers, H., Arlaud, C., Bryant, C., and McFarlane, S. (2009). The significance of organic carbon and nutrient export from peatland-dominated landscapes subject to disturbance, a stoichiometric perspective. *Biogesciences*, 6:363–374.

Waldron, S., Scott, E. M., and Soulsby, C. (2007). Stable isotope analysis reveals lower-order river dissolved inorganic carbon pools are highly dynamic. *Env. Sci. Tech.*, 41:6156–6162.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya, Ser. A*, 86:342–361.

Weissman, I. and Novaks, S. Y. (1998). On blocks and runs estimators of the extremal index. *J. Stats. Planning Inf.*, 66:281–288.

Wetzel, R. (2001). *Limnology: lake and river ecosystems.* San Diego; London; Academic Press.

Wood, S. (2005). mgcv: Gams and generalized ridge regression. *R. R. News*, 2:20–25.

Yucheng, W., Zhe, L., Huiwang, G., Lian, J., and Xinyu, G. (2011). Response of salinity distribution around the yellow river mouth to abrupt changes in river discharge. *Continental Shelf Research*, 31:685–694.

Zeger, S. L. and Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Journal of Biometrics*, 50:689–699.