University of Southampton

Faculty of Social Science
Mathematical Sciences

Bayesian Multivariate Normal Modelling

by

Sthaporn Thepsumritporn

Thesis for the degree of Doctor of Philosophy

December 2021

University of Southampton

Abstract

Faculty of Social Science
Mathematical Sciences

Doctor of Philosophy

Bayesian Multivariate Normal Modelling

by Sthaporn Thepsumritporn

Water pollution is an important global problem. Mostly, it is contaminated by excessive artificial chemicals from human activities. As a result, it will be harm to people and ecosystems. In order to prevention, we need to estimate the quantity of these pollutants which have been circulated along water cycle and end up at water resources.

We use the data provided by Christchurch Harbour Macronutrients Project in the UK and and the Hubbard Brook Ecosystem Study in US to construct a statistical model describing the amount of chemicals in rivers and precipitation. Moreover, we have to deal with uncertainties which affect to chemical concentration such as seasonal, storm, etc. In addition, examining the water sample is expensive then the most collected data are limited and ecology data itself has many specific problem such as extreme outlier, missing value, or truncated data. This uncertainty may be dealt with the Bayesian hierarchical model.

We are focusing on constructing Bayesian multivariate Normal model to use the most information based on geography of catchment area. As a result of fitting Bayesian multivariate normal model,the model has better performance comparing with individual models. Moreover, it shows a better performance if more sources are included into the same Bayesian Multivariate normal model. In addition, the annual chemical loads are calculated from the selected model including with credible intervals of total loads which can not be obtained from individual models.

Contents

List of Figures

List of Tables

Declaration of Authorship

Print name: Sthaporn Thepsumritporn

Title of thesis: Bayesian Multivariate Normal Modelling

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission;

Signed:..... Date:.....

Acknowledgements

I would like to thank my supervisor Professor Dr. Sujit K Sahu for his great understanding, suggestion, supporting, encouraging, and patients while working.

I am very grateful to Professor Dr. Dankmar Böhning for his kind, advice, and appreciation to me.

Special thanks to my officemates for sharing time and knowledge together. Also thanks to Kulvir , postgraduate research administrator, for her cordial response to me.

Finally, thanks to my parents and my family for their support and encouragement.

Chapter 1

Introduction

This thesis mainly focus on analysis and modelling of chemical elements such as Nitrate and Sulphate concentrations in ecological systems, particularly with Bayesian hierarchical modelling to deal with uncertainty events. We also aim to quantify these element fluxes in river and precipitation. Large amounts of these elements (sometimes called nutrients or solutes) are consumed by plants for their growth. However, the excess amounts may harm to the ecosystem and results in human health and well-being. In addition a case of small samples, including more information by geography into modelling may provide the better in prediction and flux estimation. Next section is giving details about importance of chemical elements and its transportation within ecosystem and surrounding environment.

1.1 Chemical and particles in water

Water is a crucial element for every ecosystem. It helps to maintain living bodies and flows the system. It is constantly and continuously recycled in the ecosystem through the process, called the water cycle (Figure 1.1) as a part of biogeochemical cycles.

It is not only the water but also chemicals are passing back and forth between living organisms and non-living matters while being recycled through nutrient cycles (Figure 1.2). Nutrients are essential chemical substances for body growth and maintaining functions of living organism. They are conserved and recycled in different forms through cycling. It can also be transferred to other ecosystems such as the Earth system. Mostly, nutrients are transporting with carrier such as water which is also a chemical nutrient. Figure 1.3 displays sources, transportation, and Removal of nutrients as a part of chemical nutrient cycling. Sources of nutrients may come from nature such as leaching of calcium from soils, otherwise from human activities such as agriculture and industry. The surplus man-made chemicals may disturb and change nutrient concentrations in

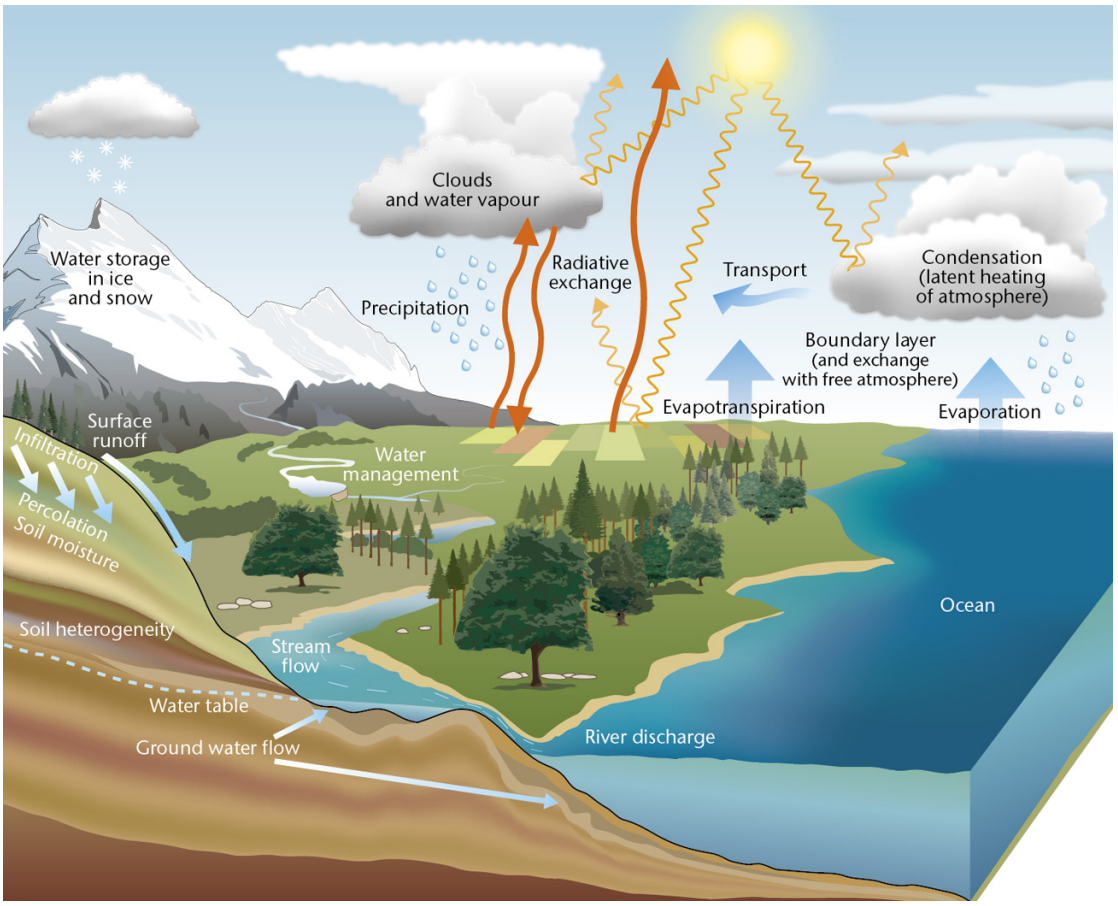


FIGURE 1.1: The water cycle (Image source: Met office, UK)

nutrient cycling. Consequently, it can cause a health problem of both human and environment.

As seen in the water cycle, chemicals can deposit from atmosphere with precipitation. Precipitation (such as rain, snow, etc.) is one of fresh water resources. It is a part of water cycle in all water forms that fall from the sky in liquid or solid such as rain, snow, hail, and fog. It falls down from atmosphere to the earth surface and moves into stream, rivers, water reservoirs and back to the ocean. Precipitation forms within cloud (or atmospheric water vapour) before condensing from atmosphere down to the Earth's surface under gravity. Hence, the atmospheric substances can be transferred to aquatic and terrestrial ecosystems.

Chemical reaction of some compounds such as sulphur dioxide and nitrogen oxides can form acidic pollutants in the air then becomes acid rain or acid deposition. It flows across the surface water reservoirs through the water cycle and causes a great impact on aquatic and terrestrial ecosystem. It pollutes the water toxic to aqua animals. Consequently, it may seriously affect the health of mammals, fish, and birds. It also depletes essential nutrients from the soil and release aluminium which harm to plant hydration process.

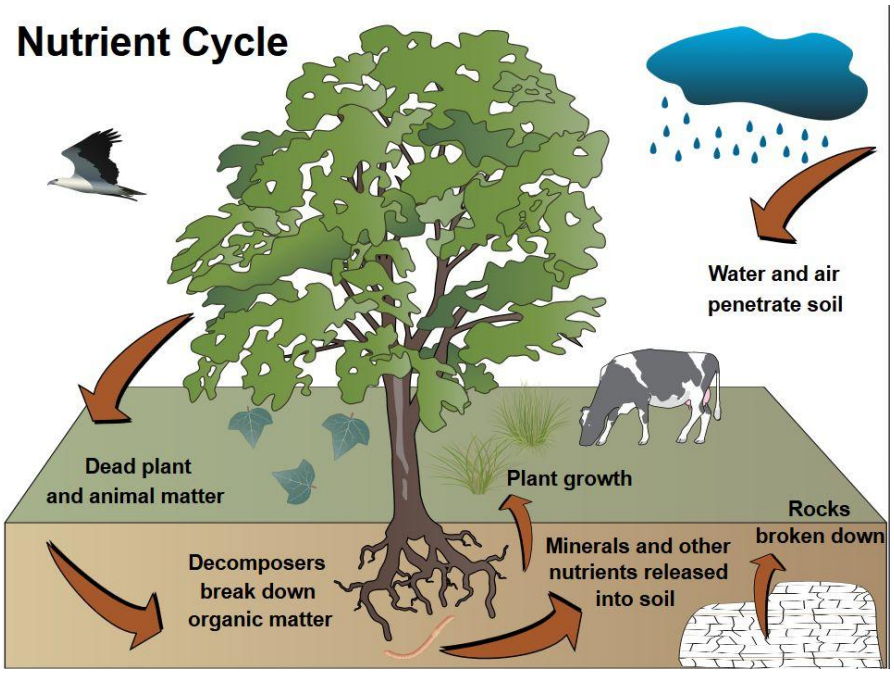


FIGURE 1.2: Nutrient cycle (Image source: Earth Sciences, Freie Universität Berlin)

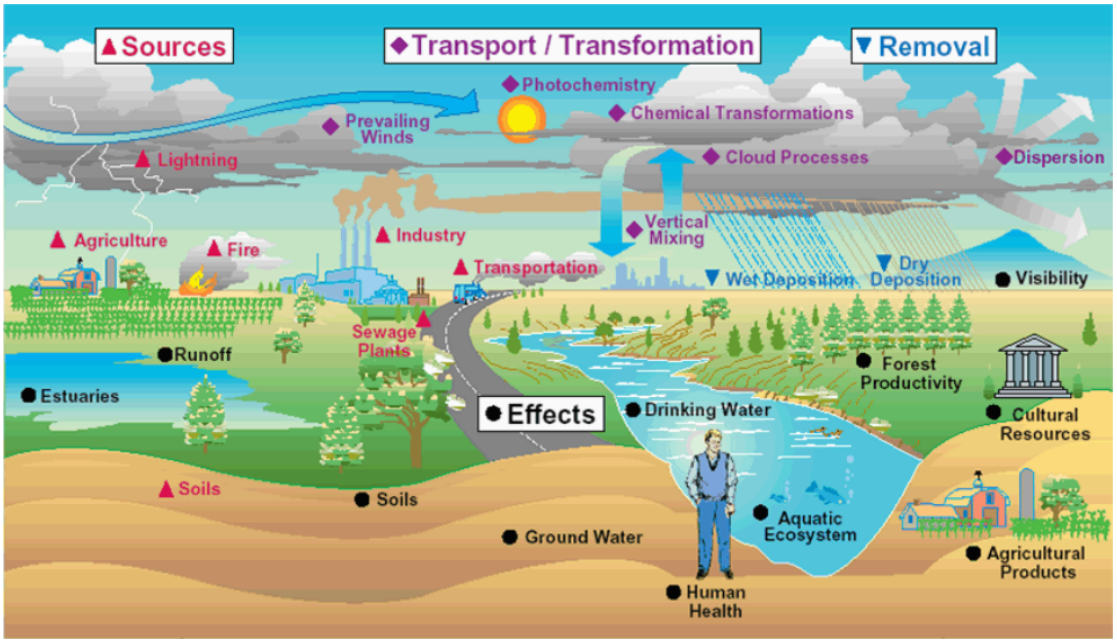


FIGURE 1.3: Pollution pathways (Image source: IntechOpen,UK)

Moreover, it appears an interaction with climate change and ozone depletion. (Galloway & Cowling 1978, Gorham 1998)

In addition, the excess artificial nutrients such as nitrogen and phosphorus from human activities may induce excessive growth of algae in water source which may result in water quality degradation and an aquatic pollution from nutrient enrichment, Europhication (Figure 1.4) .

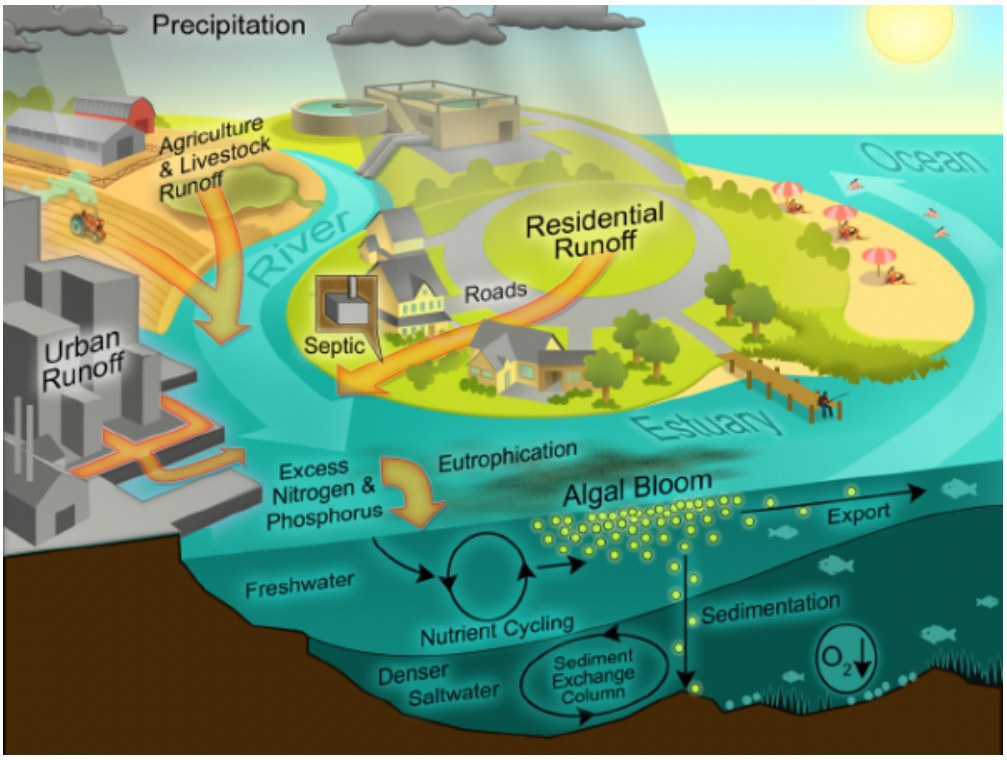


FIGURE 1.4: Eutrophication and nutrient cycling in aquatic system (Image source: https://projecteutrophication.weebly.com/)

1.2 Problem statements and purpose of study

Chemical pollution in water sources is a major problem spreading around the world. Living thing can not survive without water. However, chemical contaminated water can be harmful to creature life and ecosystems. Therefore, we are interested in modelling these chemical concentrations and quantify the annual fluxes to the ecosystem. This may provide an alternative point of view to understand effect of environment such as storm, natural substances, and human activities to chemical concentrations in water sources, especially in river and precipitation.

In this study, we use two datasets from the Christchurch Harbour Macronutrients Project and the Hubbard Brook Ecosystem Study to estimate chemical loads in rivers and precipitation, and calculate total loads transporting into target areas.

The Christchurch Harbour Macronutrients Project is an example study which focus on chemicals transferring from rivers to the coastal area. Figure 2.1 shows the geography of the study area and monitoring sites which macronutrients are loaded from the two major rivers (the Avon and the Stour) before sinking at Christchurch Harbour in Dorset and partly exchanging to the sea. In order to estimate macronutrient loads, We mainly consider the first part of the transferring process or quantifying the amount of macronutrients in rivers which mostly are from run-off the land. Typically, summing up each

macronutrient concentration over year, it provides annual macronutrient loads. However, if we need to predict that annual loads, we may fit macronutrient models to each rivers independently and estimate annual macronutrient loads as a point estimation. Unfortunately, it will not offer a tolerance of total annual load estimates from both two rivers. In addition, due to a small data issue as using only a year of weekly data, we cannot capture the seasonal pattern as it occurs once a year. Thus, we cannot assess the variation of season though there is evidence of seasonality. Moreover, macronutrient concentrations may be affected by uncertainties or natural disturbances such as local heavy storms which usually occurs in this area that may stair macronutrients and sediments from the river bed increasing concentrations in water samples. Furthermore, with axillary data from data collection such as river flow rate, water temperature, and conductivity, we may gather important information to construct the model.

Next, the Hubbard Brook Ecosystem Study is focusing on the effect of chemical contents to the forested ecosystem. Rain gauge stations were set up over watersheds in valley to collect precipitation as a chemical transporter from the atmosphere providing long-term and well-controlled collections. Due to lacking of related variables, we may consider time series approach to model chemical solutes. We use the most recent and complete weekly chemical concentrations data from several rain gauges. Similarly with previous dataset, it is likely to produce annual loads each solute each rain gauge but not interval estimation. In addition, we may take advantages of moderate correlation among selected solutes such as Calcium, Sulphate, and Nitrate, as additional information for model constructing.

In order to better coping with some sorts of uncertainties, we are interested in constructing chemical concentration models using Bayesian approach. Moreover, instead of providing individual models to quantify each chemical each monitoring station, combining related chemical models are also constructed and investigated the performance of modelling. In other words, combining model may take advantage of additional information based on relationship among chemicals in different views such as similar natural environment in modelling. This is supposed to provide better estimates of chemical concentrations, even if it is a small sample size sampling. Regarding the best model, we can calculate the annual chemical loads with credible intervals of estimation.

1.3 Review of Modelling strategies for chemical concentrations

Emerging of excess chemicals problem to the environment and human, it is vital in tracking the fluxes and quantities of chemical nutrients in an ecological system depending on the nutrient state. For instance, estimating nutrient loads from rivers to an estuary or a coastal line to understand the nutrients delivery process. This refers to the annual

flux of chemical nutrients. In addition, uncertainties or natural disturbances, for example, seasonal and climate change can cause temporal dynamic changes of nutrient concentrations.

There are a number of methods used to investigate and quantify the nutrient fluxes, for example, mathematical equations or analytic models, simulation models, and statistical models. The analytic model is suitable for a simple system and can be described by a set of mathematical equations. The widely used simulation model is more complex and realistic as the models are validated by accurate acceptances with the target system. Most of them are deterministic models with a set of input and output equations; for example, different nutrient sources, removals, and water discharge. It is an effective model to cover the nutrient cycle, but less flexible and may not respond to nutrient dynamic uncertainty. To simplify nutrient models and deal with uncertainties, it is interesting to model nutrient concentrations using statistical models.

It is not only uncertainty issue, but some characteristics also happen to water resources data such as non-negative data, positive skewness, non-normal data, outliers, missing values, autocorrelation, and dependence on uncontrolled variables. (Helsel & Hirsch 2002) Besides these issues, there are some temporal disturbance such as local storm events and climate change effect.

Some statistical methods are applied to model the nutrient concentration from traditional methods, for instance, multiple regression and time series analysis (Schoch et al. 2009, Stenback et al. 2011); to advanced methods of nonlinear model that response to the mean shifting and dynamic events. (Pirani et al. 2016).

A Bayesian approach provides a possible way to deal with the uncertainty in the model through the prior distribution. It also allows to estimate missing values with the posterior distribution. In the other word, it allows to evaluate the uncertainty and describe it in a probabilistic manner.

In addition to estimate the annual chemical loads from several sources, we may construct a model to produce interval estimation using the benefit of data dependency such as a similar fluctuation in chemicals and corresponding data from different samples different monitoring stations. It is possible to combine those models by including dummy variables as indicator of different sources. However, the tolerance of total chemical loads across sources is still not being estimated. This suggests us to consider the combination of random vectors through the joint distribution under Gaussian process with covariance matrix description, the multivariate normal distribution for our interest. In addition, we may consider this distribution as a general form of one-dimensional normal distribution with zero covariance for independent data.

The multivariate normal distribution is widely used in machine learning, economic and business data for latent construct of correlation among the dimensions. As far as the

literature is concerned, there is no presence of multivariate normal distribution for joining chemical concentration models. However, there is widely uses with the same data of independent variables for different responses, for example, pre and post treatment results of the same patients. (Tiao & Zellner 1964, Geisser 1965, Tiao & Box 1981, Hervé et al. 2018)

This thesis focuses on analysis and modelling of chemical concentrations in the ecological system. In particular, we study chemical levels in certain water carriers such as rivers and precipitation delivering those contents to the system. We are also interested to estimate annual loads of the chemical concentrations. Naturally, these chemical contents are enough to support ecological cycle. However, human activities, for example industry, harvesting, farming and living, produce a lot of artificial chemical compounds. This excess volumes release to the environment with or without control, become pollutant of the ecological system. A simple reflection to human is the degradation of fresh water quality for drinking water.

As a response, there are several projects and collaborative researches been established to understand any circumstances dealing with this environmental problem, for example, the Christchurch Harbour Macronutrients project in United Kingdom and the Hubbard Brook ecosystem study (HBES) in United States.

Chapter 2

Data Description

This chapter provides details about two study datasets: the Christchurch Harbour Macronutrients Project and the Hubbard Brook Experiment Study. The former project offers one year of weekly data of measurements and laboratory results of water samples from two monitoring sites on two rivers; consists of two nutrient concentrations in rivers, nitrate (NO_3) and phosphate (PO_4) concentration, also related data such as river flow and water temperature. The latter study offers 11-year time series weekly concentrations of 7 solutes in precipitation.

2.1 The Christchurch Harbour Macronutrients project

The Christchurch Harbour Macronutrients project[1] under macronutrients cycles programme (MC) funded by National Environment Research Centre (NERC). The project aims to understand the behaviour of macronutrients cycle and transportation from the catchment of rivers to the Christchurch Harbour estuary, under the assumption that there is a significant impact of stochastic storm-driven events. In addition, teams of scientists planned to study sediment re-suspension and the role of phytoplankton in macronutrient cycles within the estuary.

2.1.1 Study Sites

The Christchurch Harbour estuary is a natural harbour located on the south coast of England in Dorset, covers an area of 2.39 km^2. The estuary is loaded with freshwater from two rivers: the Hampshire Avon from the North and the Dorset Stour from the North west. They joins in Christchurch district before feeding the estuary. Water in the exchanges with coastal water via a channel named "the Run" at Mudeford Quay.

[1] Project web pages: http://www.christchurch-macronutrients.org.uk/

The Hampshire Avon (Avon for short) is a 96 kilometres long river that flows through the county of Wiltshire and is fed by Chalk springs covering the 1,706 square kilometre catchment area (Nedwell et al. 2002, Jarvie et al. 2005). There appears little amount of sediment in water content and the flow rate is steady. Water resources in catchment area are mostly used in water supply e.g drinking water for the county of Salisbury and Warminster. Consequently, the water pollution on Avon is of great concern since the publication of the Hampshire Avon Catchment Management Consultation Report prepared by Wessex Region in October 1992; see reference.[2] A monitoring site is set up at Knapp Mill adjacent to the Environment Agency water gauging station (50°44' 8852 N, 1°46' 8119 W) within the Bournemouth Water industrial site.

The Dorset Stour (Stour for short) is a 97 kilometres long river which rises at artificial lakes at Stourhead in Wiltshire and flows south into Dorset across land areas of clay soil (mainly chalky clay). Consequently, the water body is enriched with suspended sediment and mean river flow varies in large scale from very low flow in summer to often high flow and flood in winter. The 1,073 square kilometre catchment area is a mixture of cultivation land producing crops and a small number of livestock firms (Jarvie et al. 2005). A monitoring site is located at Throop the lowest gauging station of the Environment Agency (50°45' 8290 N, 1°50' 5158 W).

Figure 2.1 shows map of both river monitoring sites. There is also an estuary monitoring site at Mudeford Quay. Since our objective is to investigate and predict nutrients fluxes from both rivers, the Mudeford Quay data is not our interesting. It is also unavailable during the study period.

2.1.2 Data Collection

Data has been collected under two monitoring programmes: one intensive high frequency programme and another low frequency weekly programme. The intensive programme requires continuous monitoring techniques providing *in situ* high frequency data, for example, automatic instruments named YSI EXO2 multiparameter water quality monitoring sonde (Xylem, UK) detecting water qualities using sensors and ISCO automated water samplers (RS Hydro, UK) taking water samples automatically on specified time intervals e.g. every hour. The weekly programme is a typical manual method that collects water samples on weekly basis at low tide times from each of these monitoring sites for further analysis. In this report, we first analyse the weekly data.

Water samples are analysed in laboratory for quantifying chemical species. The results are classified into 3 groups: physical (water temperature and turbidity), chemical (conductivity and dissolved oxygen), and biological (chlorophyll *a*) characteristics. In addition, a physical characteristic, daily mean river flow rate within the collection period

[2] http://www.environmentdata.org/archive/ealit:2542/OBJ/20000998.pdf

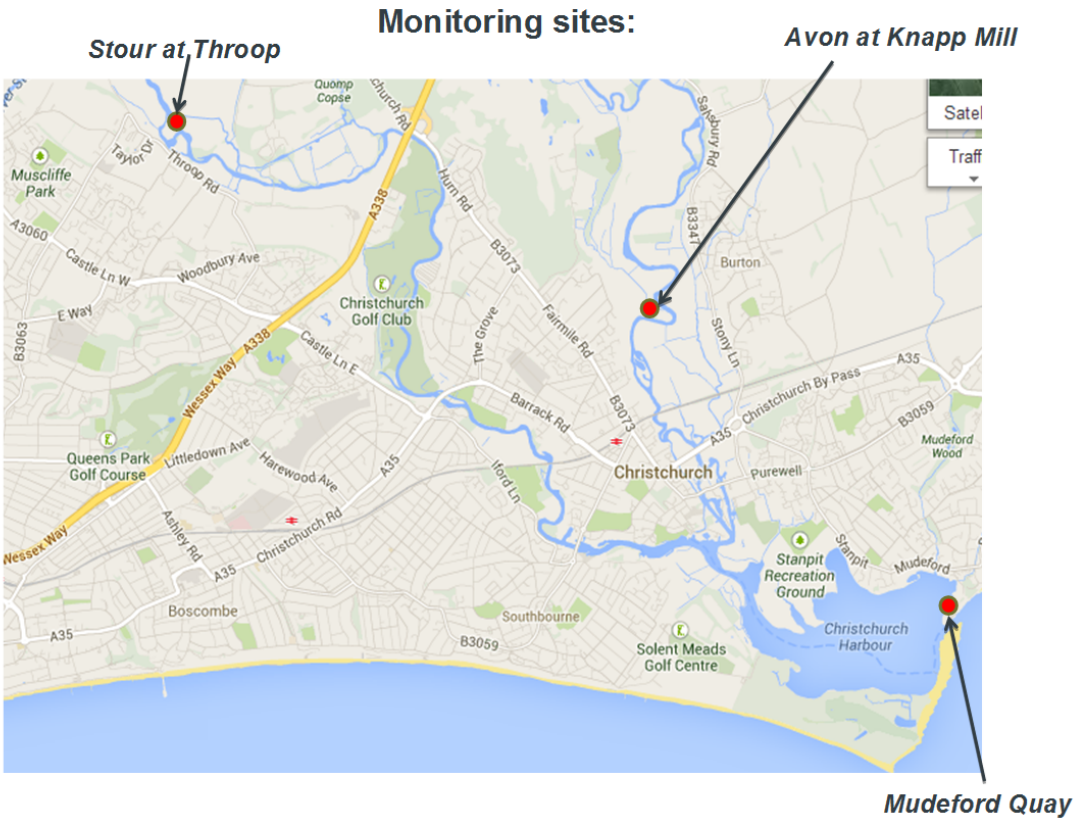


FIGURE 2.1: Monitoring sites for the Christchurch Harbour Macronutrients project

obtained from the Environment Agency (EA) has also been included into this study. For weekly data analysis, only flow rates on the collecting date would be represented for each week.

Weekly sampling was in operation for a year from 26 April 2013 to 10 April 2014. This provides 50 samples without sampling during the Christmas week (week 36) since laboratories were not available and the water samples cannot be preserved for further analysis. To maintain time series structure, the lost week will be taken into the analysis with the available mean river flow rate data.

2.1.3 Exploratory Data Analysis

To study riverine macronutrients behaviour, summary statistics of the main characteristics have been explored and obtained as a modelling guidance. We also concern on seasonal variation and storm events which may take permanent or temporal affect to in-stream macronutrients levels. In our study, we use weekly data that collected as part of the Christchurch Harbour Macronutrients project during 26 April 2013 to 10 April 2014 to develop stochastic models for nitrate and phosphate concentration as major nutrients.

It is basically that season accounts for the great change in ecosystems. It may also directly or indirectly relate to nutrient levels and water qualities through varying natural events and human activities. In the UK, there are four primary seasons: Spring (March-May), Summer (June-August), Autumn (September-November), and Winter (December-February). In brief, Spring is a transition season from winter to summer which reverses in Autumn. Spring is a time for blossoming and flowering. The warmest season is summer when there is a possibility of thunderstorms and heatwaves. In Autumn, the weather slowly approaches winter, temperature dips, and slowly becomes unpredictable. Atlantic depressions in late Autumn result in storms and strong wind. Most weather disturbances happen during the winter, the coldest season. The strongest winds mostly occur in stormy winter caused by depressions.

In accordance with severe weather, there were major storms appear within the sampling period, from mid December 2013 to early January 2014 and from late January to mid February 2014.[3]

Descriptive statistics of all measured water qualities from both monitoring sites (51 records each) are shown in Table 2.1 including those time series plots in Figure 2.2 to 2.8 for visual inspection plotted with seasons (vertical dotted guideline) and storm period (grey zone).

We briefly describe characteristics of water qualities in following paragraphs. Starting with macronutrients concentration as our interested variables and then other five water qualities related with those nutrients as essential explanatory variables.

TABLE 2.1: Descriptive statistics of water quality characteristics

Variable	**Unit**	**Min**	**Median**	**Mean**	**Max**	**SD**
Nitrate	Milligram per litre (mg/l)	4.00	6.11	6.19	9.43	1.23
Phosphate	Milligram per litre (mg/l)	0.0028	0.0886	0.1841	0.7795	0.201
Daily mean river flow	Cube metre per second (m^3/s)	2.43	11.15	24.75	119.90	28.45
Temperature	Celsius (°C)	5.74	12.48	12.68	22.96	4.73
Turbidity : SPM	Gram per litre (g/l)	0.0006	0.0049	0.0103	0.1346	0.016
Conductivity	Millisiemens per centimeter (mS/cm)	0.28	0.42	0.43	0.59	0.07
Dissolved Oxygen	Milligram per litre (mg/l)	7.29	10.38	10.42	15.73	1.7

In Table 2.1, nitrate concentration approximately averages 6.19 mg/l (sd = 1.23 mg/l). Considering Figure 2.2, we observe that average level of nitrate at Throop is higher than Knapp Mill. Both time series also exhibit similar variation pattern. Seemingly, nitrate

[3]Sources: http://www.metoffice.gov.uk/climate/uk/interesting

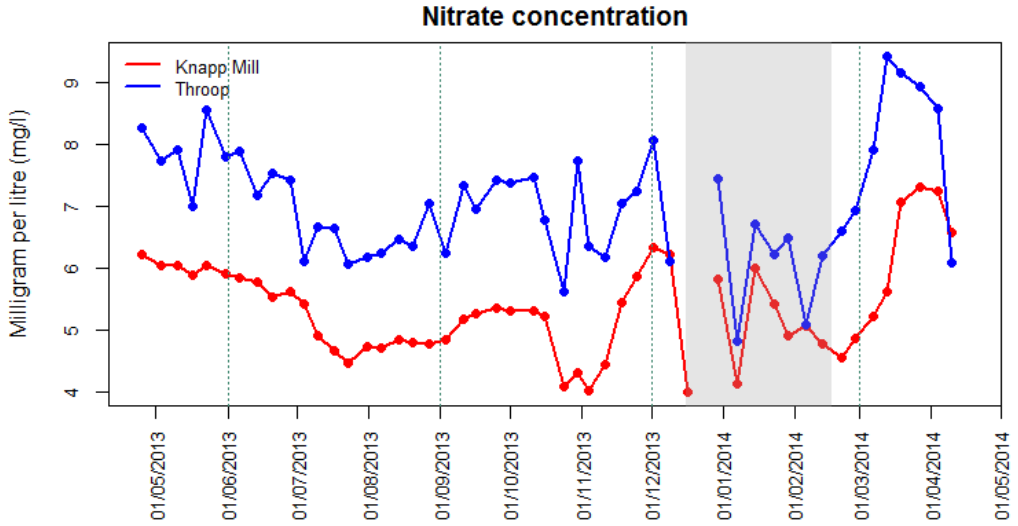


FIGURE 2.2: Time series plot of the weekly sample nitrate concentration with storm period (grey zone)

concentration steadily decreases in spring towards summer as aquatic plants grow well in warm weather and then turn upwards through autumn and winter in which application of fertilisers may be a reason of nitrate supply. Nitrate dilution slightly appears during storm periods (grey zone) which may be caused by heavy rainfalls and floods in winter. As a result, nitrate fluctuation shows a strong seasonal variation.

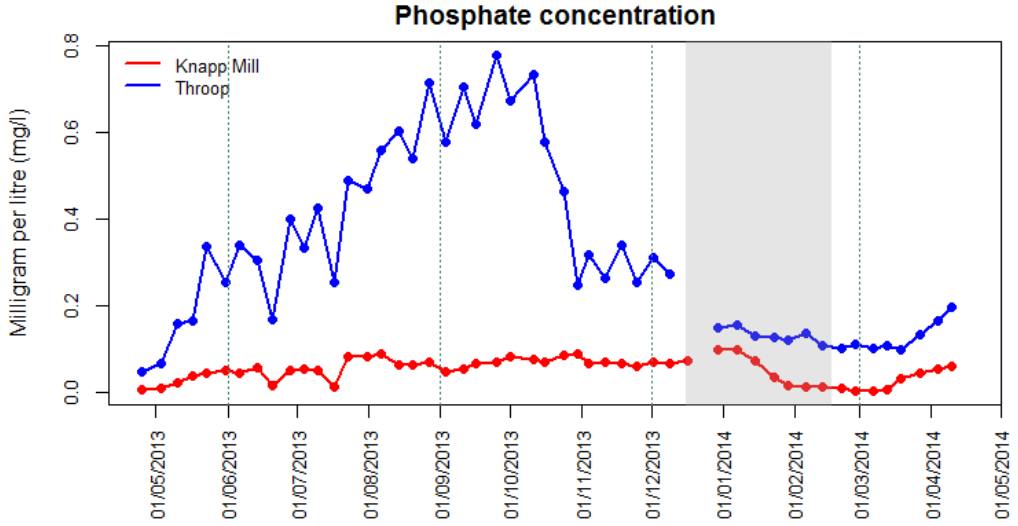


FIGURE 2.3: Time series plot of the weekly sample phosphate concentration with storm period (grey zone)

According to Figure 2.3, the level of phosphate concentration at Throop is much higher than Knappmill with larger dispersion. This difference seems to be a result of drinking water resource preservation on River Avon. The fact that most of the surplus phosphate concentration comes from farming, households and industries in forms of fertiliser, pesticide and detergent, increasing in phosphate concentration starts in spring through

summer. Moreover, the excess untreated phosphorus compounds might also be released into the river from a nearby wastewater treatment works that lead to a dramatic increasing of phosphate at Throop.

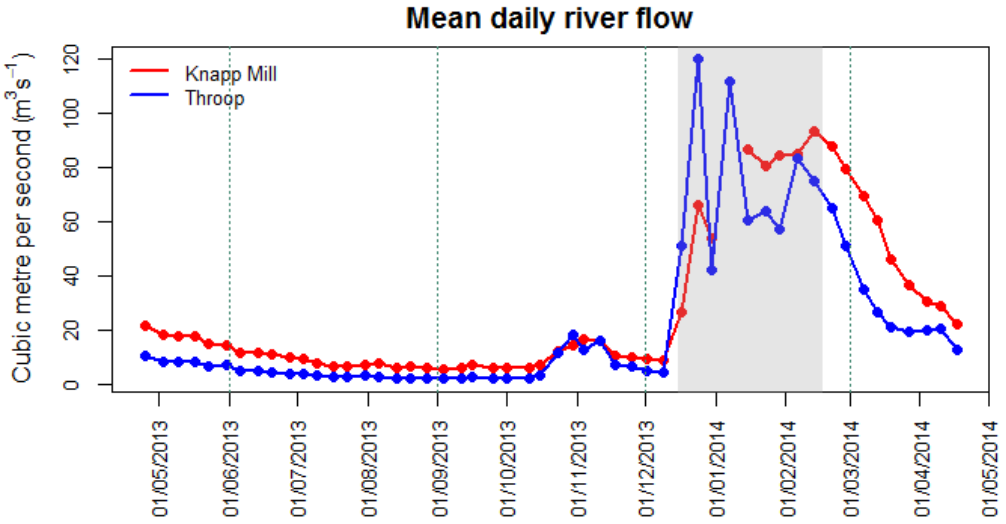


FIGURE 2.4: Time series plot of daily mean river flow with storm period (grey zone)

Most daily mean river flow rates vary in small range below 20 m^3/s. Commonly, flow rates at the Avon is higher than the Stour except in winter with a number of annual flood reports. Extreme flow rates driven by floods and storms during winter as seen in Figure 2.4 reach to a peak at 119.90 m^3/s. As a result in the large difference between median and mean (mean > median) in Table 2.1, it indicates a right skew distribution of daily mean river flow.

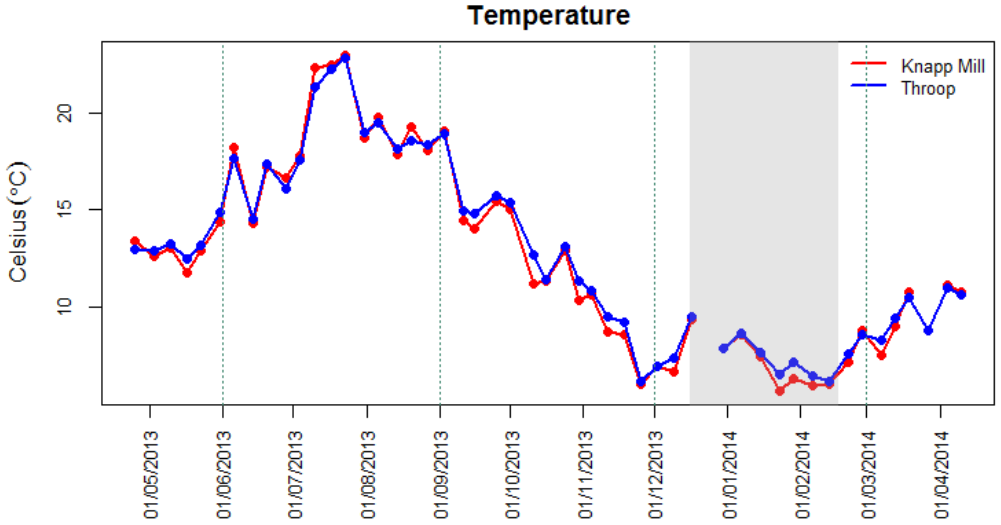


FIGURE 2.5: Time series plot of temperature with storm period (grey zone)

As expected in Figure 2.5, water temperature from both sites display similar seasonal behaviours with average 12.68 (sd = 4.73) in degree Celsius (°C). The temperatures are

higher in summer months and lower in winter. In between, the transition of temperatures and seasons appear during autumn and spring which tending to colder or warmer weather respectively.

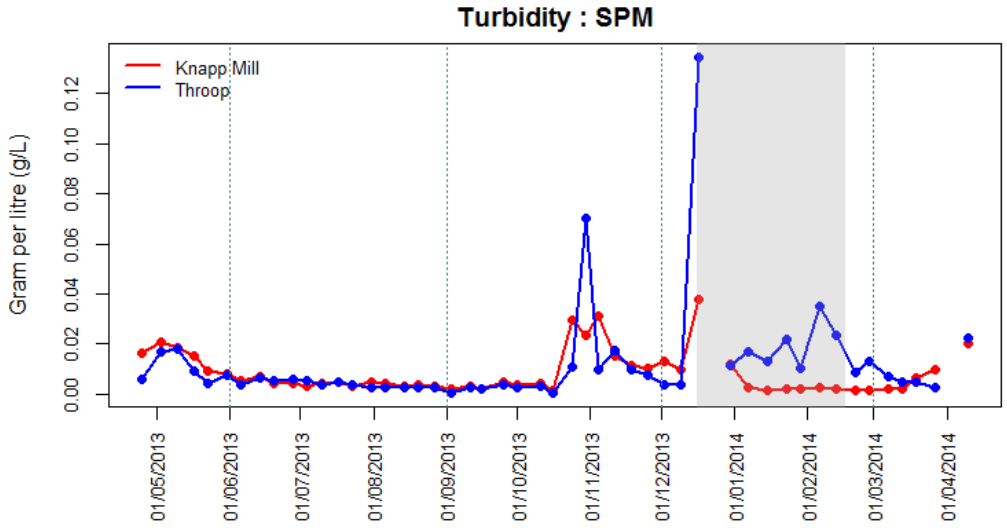


FIGURE 2.6: Time series plot of turbidity (SPM) with storm period (grey zone)

Suspended particulate matter (SPM) is a physical characteristic that indicates how aquatic organism, silica, metals, clay or slit can cause water to be turbid. A lower value of SPM indicates a better water quality. Figure 2.6 shows that SPM value remains stable throughout the year and similar at both sites. However, there is a temporal change when sediments in river bed are stirred up during storm periods and can be increase to 10 times, especially at the Stour.

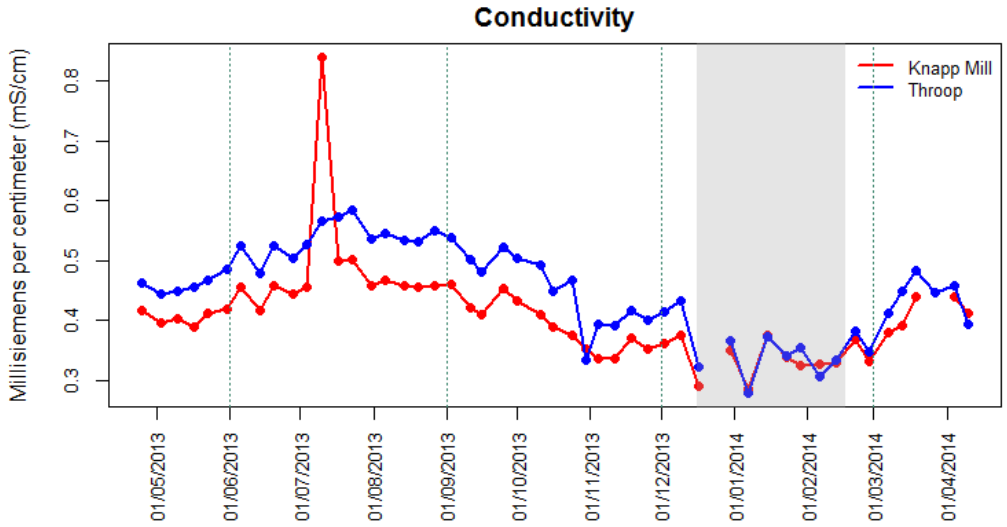


FIGURE 2.7: Time series plot of conductivity with storm period (grey zone)

Conductivity is a chemical variable that refers to mobility of ionic as a good indicator of total salinity or concentration of salts dissolved in water. It has a positive relationship

with water temperature in common which can be seen in comparison between Figure 2.5 and 2.7. The conductivity average on both rivers is about 0.43 mS/cm (sd=0.07 mS/cm), although a higher conductivity value present at Throop. The values fluctuate in a narrow range, although there are some seasonal variations. In addition, there appears an outlier at Knapp Mill in July. This may caused by measurement error.

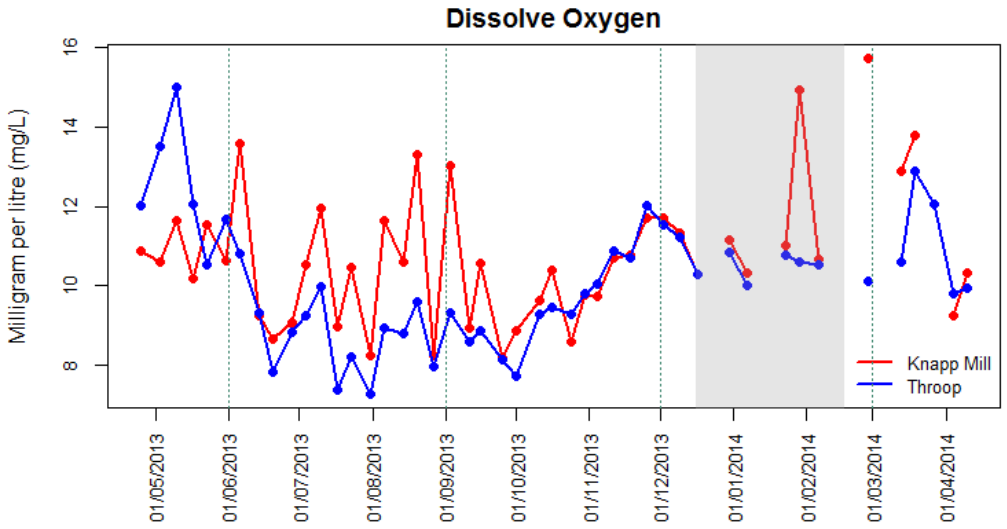


FIGURE 2.8: Time series plot of dissolved oxygen with storm period (grey zone)

Lastly, a chemical characteristic, dissolved oxygen, oxygen in water body survives aquatic animals and bacteria. It is produced by green plants and some bacteria in photosynthesis under the presence of light and chlorophyll. However, those producers also respire or uptake oxygen all the time to energise their cells. Hence, dissolved oxygen will be replenish in day time but consumed all day which may cause oxygen depletion or hypoxia in water in long cloudy days (e.g. in winter). This situation also occurs in warm weather when there is an overpopulation of aquatic plants and animals in the area which a great amounts of oxygen is required by bacteria for decomposition process when those living things die. As a response, most of dissolved oxygen result from biological activities.

Turning to our data, the average level of dissolved oxygen is approximately 10.42 mg/l (sd = 1.7 mg/l). It is seen in Figure 2.8 that there appears a large tolerance and lower oxygen levels in spring/summer months before reaching back to the average level in autumn. However, it cannot describe the oxygen level clearly due to lack of data in winter.

2.1.4 Data Preparation

As seen in Table 2.1, there are some concerns regarding the data arising from various measurement units of water qualities and large difference between mean and median of mean river flow. These require some applications of data transformation. Some common

transformation techniques, for instance, standardising, logarithm, square root, cube root are considered. Standardisation and natural logarithm transformation have been used since these produced the best results. Moreover, it is used most in the literature review.

Water qualities (except daily mean river flow) each rivers have been standardised to zero mean and one unit of variance provided comparative coefficients of various measurement unit variables. It also allows considering different data sources simultaneously for convenience.

Natural logarithm transformation is applied on both macronutrient concentrations and daily mean river flow to obtain constant variances and reduce skewness in the data.

Next, scatter plots and histogram of macronutrients and water qualities on transformed scale from both rivers are shown in Figure 2.9 including the Spearman rank correlation among those water qualities. It shows nonlinear relationship between macronutrients and water qualities. We can also see that the symmetry has been achieved for nitrate concentration; while appearing a slight left skew for phosphate concentration in river.

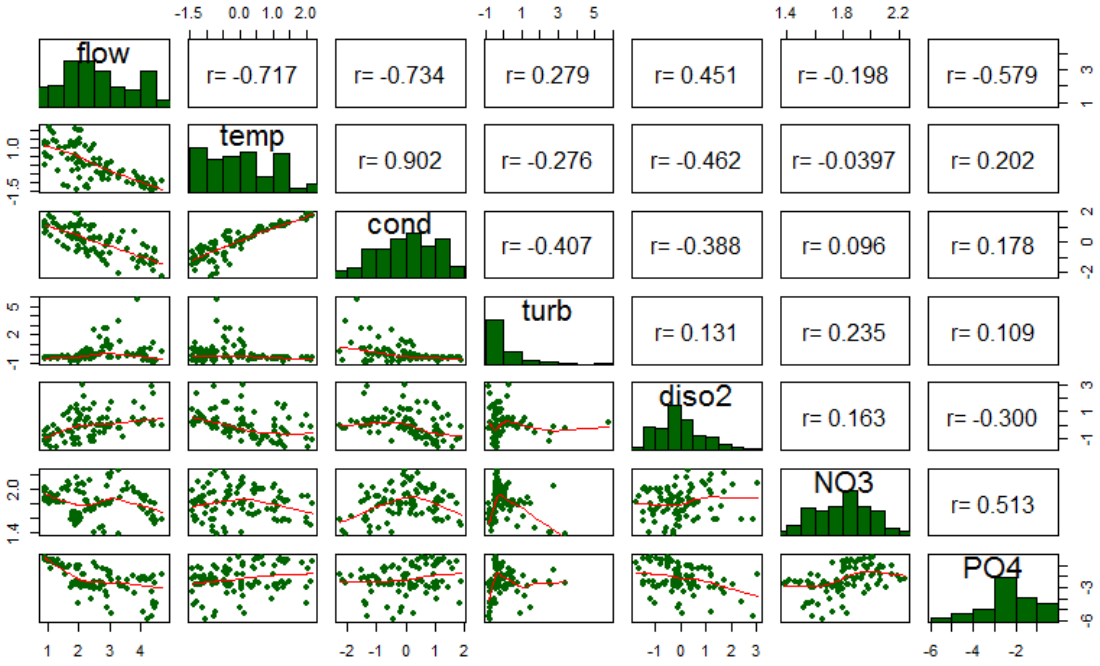


FIGURE 2.9: Pairwise scatter plots of macronutrient levels and water qualities on transformed scale (log scale on macronutrients and daily mean river flow, others in standardised scale) and correspondence correlation coefficients

There is a high positive correlation between water temperature and conductivity (0.902), followed by negative moderate correlations between mean river flow and the two water qualities: conductivity (-0.734) and water temperature (-0.717).

Furthermore, it appears outliers in the data. This may be caused by human errors and/or stochastic events e.g. the winter storm. However, it would not be treated due

to complex structures and lack of data but the obvious one e.g. conductivity shown in Figure 2.7 will be replaced with an average of weekly conductivity of before and after that problem week.

Moreover, some water quality attributes cannot be obtained due to instrument problems and quality of water samples which are primary reasons leading to missing values. In order to treat the data, a sufficient number of data are needed to examine data patterns incorporating with seasonality. Unfortunately, the pattern cannot be captured clearly with one year of weekly data. However, it is possible to model the data with statistical techniques such as Bayesian approach.

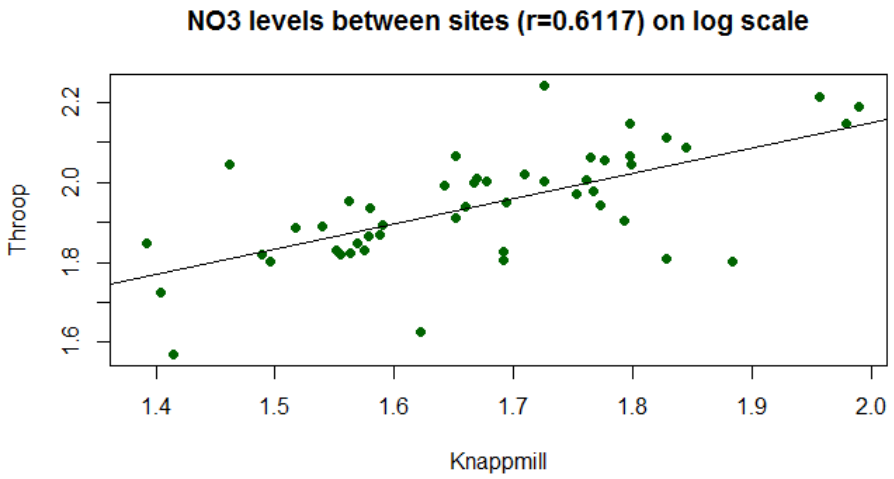


FIGURE 2.10: Scatter plot of log nitrate concentrations between Knapp Mill and Throop sites

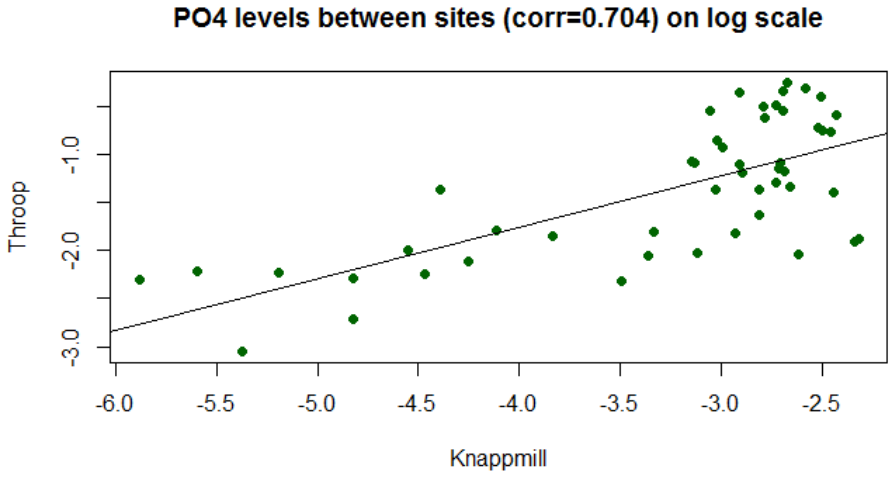


FIGURE 2.11: Scatter plot of log phosphate concentrations between Knapp Mill and Throop sites

In addition, Figure 2.10 and 2.11 shows the relationship between macronutrient levels on log scale from both rivers with $r = 0.6117$ for nitrate and 0.704 for phosphate. This suggests the possibility of bivariate linear regression model.

Overall, weekly data are collected and sampled from two monitoring sites: Knapp Mill and Throop represented the two major rivers: the Hampshire Avon and the Stour respectively during 26 April 2013 to 10 April 2014. The sample is measured and analysed to obtain riverine macronutrients (nitrate and phosphate) concentration and water quality characteristics. Daily mean river flow on sampling date from the EA also included to the dataset as an important variable driven macronutrient levels.

To achieve our aims, the data have been explored their attributes and relationships. Time series plots show similar fluctuations on both rivers for the most variables except phosphate concentrations. It also shows sudden changes in Winter which may be affected by uncertainty events such as higher river flow rate, floods and local storms. It may introduce seasonal and temporal effects on time series data. In addition, scatter plots show the nonlinear relationship between macronutrients and water qualities. It also appears inter-relationship of macronutrients between rivers.

2.2 The Hubbard Brook Ecosystem Study (HBES)

The Hubbard Brook Ecosystem Study [4] is a unique public-private partnership for hydrologic and ecologic research on the Hubbard Brook Experimental forest (HBEF) established in 1955 by the USDA forest service.

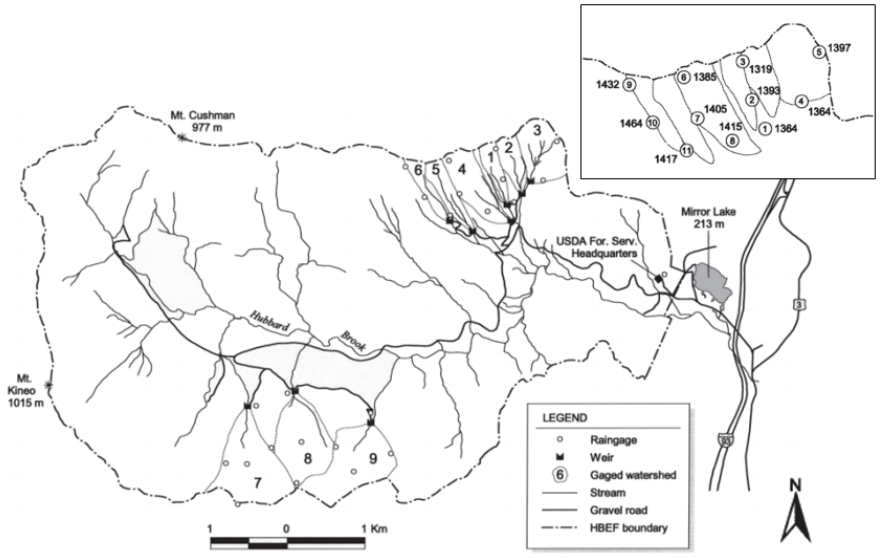


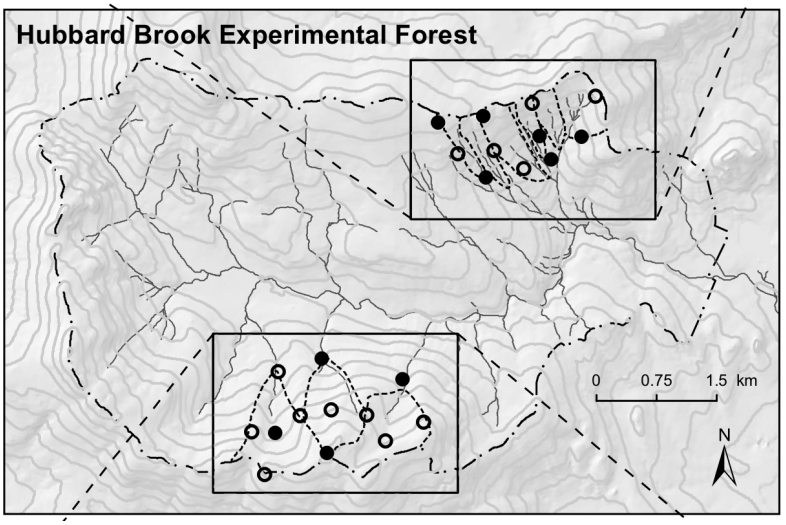
FIGURE 2.12: The Hubbard Brook Experimental Forest map

[4] https://hubbardbrook.org/

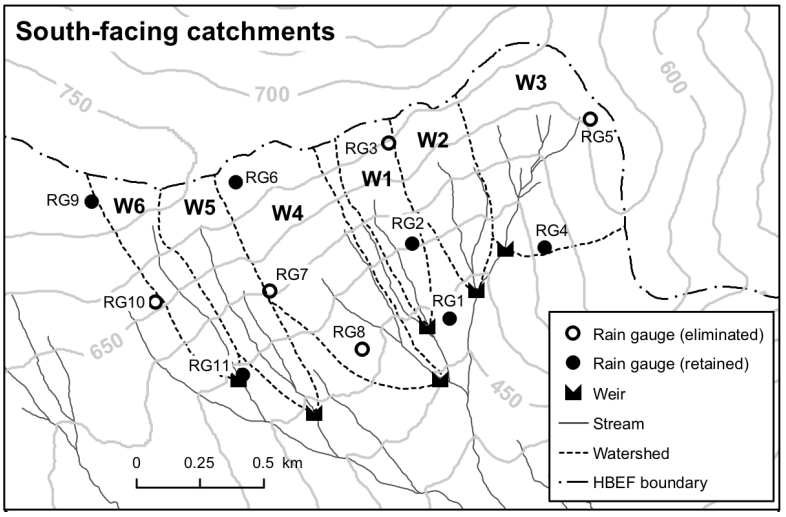
2.2.1 Study Sites

Hubbard Brook valley is a bowl-shaped area in the White Mountain National Forest, Thornton, New Hampshire, United States. It is bounded by the Mt. Cushman ridge on the north and the Mt.Kineo ridge on the south, drained by Hubbard Brook from west to east for 14 km. in distance. It covers an area of approximately 4,000 hector(ha). It is also the location of a 8000-acre experimental forest.

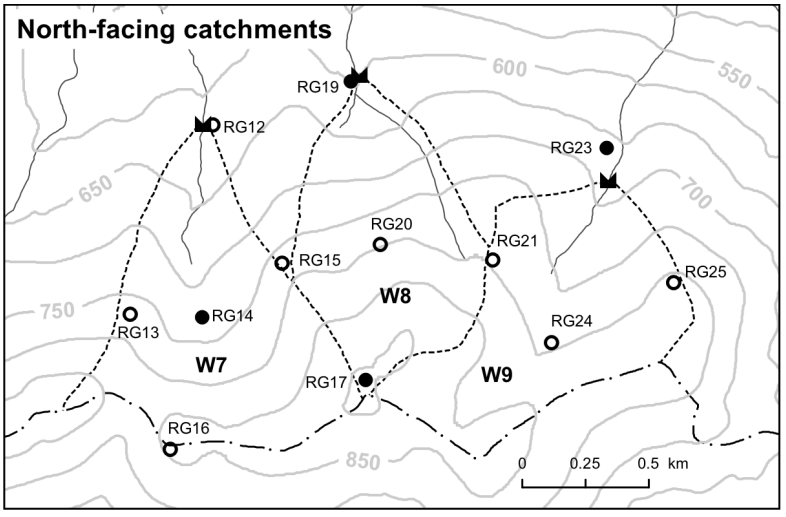
Different types of monitoring stations have been deployed for collecting data since the 1950s, including the small watersheds collection started in the early 1960's. In particular, precipitation has been measured continuously providing long-term records from rain gauge stations installed in watershed areas (Figure 2.12-2.13, Green et al. (2018)).



(a)



(b)



(c)

FIGURE 2.13: (a) The Hubbard Brook Experimental Forest (Watersheds and Rain gauge locations) (b) South-facing catchments (c) North-facing catchments

2.2.2 Data Collection

Precipitation (e.g. rain and snow) is the primary process that cycles water from the atmosphere to the ground, transfers substances in the atmosphere to aquatic or terrain ecological systems. It may include gases, aerosols, and large particles varying by area or storm with different sources from ocean, terrain, gaseous pollutants, and volcano. It is possible to be organic or inorganic composition of precipitation which causes the acid rain as a solution of nitric and sulphuric acids. This acid rain affects to aquatic system and may be harmful to fish and wildlife. The trees are less able to absorb the sunlight. Mineral and nutrients are also removed from the soil. This may cause dead of trees in acid rain area.

Precipitation is collected weekly commonly using "bulk precipitation collector". The collection process was well designed to avoid contamination in the samples (see Galloway & Likens (1978) for more details). Once the sample is collected, it is analysed for chemical contents such as pH, base cations, anions, aluminium and more. This long-term records provide a trend of acid rain within New England region. More details about hydrology, ecology, biogeochemistry and analysis are given by Buso et al. (2000), Bailey et al. (2003), and Likens (2013).

In this study we use weekly collected and analysed data (precipitation and chemical concentrations) from samples between 3/6/2002 and 2/6/2013 for 574 weeks of 11 water years collection from three rain gauges: RG-11 located in South-facing catchments, RG-22 located at the U.S. Forest Service headquarters, and RG-23 located in North-facing catchments. RG-11 is on watershed WS6 (used as a biogeochemical reference watershed) lying in the southern part of the watershed. Noted that RG22 is sometimes excluded from analysis as it is located far from the other catchment areas, close to USDA for service headquarters. Analysis will be studied on water-year basis which runs from June 1 to May 31, when water storage is the most stable than common year.

2.2.3 Exploratory Data Analysis

In order to quantify precipitation solutes at the HBEF, we first examines the basics of the weekly data. Descriptive statistics of precipitation and chemistry solutes are described in table below.

Table 2.2 provides descriptive statistics of precipitation and chemistry samples. The different measurement values among solutes issues usage of standardised solutes. The deviation of median from the mean indicates the right skewness of measurements as the natural property. A variety of data plots can be analyzed to gain a better understanding of data behaviour or data pattern. For example, some extreme measurements can be seen in Figure 2.14(a) and 2.14(b). Precipitation and chemical solutes data are presented

with two different plots. First, a weekly time series plot and a box plot of the entire series are supplied to visualise the data variation and distribution. Second, a box plot of time series by month are also given to consider seasonality pattern as seen in Figure 2.14-2.21. These plots use data from rain gauge 11 (RG11) for example. Results from the other two rain gauges show similarity of RG11

TABLE 2.2: Descriptive statistics of precipitation and chemical solutes

Variable	**Min**	**Median**	**Mean**	**Max**	**SD**	**NA**
Rain gauge 11 (RG-11)						
Precipitation	0.000	22.850	28.920	159.800	26.763	-
Calcium (Ca)	0.005	0.060	0.102	1.660	0.144	108
Sulphate (SO4)	0.020	0.880	1.216	10.300	1.164	108
Ammonium (NH4)	0.005	0.100	0.178	2.120	0.228	108
Nitrate (NO3)	0.01	0.87	1.17	10.20	1.135	108
Potassium (K)	0.005	0.030	0.055	1.000	0.096	108
Sodium (Na)	0.005	0.050	0.101	1.100	0.14	108
Chloride (Cl)	0.002	0.100	0.184	2.250	0.241	109
Rain gauge 22 (RG-22)						
Precipitation	0.000	19.650	25.999	143.500	24.135	-
Calcium (Ca)	0.005	0.060	0.107	1.380	0.132	119
Sulphate (SO4)	0.040	0.935	1.307	7.520	1.171	118
Ammonium (NH4)	0.005	0.110	0.210	3.290	0.313	119
Nitrate (NO3)	0.010	0.940	1.265	7.510	1.197	118
Potassium (K)	0.005	0.020	0.055	1.200	0.115	120
Sodium (Na)	0.005	0.060	0.126	1.390	0.168	120
Chloride (Cl)	0.005	0.120	0.228	2.740	0.277	119
Rain gauge 23 (RG-23)						
Precipitation	0.000	23.850	30.903	170.100	28.603	-
Calcium (Ca)	0.005	0.060	0.096	1.170	0.124	96
Sulphate (SO4)	0.030	0.810	1.170	6.510	1.088	96
Ammonium (NH4)	0.005	0.090	0.167	1.350	0.192	96
Nitrate (NO3)	0.010	0.820	1.088	5.450	0.95	96
Potassium (K)	0.002	0.020	0.053	0.890	0.086	96
Sodium (Na)	0.005	0.040	0.090	1.010	0.126	96
Chloride (Cl)	0.005	0.100	0.171	1.910	0.22	98

Note: solute measurement unit is milligram per litre (mg/l).

In addition, there are some missing values or NA counted about 19% of 574-week data each chemical solutes. One reason of missing is the loss of bulk collections then the sample are rejected without analysis. Due to the bulk precipitation collectors are continuously open to the air and leaving outside during the sampling period, the sample falls under the risk of being contaminated or the collector might be damaged from such as windstorm and hail. However, these missing values will be partially treated later, see data preparation in section 2.2.4 for more details.

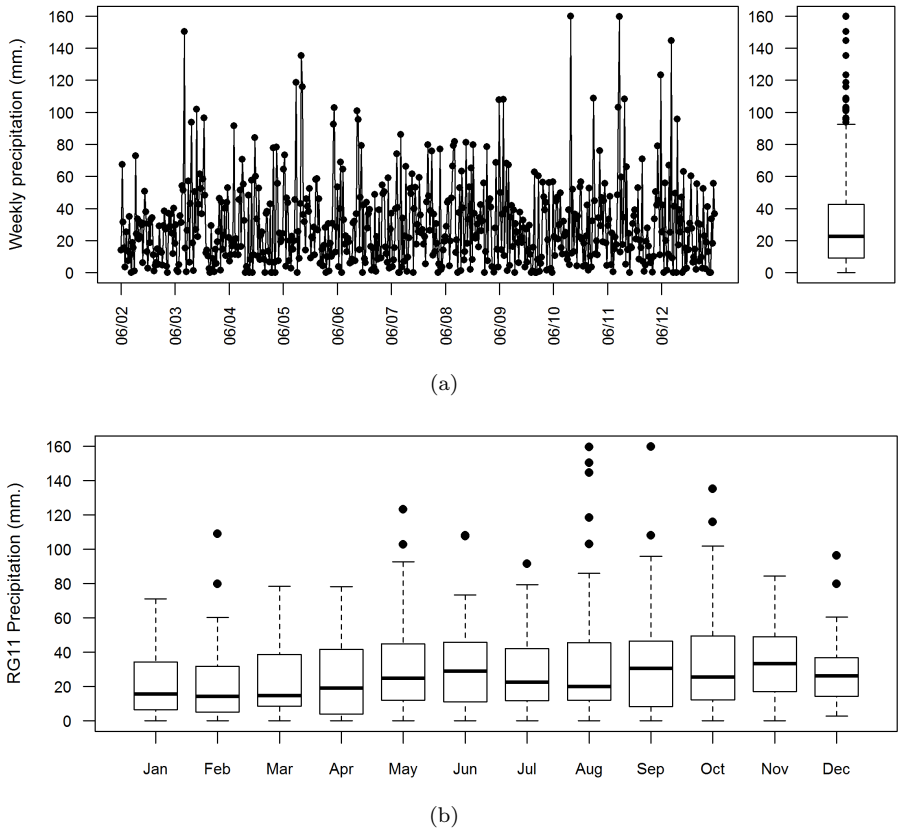


FIGURE 2.14: (a) Time series and box plot of weekly precipitation (mm) (b) Box plot of precipitation by month (mm); RG11 example

From Figure 2.14, it seems to have no trend in weekly precipitation data. Half of records are approximately lower than 25 mm while the average is 28.92 (sd=26.099). This deviation may cause by extreme measurements.

Federer et al. (1990) indicated that the sudden large amount of precipitation may cause by storms which can double the precipitation. The storm can happen at any time of the year. In addition, the high precipitation event in winter/cold months mostly occurs around freezing temperature and prevails rain events. Note that the cold months are approximately November to April, the warm months at the HBEF are approximately from May to October which there is sometime no precipitation. Moreover, orographic effect may vary precipitation by elevation and aspect. However this effect is not included to this study

However, It may consider that the most of weekly precipitation are below 50 mm and remain steady in long term. The seasonal pattern may be not well defined as it may be disturbed by temporal effects such as storm through out the year.

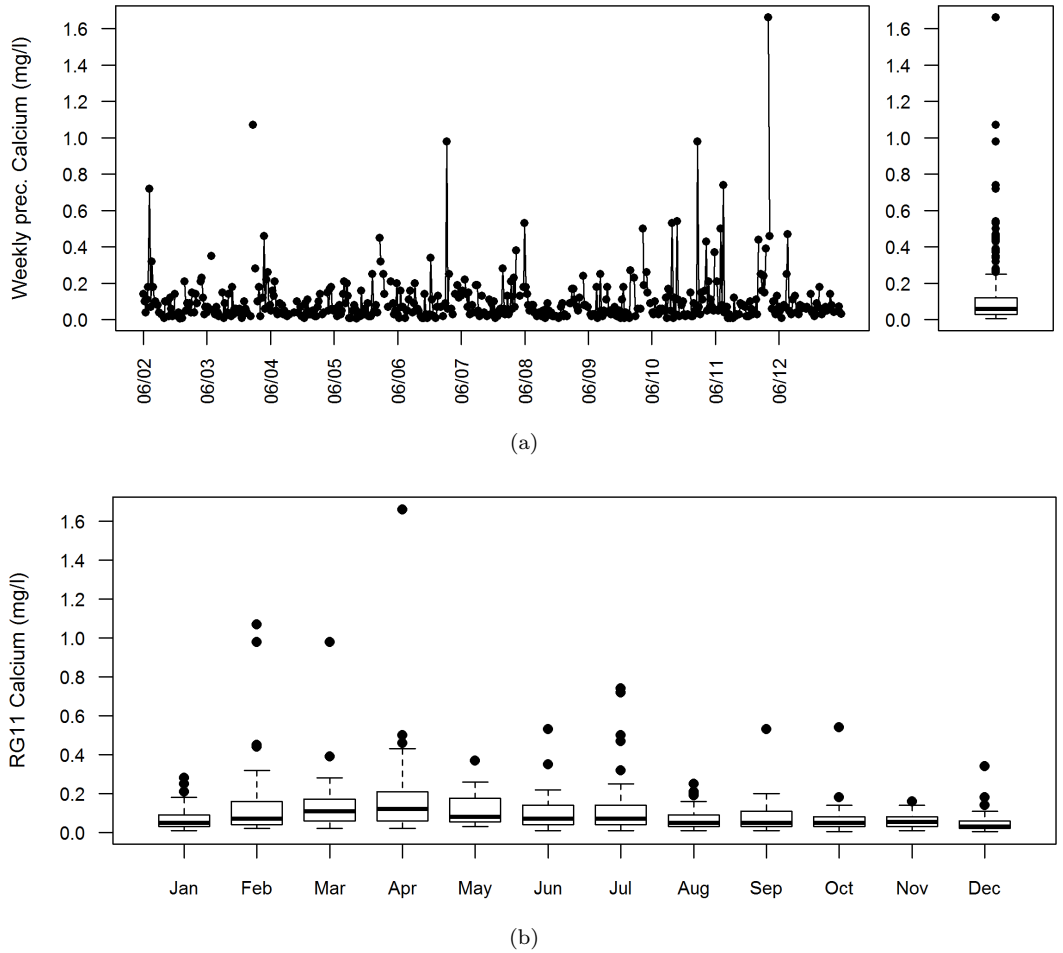
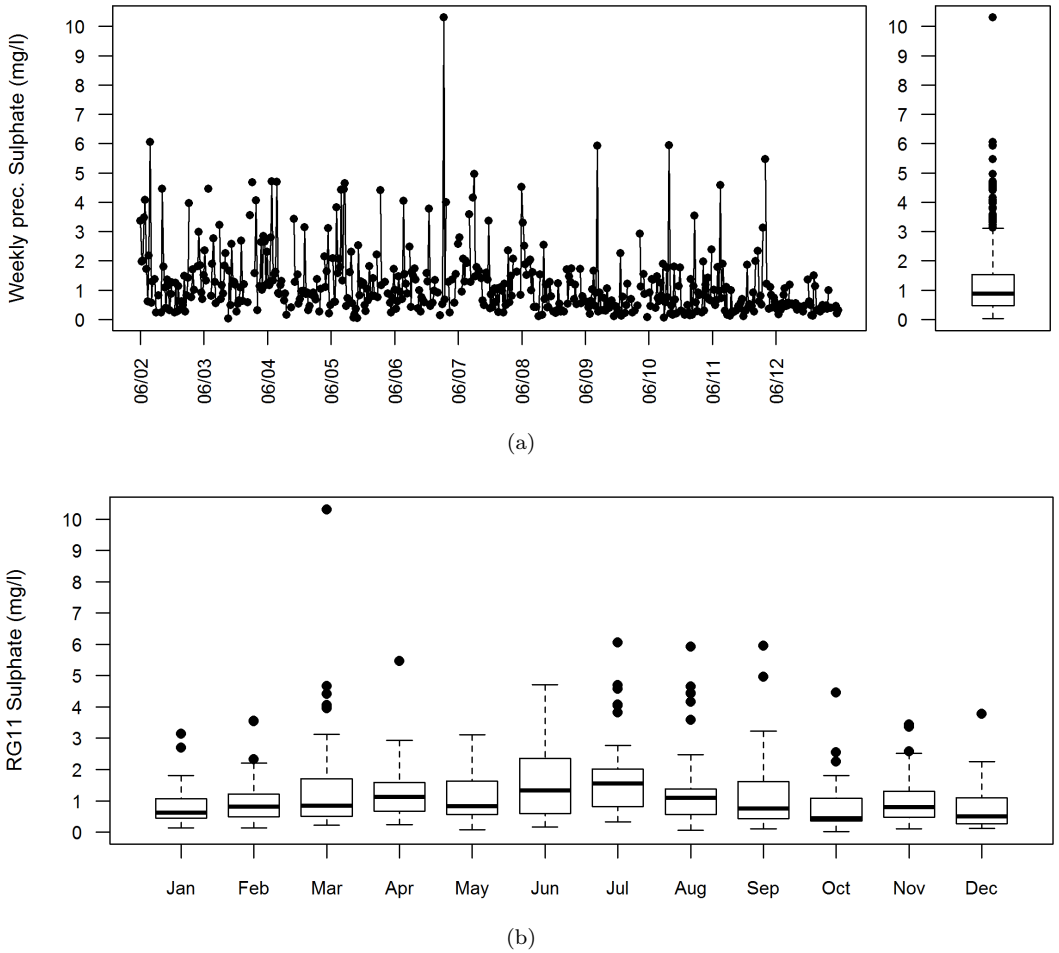


FIGURE 2.15: (a) Time series and box plot of weekly calcium (Ca) concentration (mg/l) (b) Box plot of calcium concentration by month (mg/l); RG11 example

Calcium (Ca) plays an important role in functioning of plant nutrients. It is also a key element in a number of plant processes, for example, cell wall and membrane synthesis. Many evidences place Ca supply as an important limitation in forest structure and function. The fact that Ca deposition by precipitation at the Hubbard Brook experiment forest declines significantly since 1963. More details about calcium and calcium process in terrestrial ecosystem can be found in McLaughlin & Wimmer (1999). In particular, the biogeochemistry of calcium at Hubbard Brook see Likens et al. (1998).

The median and mean of weekly calcium concentration are 0.06 and 0.104 mg/l (sd=0.142 mg/l) respectively. Considering Figure In monthly view, box plots show high positive skewness of chemical concentration each month. The median is almost in the most month except July and December. The larger box is also correspond with period of warm and cold months. This may indicates a relationship between temperature and precipitation chemical concentrations. Note that these acids are in vapour under the high temperature, then becomes droplets when the temperature falls (Singh & Agrawal 2007).



(a)

(b)

FIGURE 2.16: (a) Time series and box plot of weekly sulphate (SO4) concentration (mg/l) (b) Box plot of sulphate concentration by month (mm); RG11 example

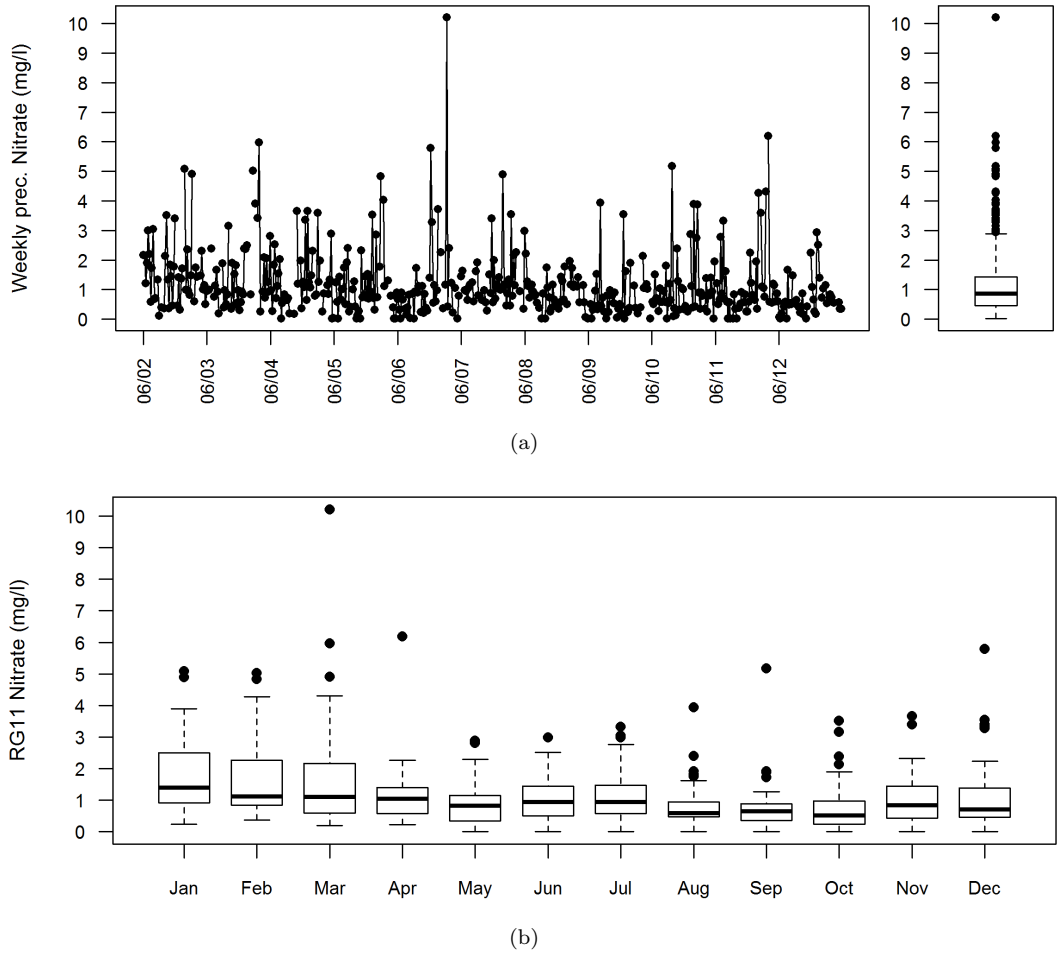


FIGURE 2.17: (a) Time series and box plot of weekly nitrate (NO3) concentration (mg/l) (b) Box plot of nitrate concentration by month (mm); RG11 example

Sulphate (SO4) and Nitrate (NO3) are important chemical compounds indicating the acidity of precipitation in form of sulphuric and nitric acid. Time series plot of sulphate concentration in Figure 2.16(a) and nitrate concentration in Figure 2.17(a) show the declining of concentration and less fluctuation in later year. This context confirms a success story of the Clean Air and Clean Water Acts and other polices in air quality management. (Sullivan et al. 2018) Decreasing in pollutant input also indicates the decline of acid deposition to the Hubbard Brook experiment forest.

The average of concentration is about 1.070 (sd=1.084) for sulphate, and about 1.103 (sd=1.061). The large standard deviation indicates large fluctuation or extreme outlier in chemical concentrations.

Box plots of both chemical concentrations by month show the higher concentrations in June and December, and some outliers through out the year. This pattern is similar with calcium concentration.

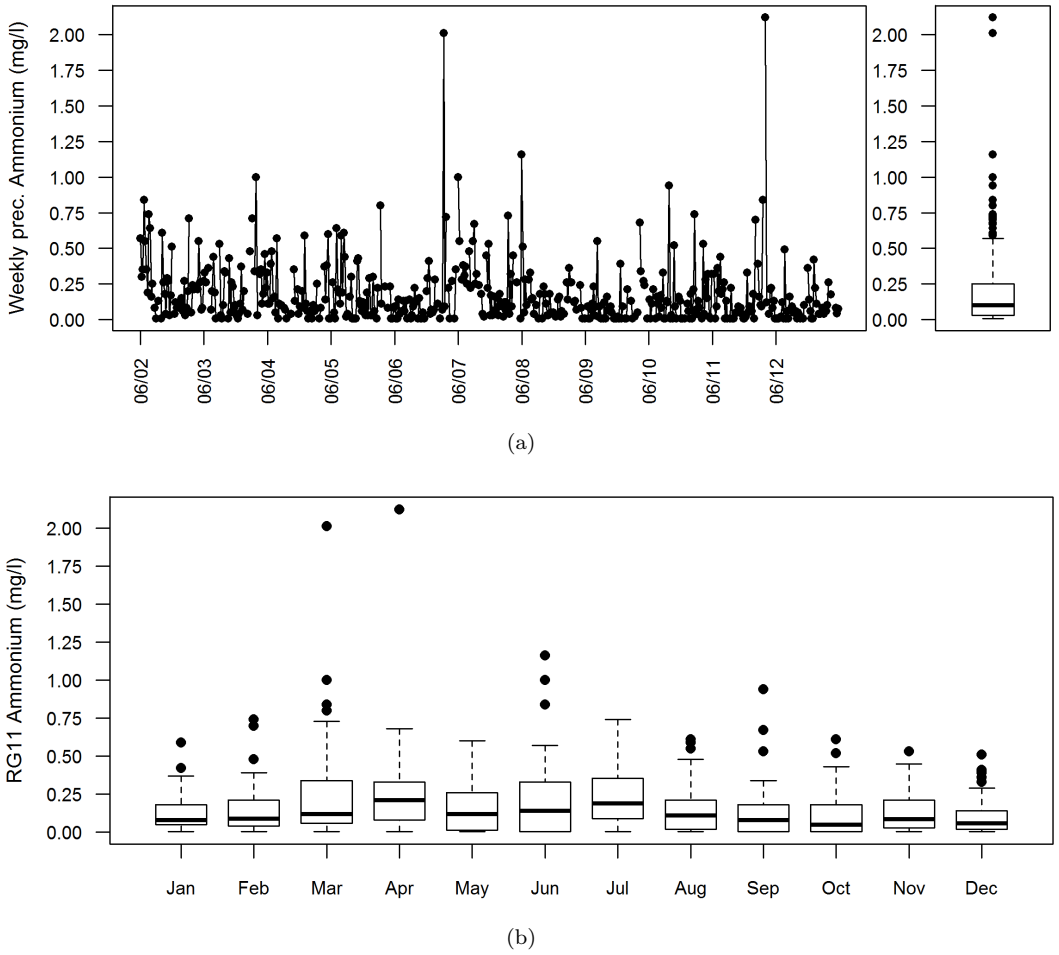


FIGURE 2.18: (a) Time series and box plot of weekly Ammonium (NH4) concentration (mg/l) (b) Box plot of Ammonium concentration by month (mm); RG11 example

Ammonium is normally used in fertilizer, cleaning agents and food additives. It is not important to health directly, even it may present in drinking water. It can cause taste and odour problems at high concentration (¿ 3 mg/l). The average of Ammonium concentrations is 0.178 mg/l (SD=0.216). Figure 2.18 shows the right skew of weekly Ammonium concentration.

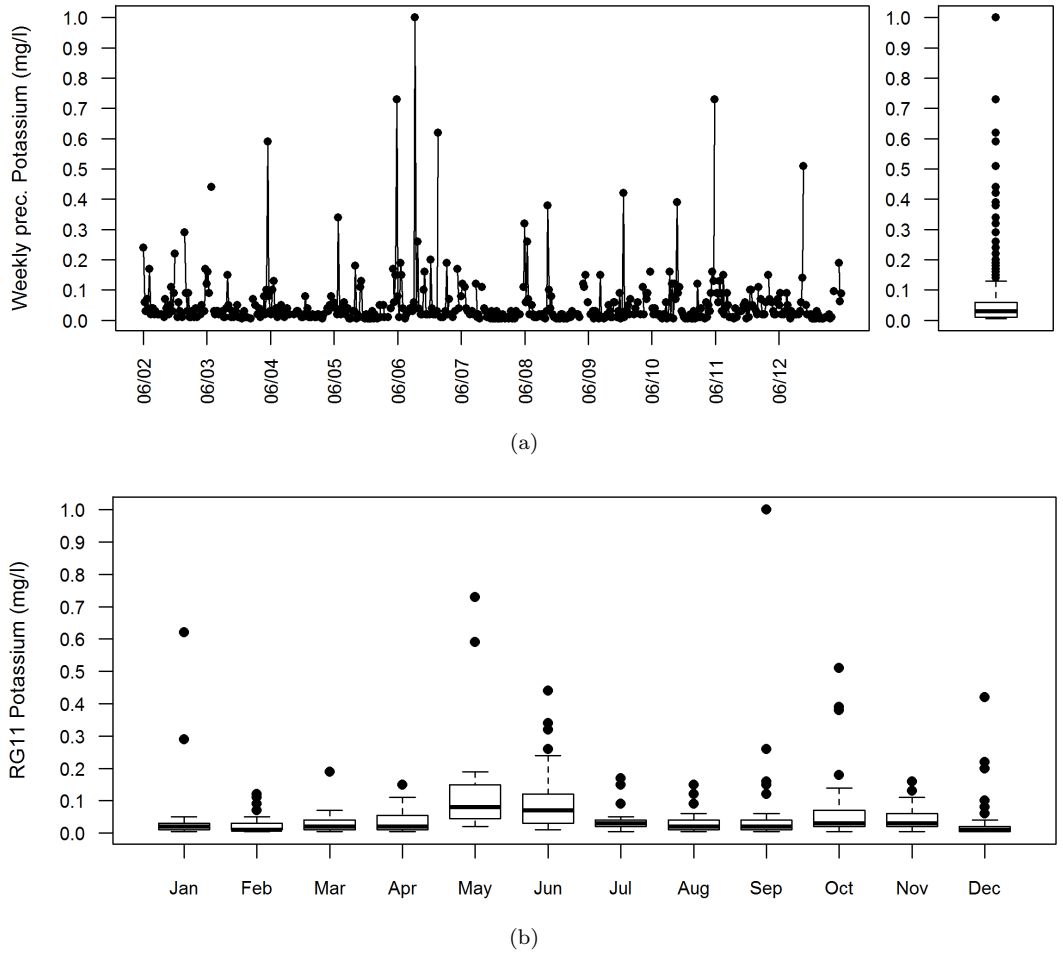


FIGURE 2.19: (a) Time series and box plot of weekly Potassium (K) concentration (mg/l) (b) Box plot of Potassium concentration by month (mm); RG11 example

Potassium is important to plant growth. It is often filled into soil as potassium fertilizer in Agriculture. It is easy widespread but less harm. It is usually formed as a compound with other components such as cyanide in potassium cyanide. Figure 2.19 shows that Potassium concentration is precipitation samples are very low at 0.058 mg/l (SD=0.091) on average with highly skew to the right.

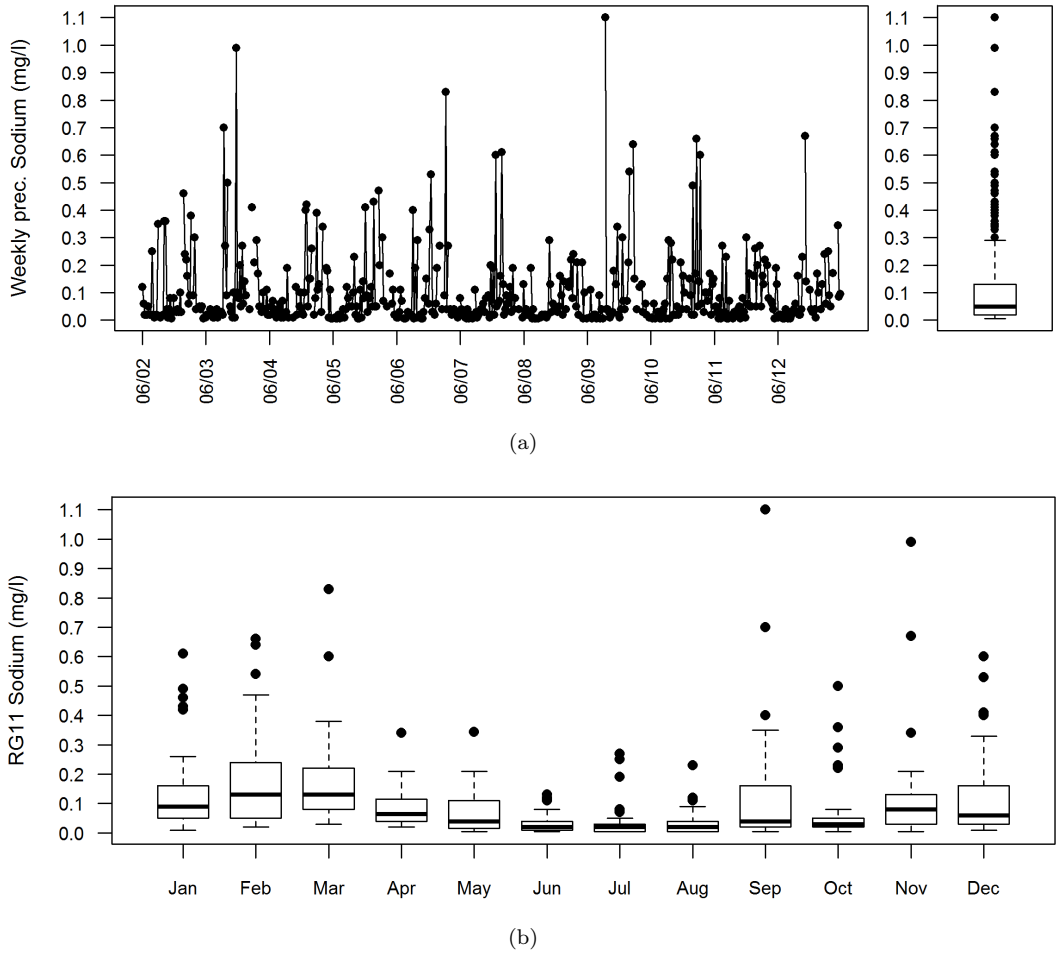


FIGURE 2.20: (a) Time series and box plot of weekly Sodium (Na) concentration (mg/l) (b) Box plot of Sodium concentration by month (mm); RG11 example

Sodium is a dietary mineral. It is an important for nerve and muscle system to function. It is normally used in daily life in form of sodium chloride or kitchen salt. Naturally, sodium is washed out from rocks and soil before end up in water sources such as sea. The Average of sodium level in precipitation samples is 0.117 mg/l (SD=0.15). It is suggest to take sodium chloride about 300 mg daily to maintain sodium level for human body. Figure 2.20 shows the moderate right skew of weekly sodium concentration. The monthly box plot suggests the lower sodium level around mid year.

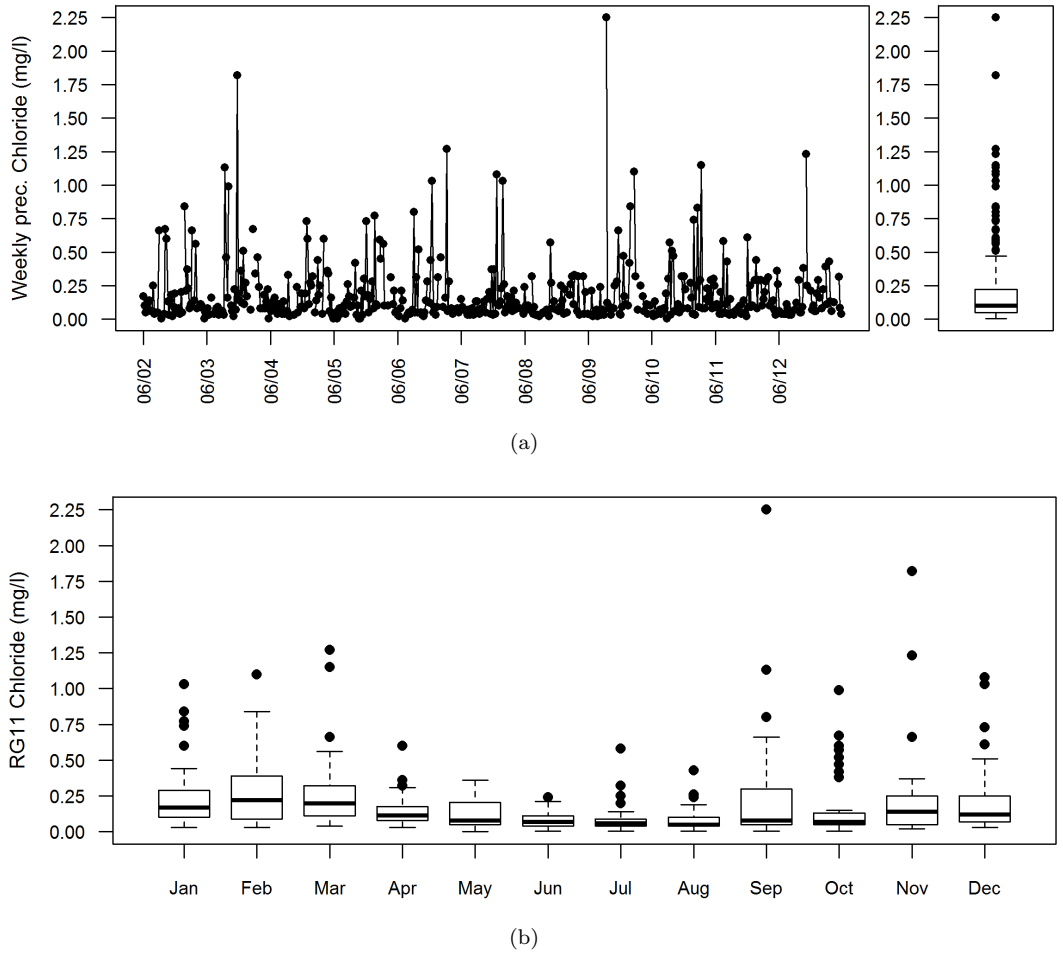


FIGURE 2.21: (a) Time series and box plot of weekly Chloride (Cl) concentration (mg/l) (b) Box plot of Chloride concentration by month (mm); RG11 example

Chloride is commonly found in tap water. It is not harm to human health at low levels. The drinking water standard suggests Chloride level not to exceed 250 mg/l, otherwise it will give taste and unpleasant odour problems. The average of Chloride concentration in precipitation samples is 0.199 mg/l (SD=0.254). Figure 2.21 shows the moderate right skew of weekly sodium level. The time series and monthly box plot suggests similar pattern with sodium concentration. This can be occur as we normally use sodium chloride compound (kitchen salt) in everyday.

The plots of precipitation and chemical solutes of Rain gauge 22 (RG22) data are presented in Figure 2.22-2.29. Those of Rain gauge 23 (RG23) are showed in Figure 2.30-2.37. It can be seen that the plots of RG22 and RG23 provide similar patterns with RG11.

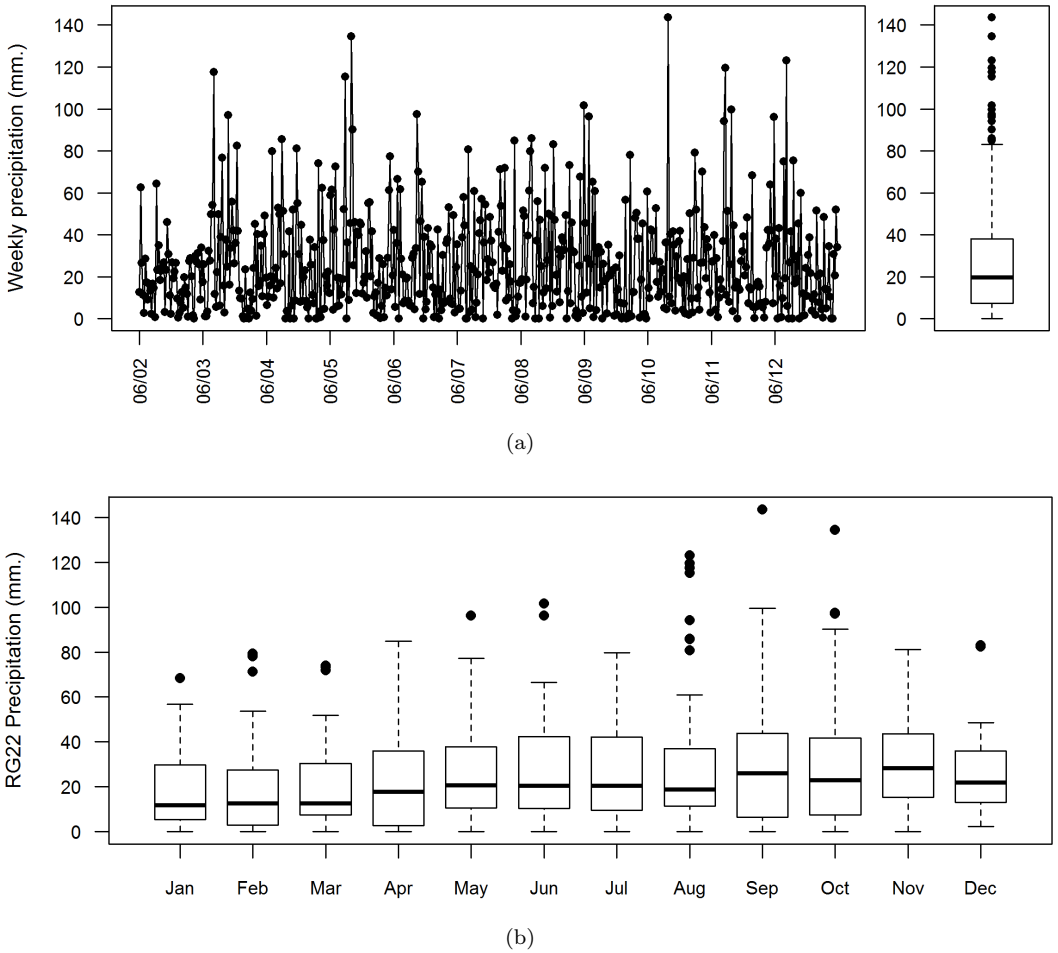


FIGURE 2.22: (a) Time series and box plot of weekly precipitation (mm) (b) Box plot of precipitation by month (mm); RG22 example

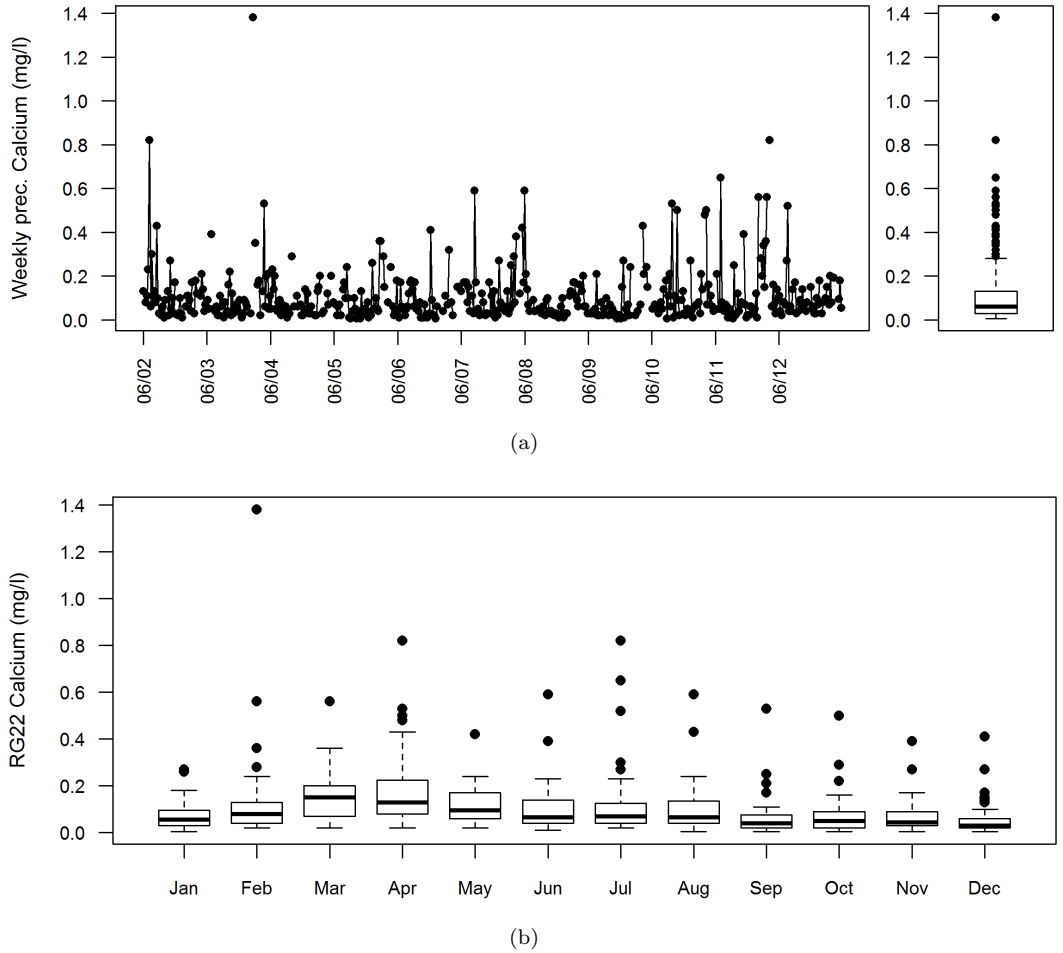
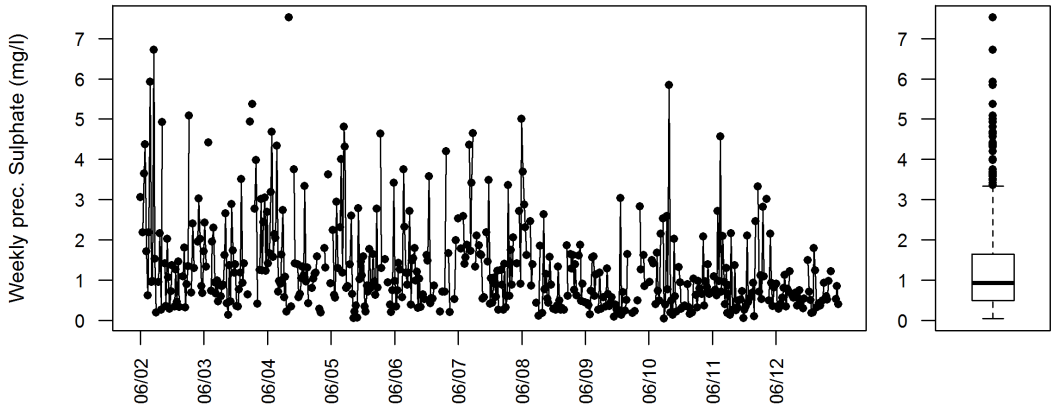
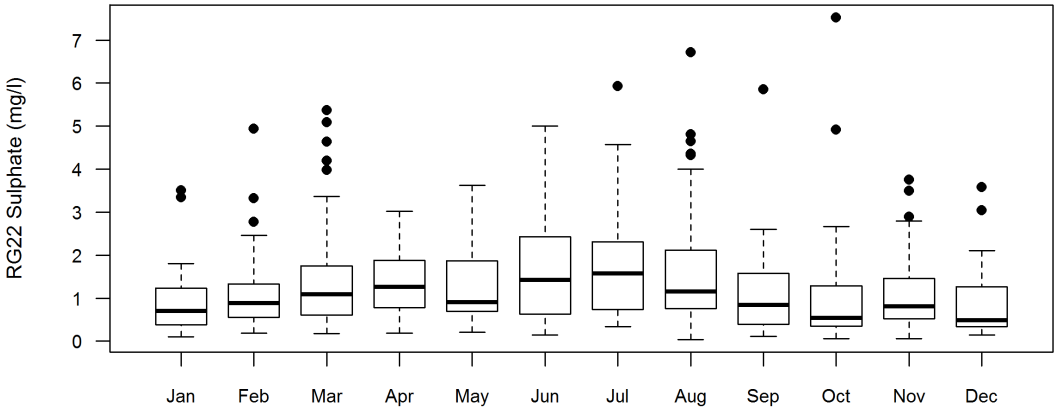


FIGURE 2.23: (a) Time series and box plot of weekly calcium (Ca) concentration (mg/l) (b) Box plot of calcium concentration by month (mg/l); RG22 example

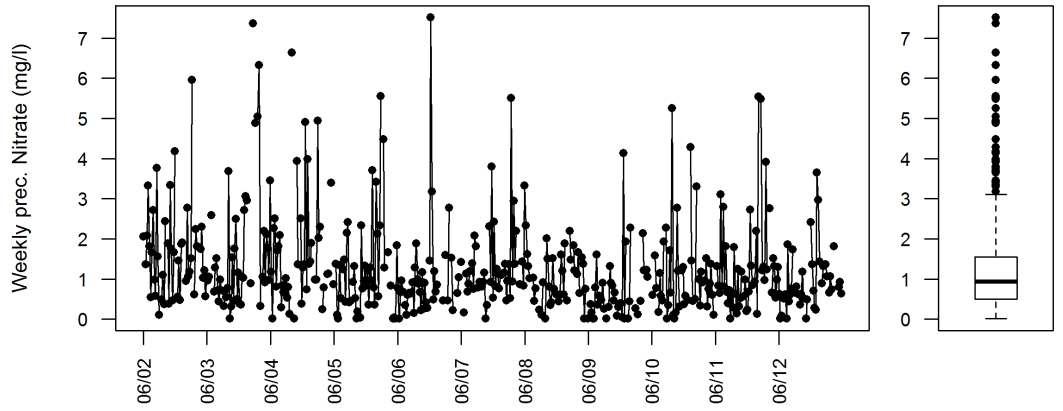


(a)

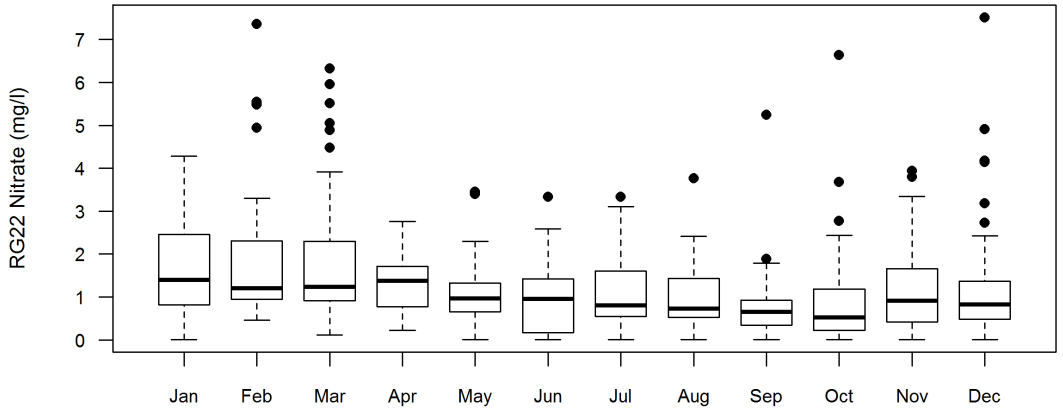


(b)

FIGURE 2.24: (a) Time series and box plot of weekly sulphate (SO4) concentration (mg/l) (b) Box plot of sulphate concentration by month (mm); RG22 example



(a)



(b)

FIGURE 2.25: (a) Time series and box plot of weekly nitrate (NO3) concentration (mg/l) (b) Box plot of nitrate concentration by month (mm); RG22 example

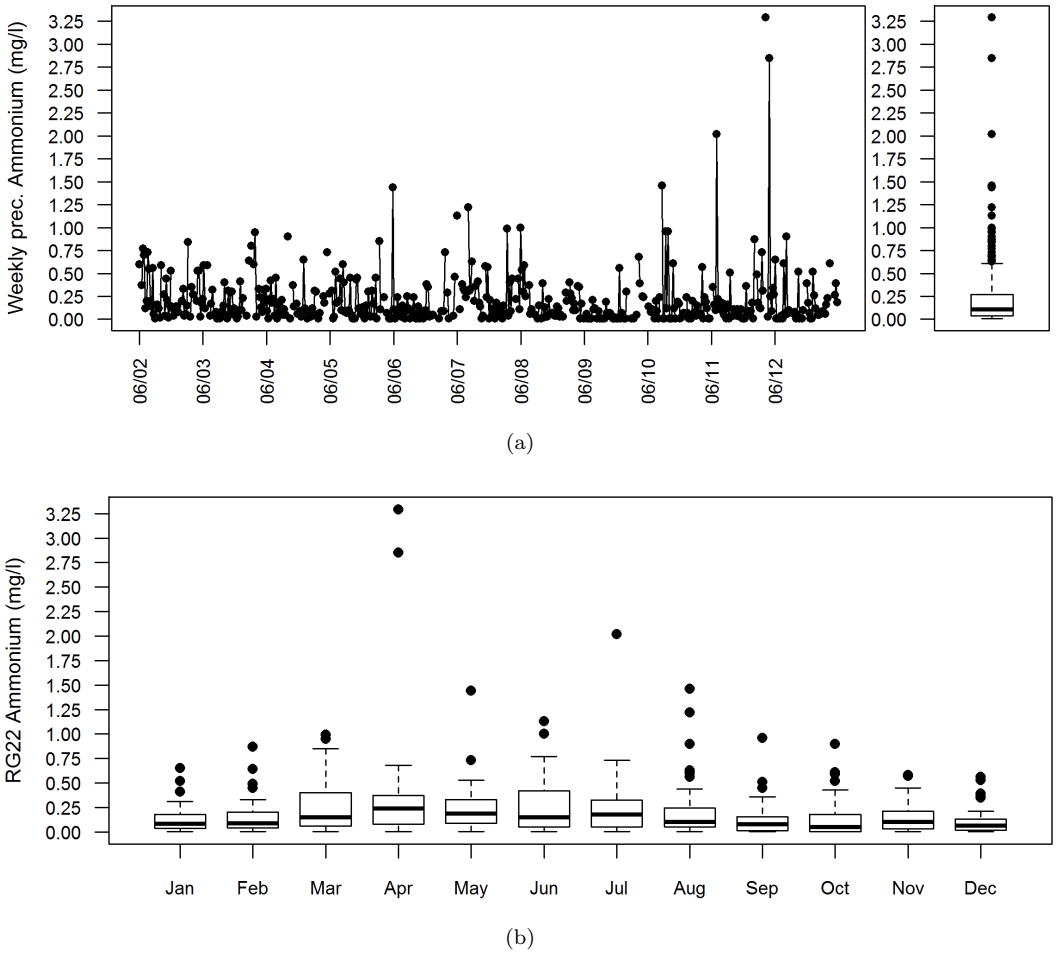
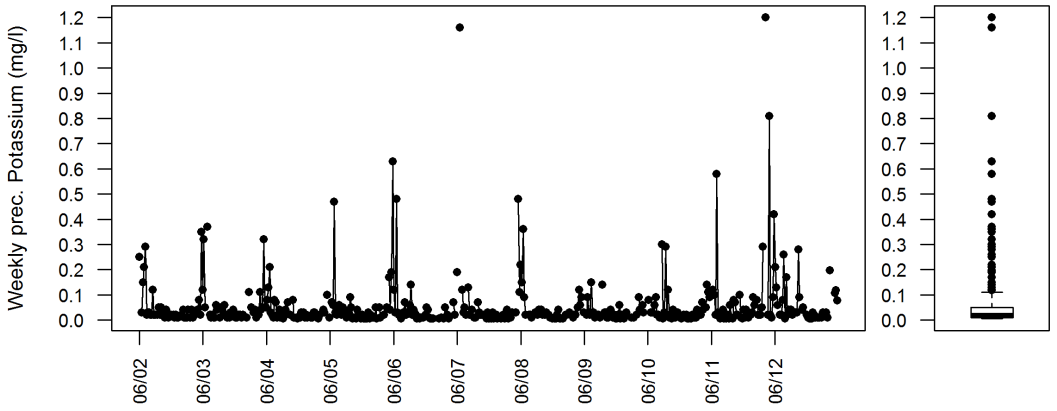
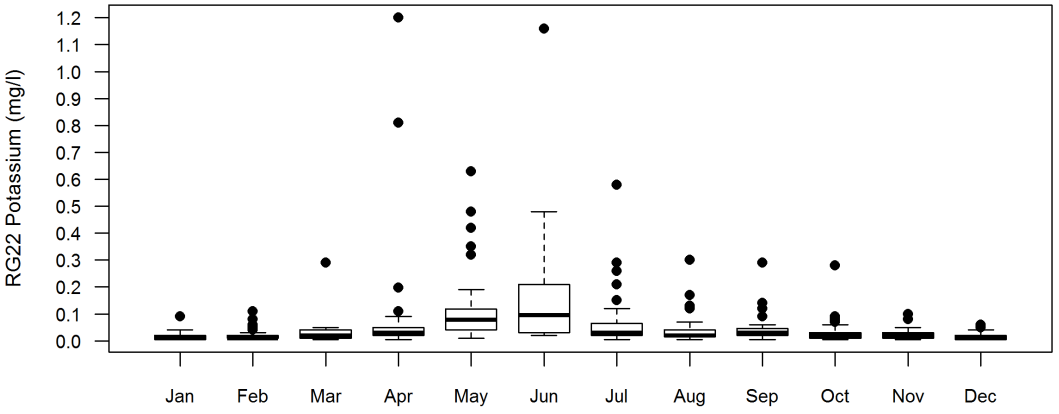


FIGURE 2.26: (a) Time series and box plot of weekly Ammonium (NH4) concentration (mg/l) (b) Box plot of Ammonium concentration by month (mm); RG22 example

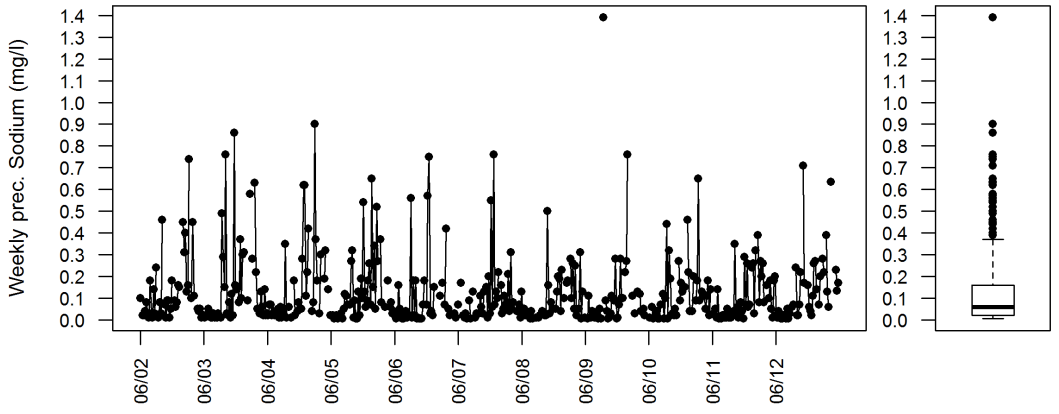


(a)

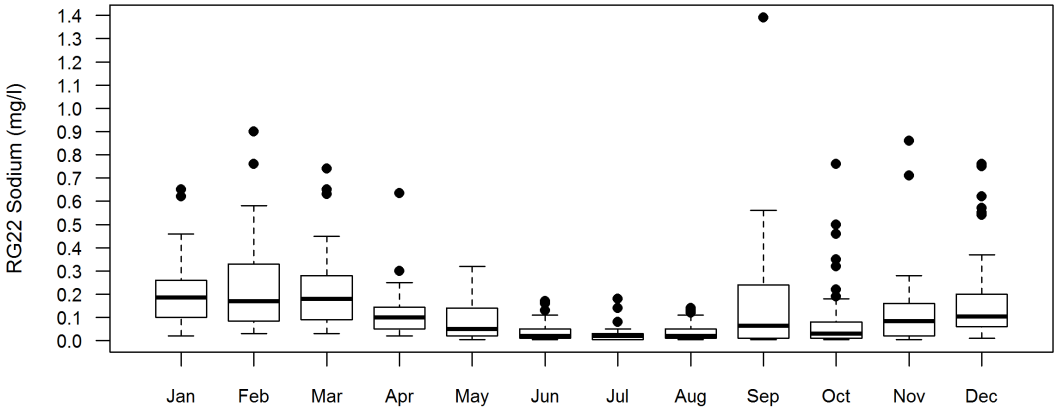


(b)

FIGURE 2.27: (a) Time series and box plot of weekly Potassium (K) concentration (mg/l) (b) Box plot of Potassium concentration by month (mm); RG22 example

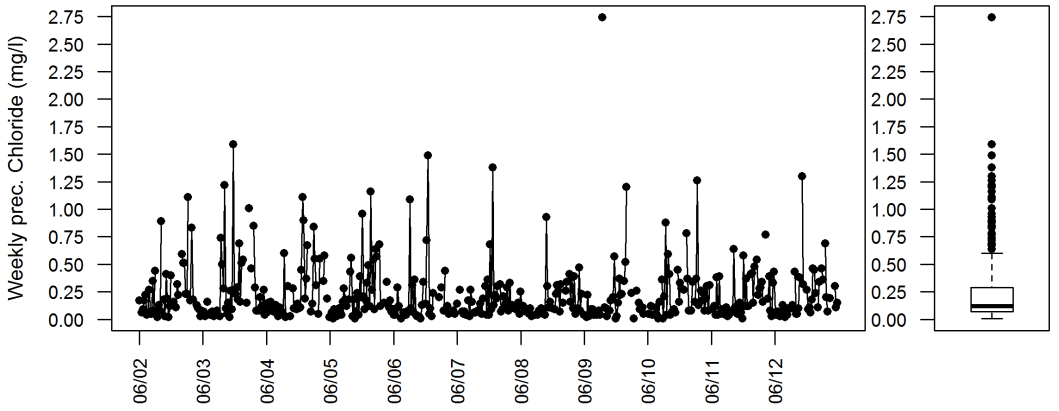


(a)

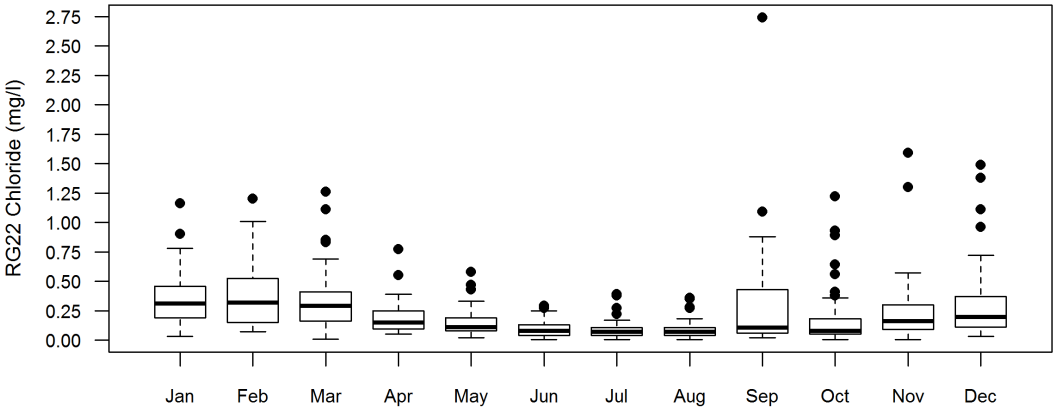


(b)

FIGURE 2.28: (a) Time series and box plot of weekly Sodium (Na) concentration (mg/l) (b) Box plot of Sodium concentration by month (mm); RG22 example



(a)



(b)

FIGURE 2.29: (a) Time series and box plot of weekly Chloride (Cl) concentration (mg/l) (b) Box plot of Chloride concentration by month (mm); RG22 example

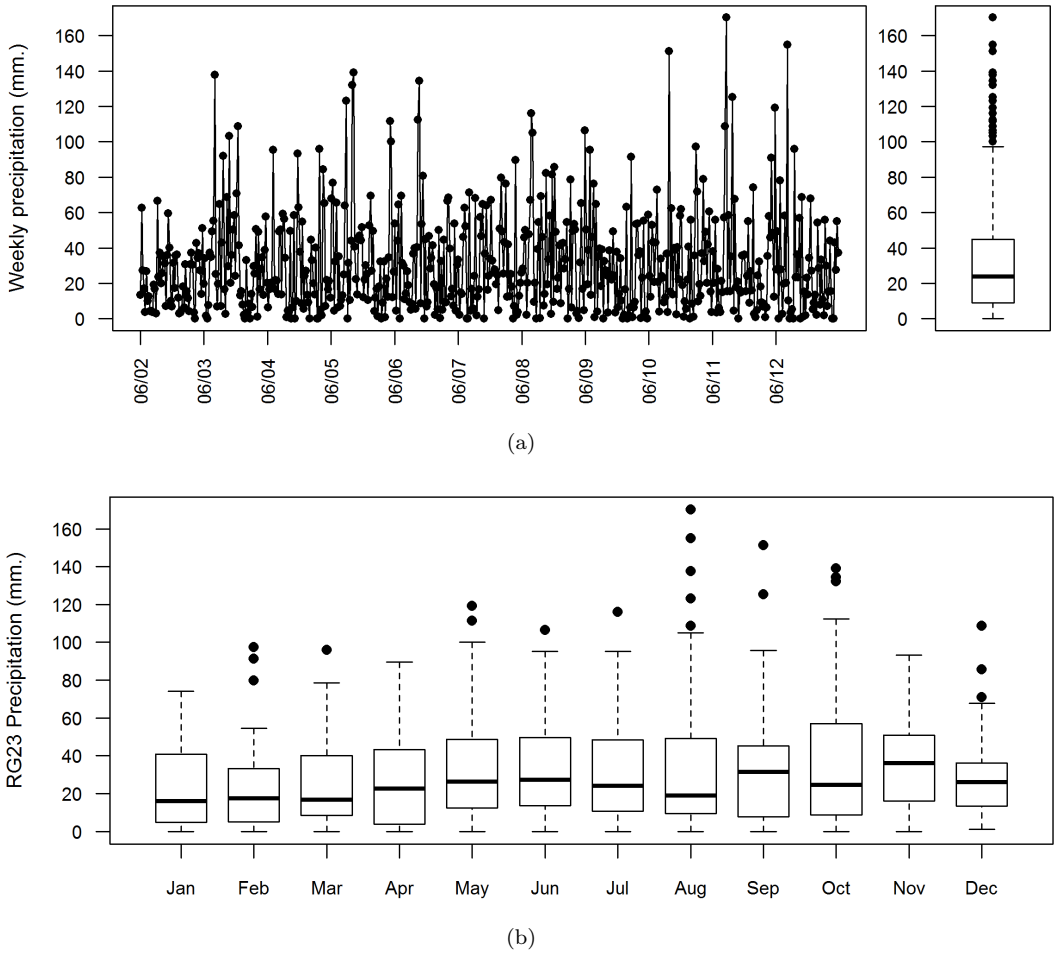


FIGURE 2.30: (a) Time series and box plot of weekly precipitation (mm) (b) Box plot of precipitation by month (mm); RG23 example

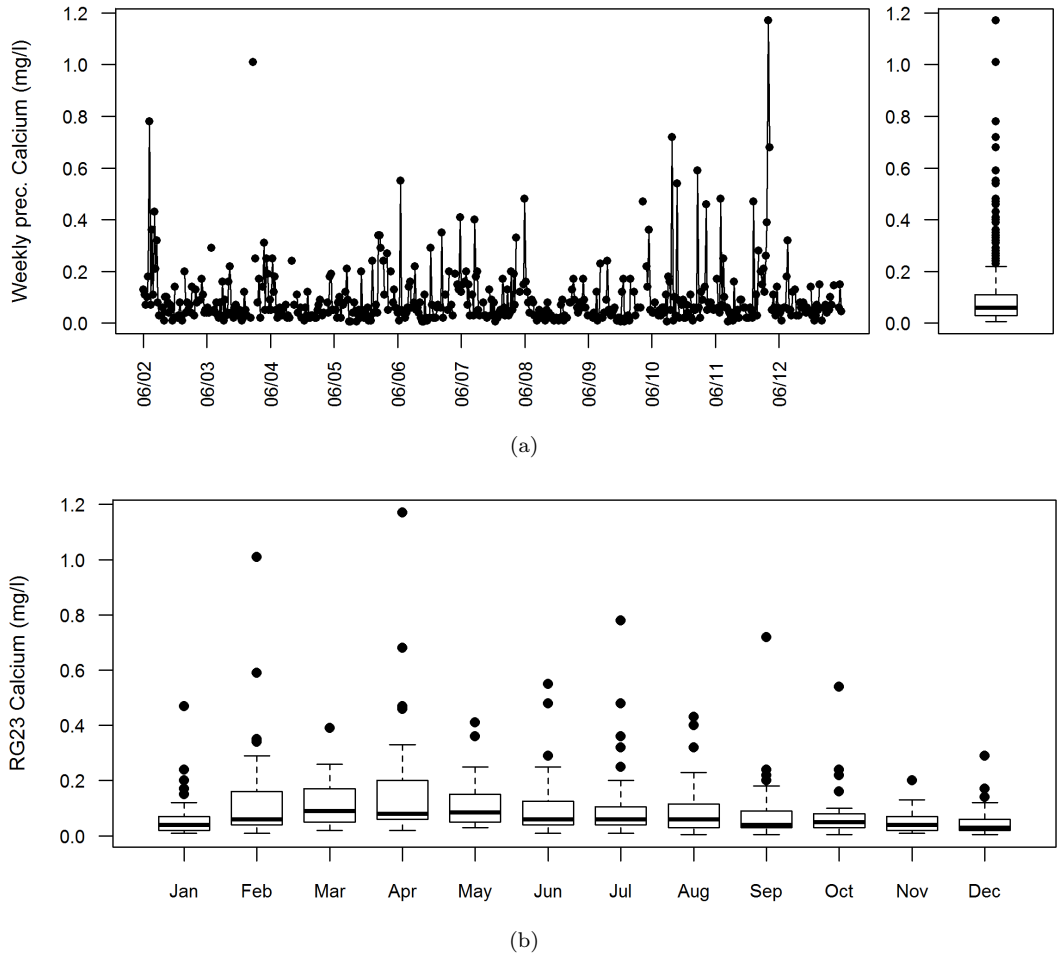
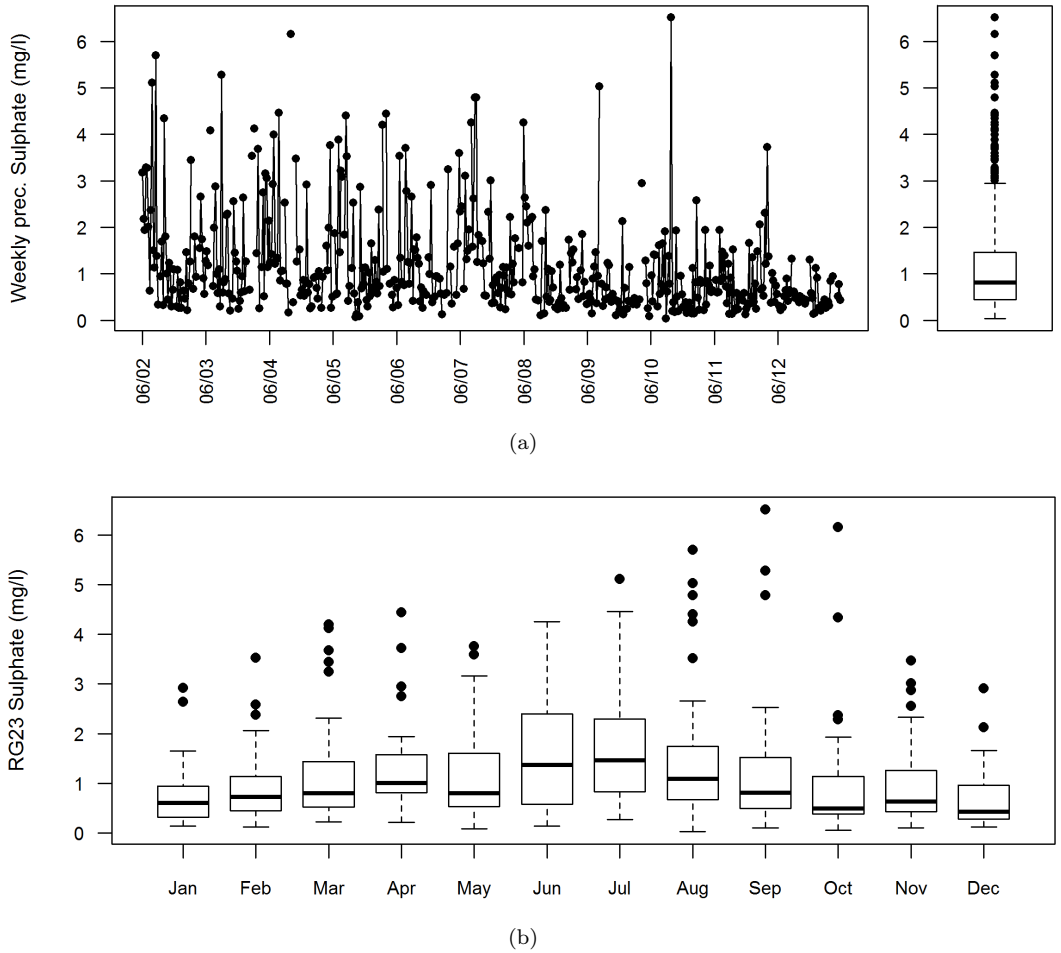


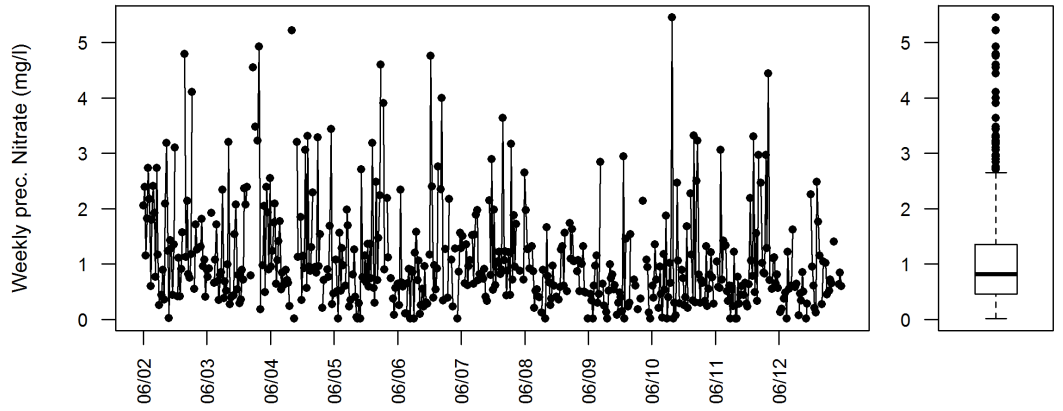
FIGURE 2.31: (a) Time series and box plot of weekly calcium (Ca) concentration (mg/l) (b) Box plot of calcium concentration by month (mg/l); RG23 example



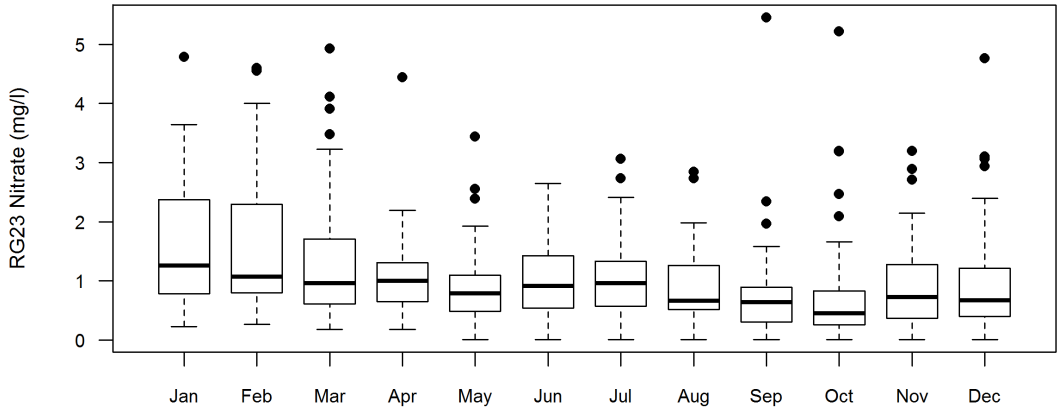
(a)

(b)

FIGURE 2.32: (a) Time series and box plot of weekly sulphate (SO4) concentration (mg/l) (b) Box plot of sulphate concentration by month (mm); RG23 example

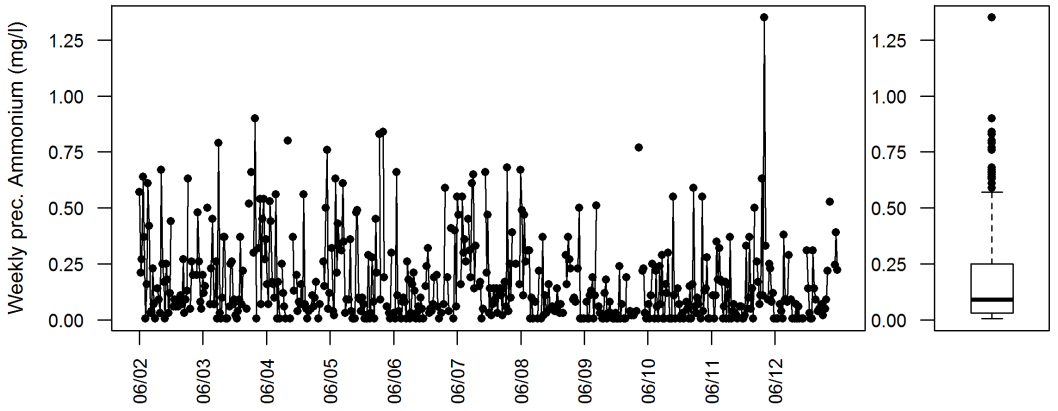


(a)

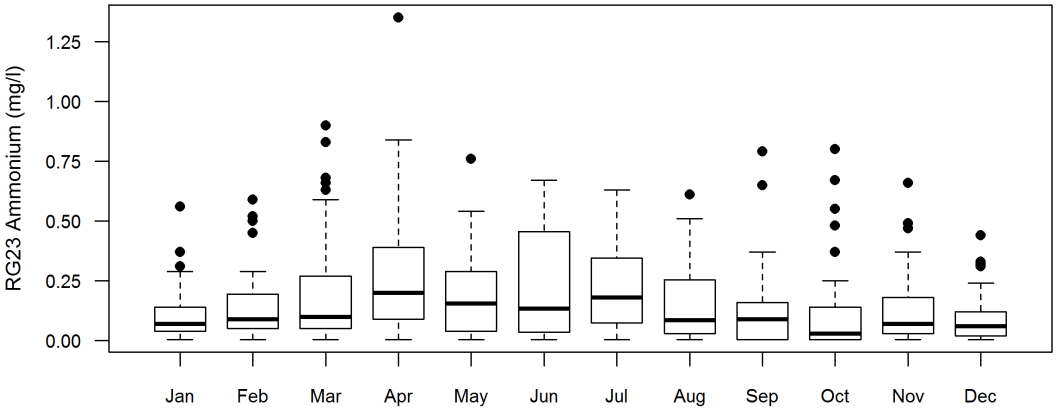


(b)

FIGURE 2.33: (a) Time series and box plot of weekly nitrate (NO3) concentration (mg/l) (b) Box plot of nitrate concentration by month (mm); RG23 example

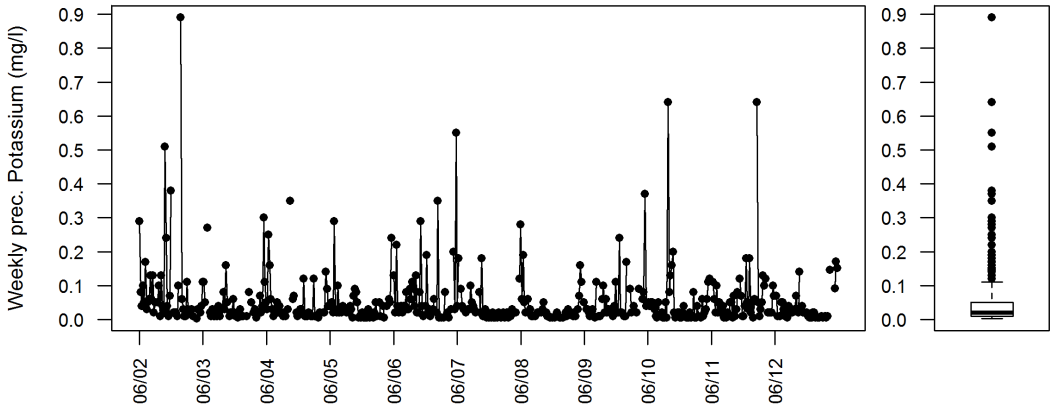


(a)

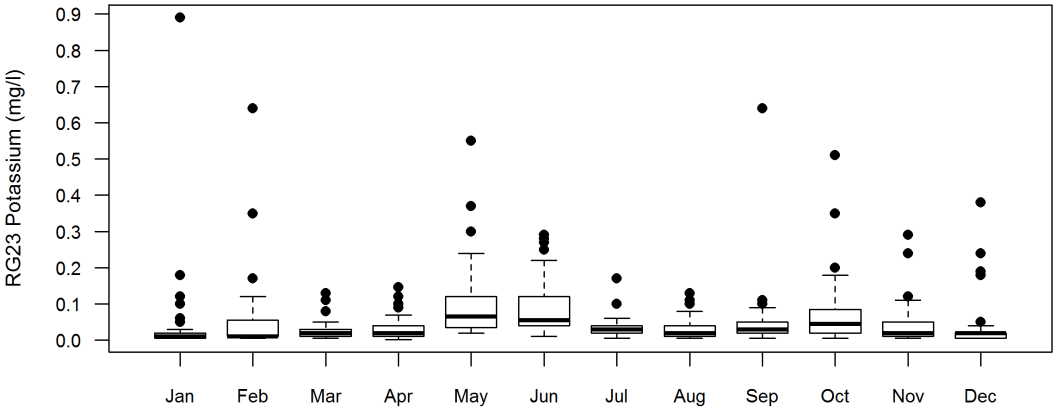


(b)

FIGURE 2.34: (a) Time series and box plot of weekly Ammonium (NH4) concentration (mg/l) (b) Box plot of Ammonium concentration by month (mm); RG23 example

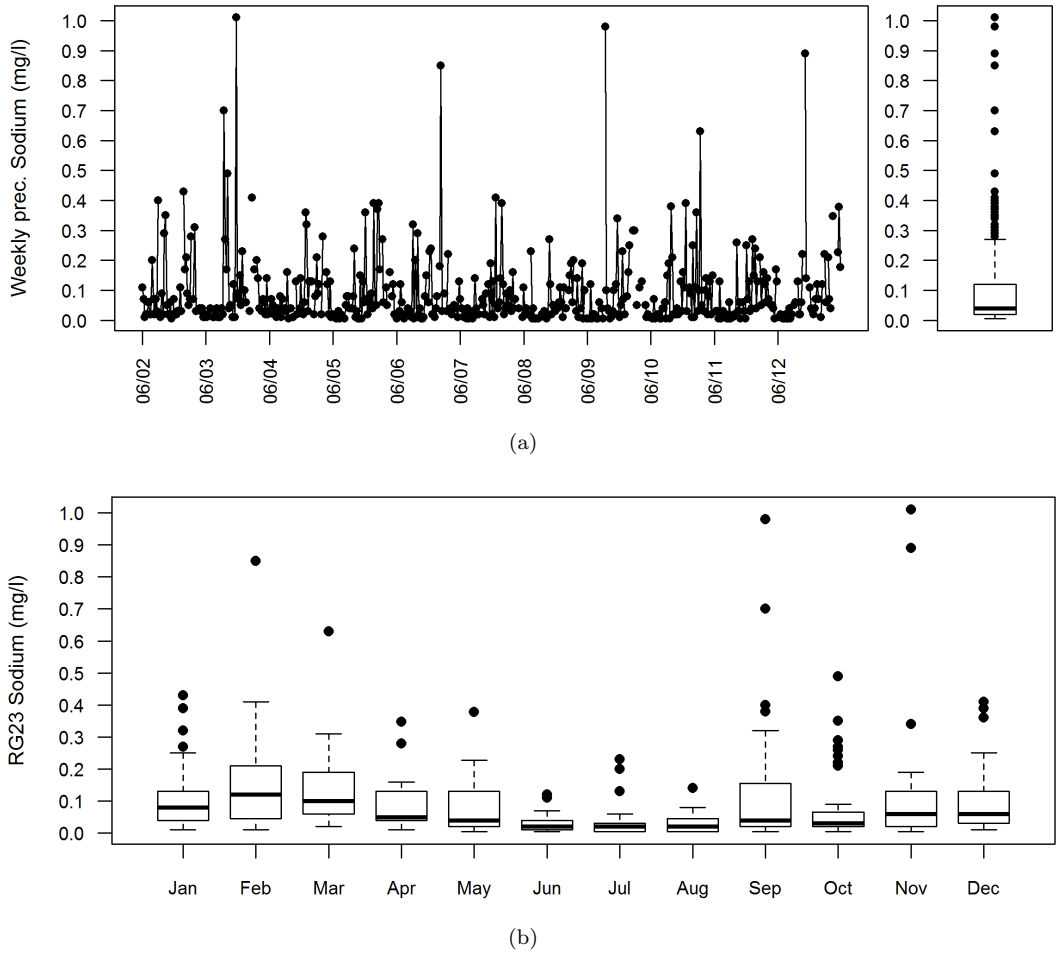


(a)



(b)

FIGURE 2.35: (a) Time series and box plot of weekly Potassium (K) concentration (mg/l) (b) Box plot of Potassium concentration by month (mm); RG23 example



(a)

(b)

FIGURE 2.36: (a) Time series and box plot of weekly Sodium (Na) concentration (mg/l) (b) Box plot of Sodium concentration by month (mm); RG23 example

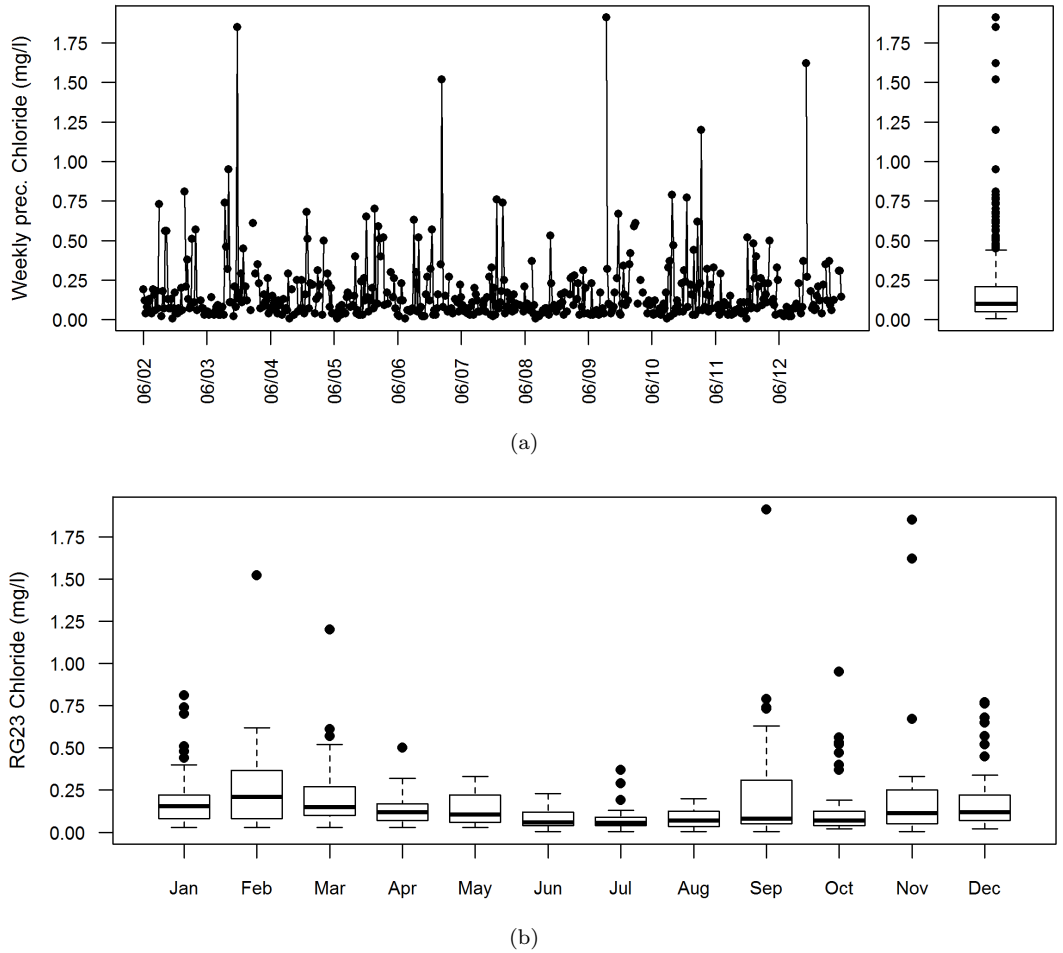


FIGURE 2.37: (a) Time series and box plot of weekly Chloride (Cl) concentration (mg/l) (b) Box plot of Chloride concentration by month (mm); RG23 example

2.2.4 Data Preparation

The deviation of mean and median in table 2.2 and box plots indicate positive skewness. The need of reducing the large variation and the effect of outliers, it is considerable to apply a data transformation, i.e. logarithmic transformation to stabilise its variance. It is also convenient for solutes as it is nonnegative and their distributions are skewed to the right and helps in consider the relation among precipitation and chemical solutes.

Figure 2.38 visualizes the relationship of data from RG11 on natural logarithm scale. Interestingly, there is a weak correlation between precipitation and those chemicals, but moderate relationship among chemicals. Figure 2.39 and Figure 2.40 present the correlation among the RG22 and RG23 data respectively.

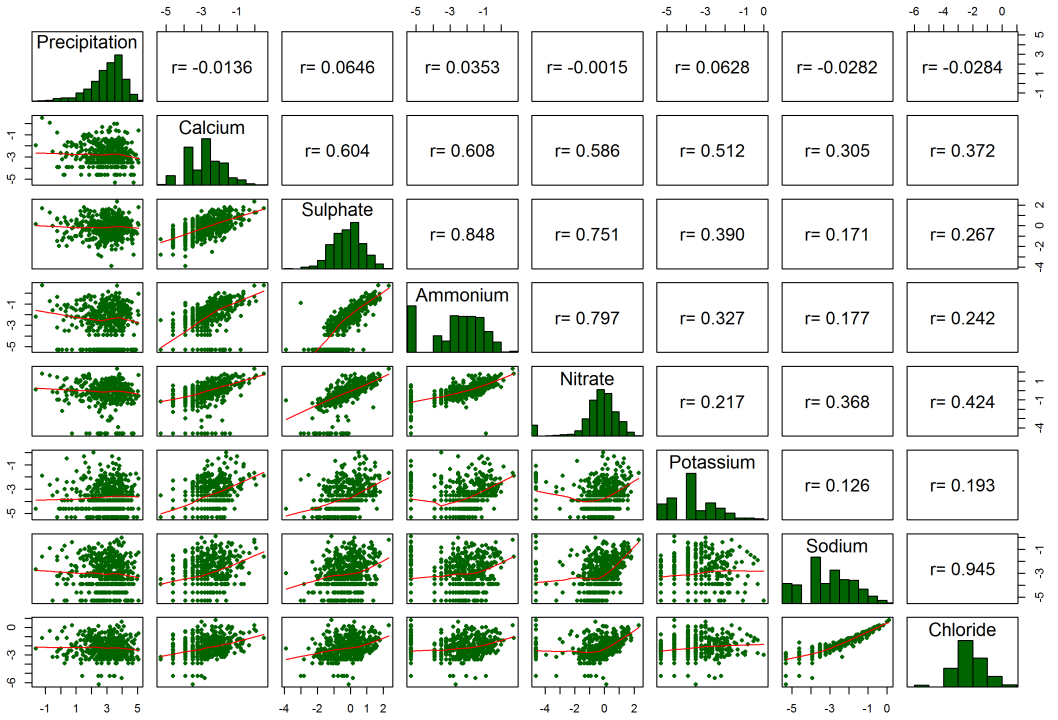


FIGURE 2.38: Pairwise plots with correlation among precipitation and chemical concentrations on natural logarithmic scale : RG11

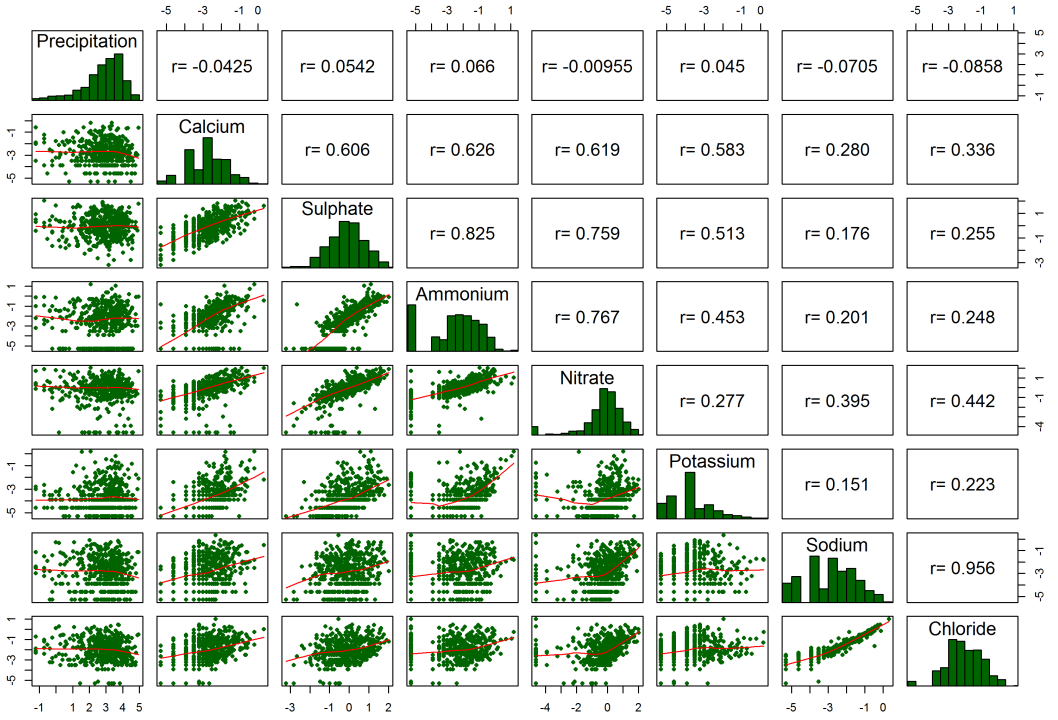


FIGURE 2.39: Pairwise plots with correlation among precipitation and chemical concentrations on natural logarithmic scale : RG22

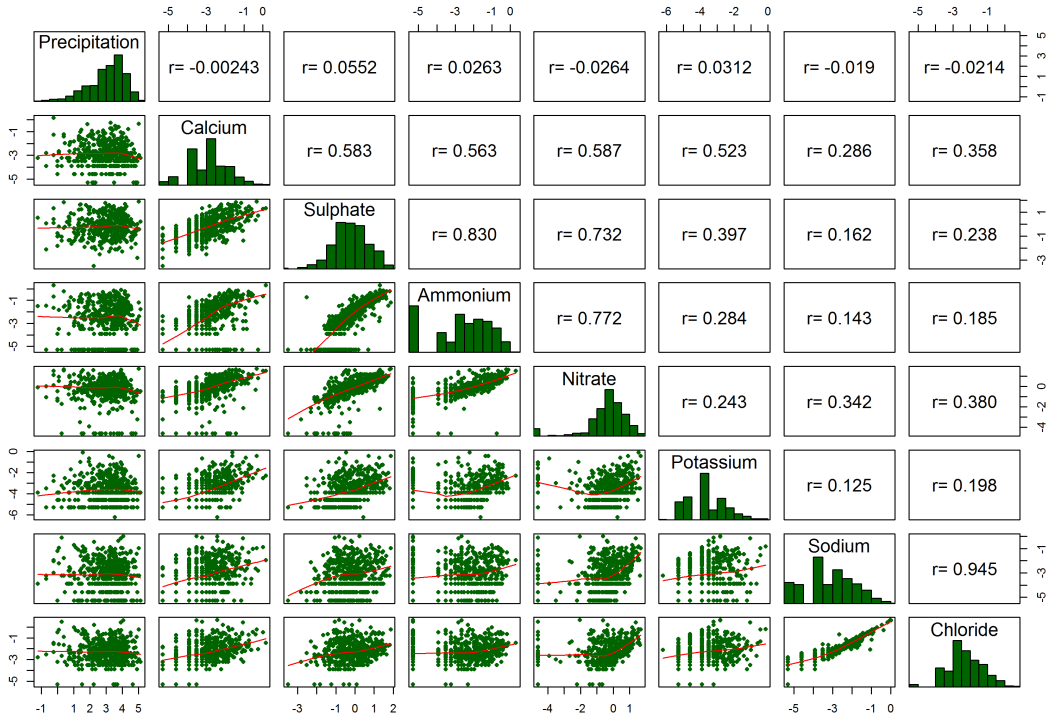


FIGURE 2.40: Pairwise plots with correlation among precipitation and chemical concentrations on natural logarithmic scale : RG23

Consequently, we consider to model chemical solutes only as time series data. Also, we are studying combined models of chemical concentrations for better model explanation based on the relationship among chemicals.

The precipitation chemistry data contains some missing values defined as NA or data not available. It needs to be treated before modelling. Due to the limitation of studying statistical methods, it is not allow partly missing of any records or weeks. Thus, there are two choices of missing values treatment: 1) replace the whole record with NA or 2) impute the NA with moving average of data before and after that NA record. In order to make a decision, all rain gauge data are combined together by week to check the missingness. If any records or weeks contain the whole NAs, it will remain the same as the first choice application. Otherwise, the incomplete or partial NA records, the missing value will be reputed with the moving average of measurements from the before and after that missing week as the second choice application. We adopts both choices to avoid unnecessary discarded data if only the first choice is applied. However, if only the second choice is used, a large amount imputed data may influence the result of data modelling. As a result, NA records reduced to 83 weeks out of 574 weeks or about 14.5% for each chemical solutes. Fortunately, Bayesian approach is able to handle missing data by treating missing data as random variables.

Figure 2.41-2.47 show upward pattern from left to right indicating a positive relationship of solutes across rain gauges (also watersheds). Hence, it is possible to consider the

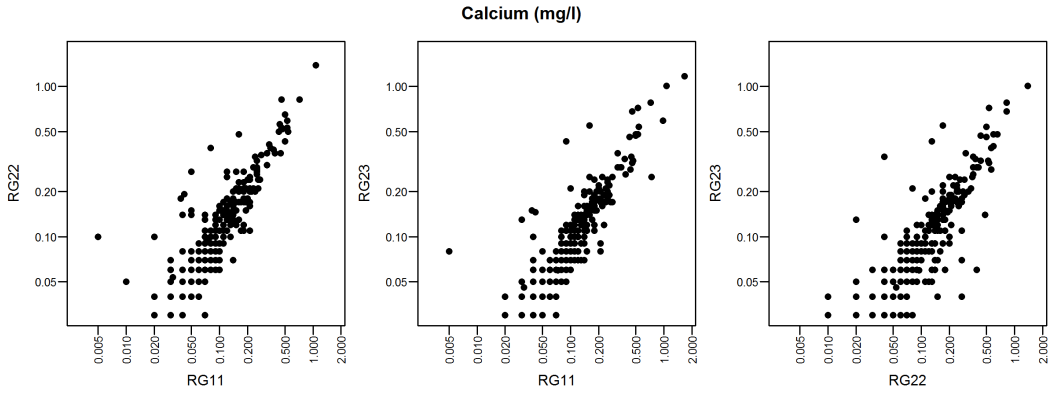


FIGURE 2.41: Scatter plots of Calcium between rain gauges on logarithmic scale

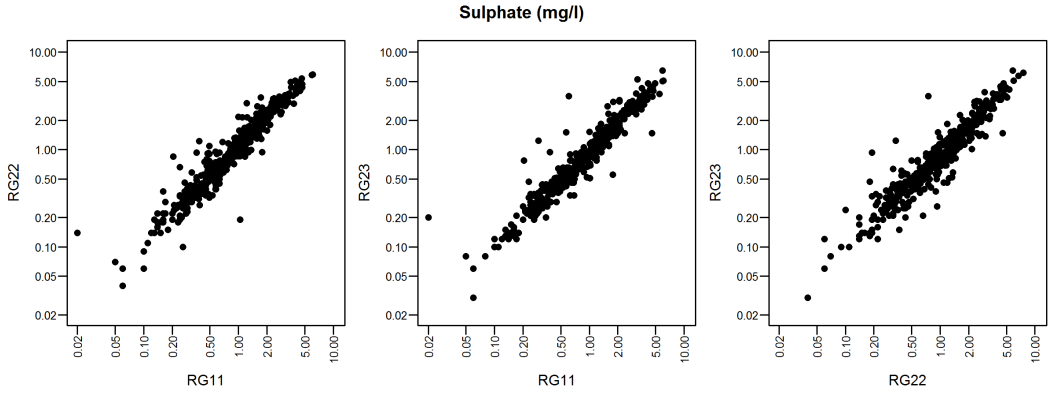


FIGURE 2.42: Scatter plots of Sulphate between rain gauges on logarithmic scale

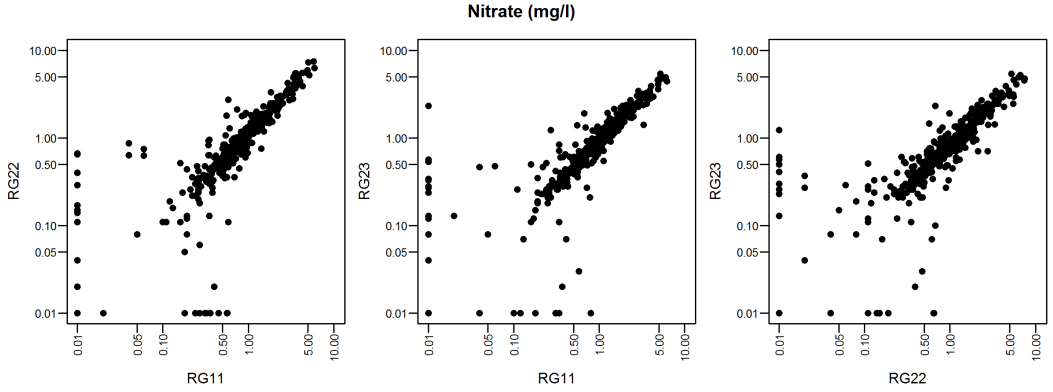


FIGURE 2.43: Scatter plots of Nitrate between rain gauges on logarithmic scale

relationship between chemical solutes as well as rain gauges to provide a combined model of chemical solutes or/and rain through their correlations.

In summary, the time series of chemical concentrations show seasonal pattern and dynamic effects. The data are treated on some missing values and transformed by logarithmic scale preparing for modelling. In order to cover uncertainties, we are interested in

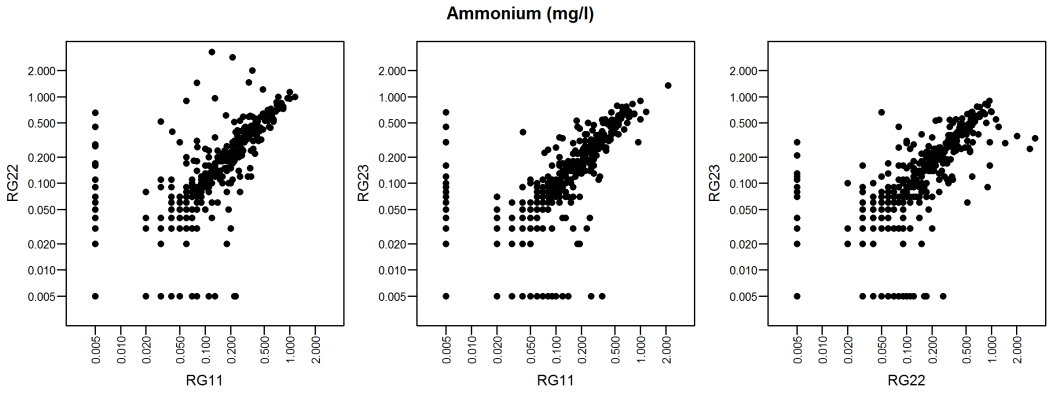


FIGURE 2.44: Scatter plots of Aluminium between rain gauges on logarithmic scale

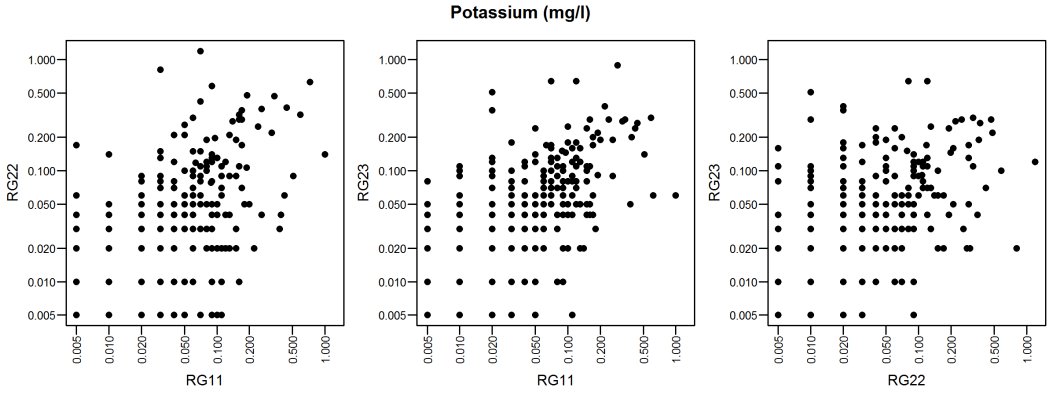


FIGURE 2.45: Scatter plots of Potassium between rain gauges on logarithmic scale

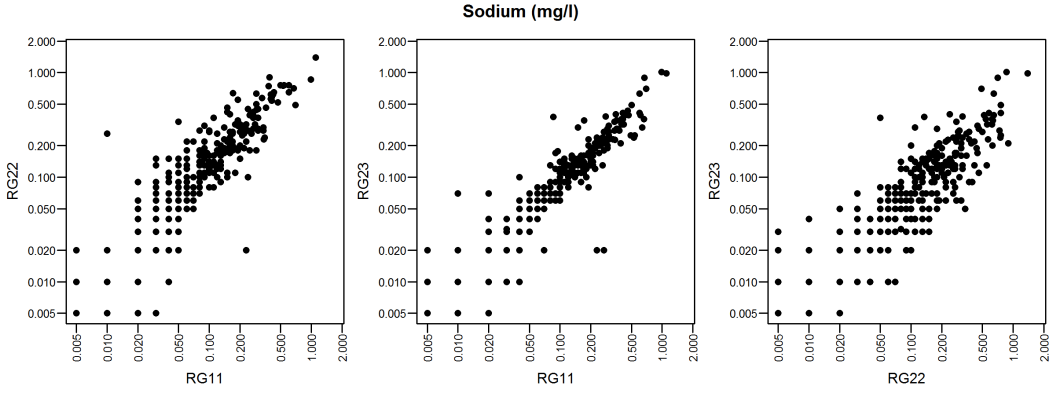


FIGURE 2.46: Scatter plots of Sodium between rain gauges on logarithmic scale

modelling the precipitation chemical concentrations based on Bayesian approach. Correlation among chemicals also suggests a combining model of those chemicals. Noted that it can be seen stacks of point in scatter plots of solutes. This is due to measurement strategy and precision of measurement devices if detecting very low level of solutes. In the past, it is often rounded while recording.

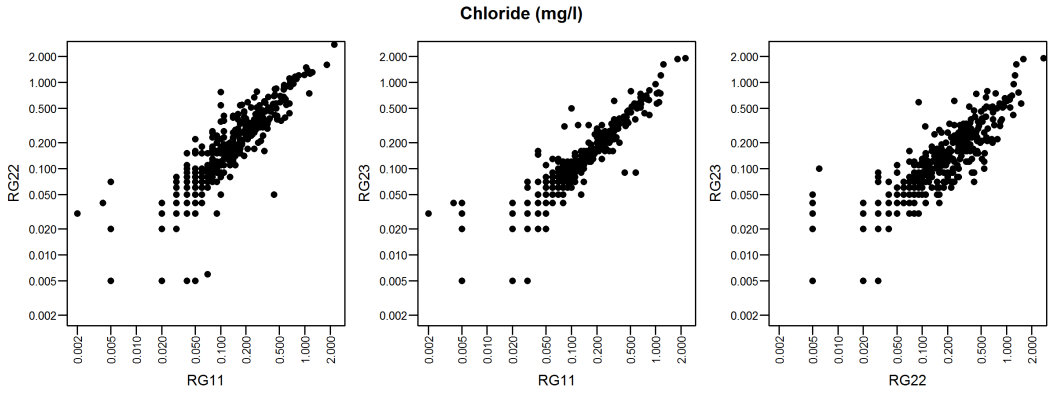


FIGURE 2.47: Scatter plots of Chloride between rain gauges on logarithmic scale

Chapter 3

Bayesian Multivariate Modelling

3.1 introduction

The Bayesian approach is used to construct the multivariate model. This chapter we provide the basic of Bayesian modelling concept. It allows to cope with the uncertainty in modelling and parameter estimation describing in term of probability manner. Especially with ecological observation that presents specific characteristics such as positive skewness, non-normal distribution, extreme outliers, and missing value due to natural uncertainty. For more information about Bayesian modelling and multivariate Bayesian statistics, see Gelman et al. (2014), Rowe (2002), Press (2005).

3.2 Bayesian modelling and computation

3.2.1 Bayesian framework

In the Bayesian inference, we may have belief or prior knowledge about parameter. This prior belief can be changed or updated with new data to the posterior distribution. Let $p(Y \mid \theta)$ be the likelihood function of parameters $\theta = (\theta_1, ..., \theta_p)'$ based on the observed data $Y = (Y_1, ..., Y_n)'$. $\pi(\theta)$ denotes a prior distribution of the parameter θ. The conditional density of the parameters given the data or posterior distribution can be obtained as

$$\pi(\theta \mid Y) = \frac{p(Y \mid \theta)\pi(\theta)}{\int_n p(Y \mid \theta)\pi(\theta)d\theta} \tag{3.1}$$

The integral term, marginal likelihood of the data Y does not contain any parameter θ, then it can be considered as a normalizing constant. Hence the posterior distribution can be written as proportional posterior,

$$\pi(\theta \mid Y) \propto p(Y \mid \theta)\pi(\theta)d\theta. \tag{3.2}$$

Bayesian Approach can be applied with some steps (1) find the likelihood of the parameters given data, (2) define prior distribution for unknown parameters, (3) derive or calculate the posterior distribution and (4) inference about the parameter using the updated information through posterior distribution.

3.2.2 Prior choices

There are several ways to choose prior distribution. Informative or noninformative prior will be used based on previous knowledge or belief. Some prior distributions lead to the posterior distribution with the same family as the prior, called conjugate priors. The most of Gaussian prior distribution is this case. However, the prior can be dominated with large data.

3.3 Methods and Criteria for model selection

For assessing and selecting models, we use a traditional coefficient of determination and adopt Bayesian predictive approaches which assess models and compare candidate models based on their predictive performance. If the model performs the poor predictions, it is lack of predictive ability about future or unseen observations. We use three model choice criteria: the analogus Bayesian coefficient of determination (R_B^2), Cross-validation (CV) and Predictive model choice criteria (PMCC).

The analogus Bayesian coefficient of determination The traditional coefficient of determination, known as adjusted R^2, is considering in Bayesian viewpoint as R_B^2. It is defined as

$$R_B^2 = 1 - \frac{\sigma^2}{S_Y^2} \tag{3.3}$$

where σ^2 is the model variance (described by explanatory variables) and S_Y^2 is the sample variance of Y (macronutrient or solute concentration). If the model is fit to the data, the model (explanatory variables) capture the most variation of Y. The variance of model will be lower than the variance of Y. Hence, a model with the larger R_B^2 is preferred as a better fit model.

Cross Validation (CV) Cross validation is a method to evaluate models on small data using resampling technique. The k-fold cross validation (k-fold CV) is a well-known technique that splitting data into k groups or folds then keeping a data fold as validation data to be predicted and remaining as training datasets to modelling. Thus, it provides k independent datasets to assess the modelling performance. The overall prediction

errors will be expressed for comparison by the root mean squared error (RMSE), the mean absolute error (MAE), and the continuous ranked probability score (CRPS).

1. The root mean squared error (RMSE) is the standard deviation of the residuals which measure the different between data points and its prediction. It is defined by:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_{st} - \hat{y}_{st})^2}$$

where y_{st} is the actual value, $\hat{y}_{st}$ is the prediction value at time t, and T is the total number of observations over time.

2. The mean absolute error (MAE) is a measure of average magnitude of the residuals without taking direction into account. It is defined by:

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|y_{st} - \hat{y}_{st}|\,.$$

3. The continuous ranked probability score (CRPS) is a comparable score in terms of forecasting. It is defined by:

$$crps(F, y) = E_F\,|Y - y| - \frac{1}{2}\left|Y - Y'\right|$$

where Y and Y' are independent of a random variable with distribution function F and finite first moment. With m hold-out observations, the overall score is given by:

$$CRPS = \frac{1}{m}\sum_{i=1}^{m} crps(F_i, y_i)$$

Regarding MCMC sample $y_i^{(j)}, j = 1, ..., J$, then

$$\hat{crps}(F, y) = \frac{1}{J}\sum_{j=1}^{J}\left|y_i^{(j)} - y_i\right| - \frac{1}{2J^2}\sum_{j=1}^{J}\sum_{k=1}^{J}\left|y_i^{(j)} - y_i^{(k)}\right|$$

where $i = 1, ..., m$. Then the estimated overall CRPS is given as

$$\hat{CRPS} = \frac{1}{m}\sum_{i=1}^{m} \hat{crps}(F_i, y_i)$$

Predictive Model choice criteria (PMCC) This approach is assessing the quality of the model fit and prediction by comparing the observed data with estimates obtained by samples from the posterior predictive distribution. It is suit to compare models with

normally distributed error. (Laud et al. 1995, Gelfand & Ghosh 1998).

$$PMCC = \sum_{j=1}^{J} (E(y_{new,j}) - y_{obs,j})^2 + \sum_{j=1}^{J} Var(y_{new,j})$$
$$= \text{Goodness of fit} + \text{Penalty for model complexity}$$

where J is the total number of observations (dataset) combined into the model, $S \times T$ where S is the number of data sources and T is the number of observation in each data source. For example, a model of a macronutrient concentration is constructed by the data from two rivers ($S = 2$), each river provides 51 observations ($T = 51$), then $J = S \times T = 2 \times 51 = 102$ datasets are included into the same multivariate model. $E(y_{new,j})$ and $Var(y_{new,j})$ are the expectation and variance of prediction on observed y_{obs} over MCMC iterations with the j^{th} dataset. The more complex model, the more goodness of fit. However, to avoid the overfitting issue, the model should has a good balance between goodness of fit and penalty for model complexity.

Consequently, the preferred models is a model with the largest value of R_B^2 and the smallest values of RMSE, MAE, CRPS, and PMCC as it provides a good prediction behaviour.

3.4 Bayesian multivariate normal model

In ecological study, observations do not meet statistical assumptions such as the data is independent and normally distributed. Typically, the environmental data are correlated over time, across the monitoring stations and depend on geological area. The data might be collected separately but correlate each other. Data is also skewed. It may contains very low measurements or truncated data. Missing data is a common problem due to natural disturbance. Mostly, the data is collected in short fixed period. These limit relationship investigations among natural factors using statistical manner. Nowadays, new statistical methods are developed to solve such concerning problems. However, it is also more complicate to apply.

The beginning of our approach come from emerging of water pollution. Scientists need to estimate chemical substances in river which delivering to an estuary each year. Seasonality may appear but cannot capture the pattern due to no replicates. The excess amount of these chemical substances such as nitrate may harm aqua and marine ecology; can cause the polluted environment and degraded water quality. Water samples were collected weekly from two rivers for one year. Sample were analysed in laboratory for macronutrient concentrations. Some related data were collected at the same time such as river flow rate. This is a case of small data and the environment data experiencing

natural disturbances such as local storms which may affect the amount of macronutrients in river.

Focusing more on temporal effects or a kind of uncertainty and small data, we need to combine the correlated data from different sources (rivers) for more information and better describe the macronutrient concentrations. This leads to the development of our approach to combine models across monitoring stations and/or chemicals based on their interrelationships (discussed in section 2.2.3) using multivariate modelling and to deal with uncertainties, temporal dynamic changes in chemical concentrations using Bayesian framework as illustrated in Figure 3.1. As a result, we can provide interval estimation of annual chemical fluxes.

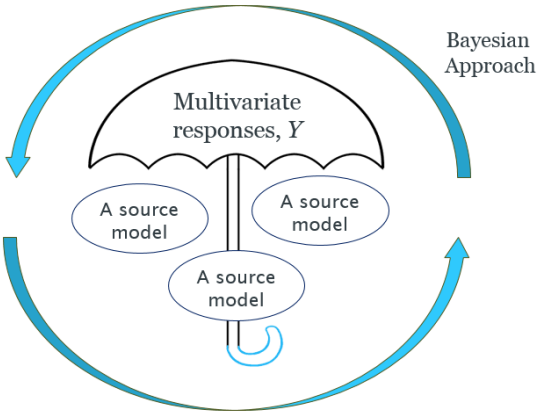


FIGURE 3.1: An illustration of Bayesian Multivariate Modelling

To investigate our approach, the model is developed for observations $\boldsymbol{y_t} = (y_{1t}, y_{2t}, ..., y_{dt})'$, which denotes the natural logarithm of chemical concentrations at time t, where $t = 1, ..., T$, T is the number of chemical concentrations each source (assume the equivalent of T for all sources at first), d is the number of chemical concentration sources and $j = 1, ..., d$ indicated the j^{th} source, for example, $d = 2$ indexing with j = {1,2} representing two monitoring sites on two rivers which are Knapp Mill and Throop monitoring site respectively.

We assume that the joint chemical concentrations are multivariate normal distributed with mean vector $\boldsymbol{\mu_t}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. The Bayesian hierarchy model

for multivariate Gaussian measurement error model can be written in general as

$$\boldsymbol{y_t} \overset{iid}{\sim} N_d\left(\boldsymbol{\mu_t}, \Sigma\right),$$

$$\boldsymbol{\mu_t} = \begin{pmatrix} \mu_{1t} \\ \mu_{2t} \\ \vdots \\ \mu_{dt} \end{pmatrix}, \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{bmatrix}$$

$$\epsilon_t \overset{iid}{\sim} N(0, \Sigma)$$

where $\boldsymbol{\mu_t}$ denotes the time varying mean and following the same choice of function of mean chemical concentration $\boldsymbol{\mu_t}$ such as regression function for all sources in the second stage of modelling hierarchy to capture the variation of chemical concentrations, $\sigma_1^2, \sigma_2^2, ..., \sigma_d^2$ are the variance of y_t and the covariance $\sigma_{ab} = \sigma_{ba} = \rho_{ab}\sigma_a\sigma_b$, $|\rho_{ab}| < 1$, $a, b = 1, ..., d$, $a \neq b$. ϵ_t is normally distributed with mean 0 , variance Σ.

We specify prior distribution of the conjugate prior distribution for the precision matrix $\Phi = \Sigma^{-1}$ is Wishart distributed, $\Phi \sim Wishart_d(\mathbf{V}, k)$ where positive symmetric matrix $\mathbf{V}_{d\times d}$, d is the dimension of the distribution (i.e the number of chemical concentrations sources), and k is the degree of freedom, $k \geq d$. Note that the lower k, the less informative distribution and the density is proper if $k \geq d+1$ (Gelman et al. 2014). A suggested choice of $\mathbf{V}$ is $d\Sigma_0$, where Σ_0 is a prior guess of the covariance (Lunn et al. 2013). In this work, we let $d = 2$, $k = 3$ for the least informative proper prior, and, $\mathbf{V} = d\Sigma_0$ where Σ_0 is $0.001 \times I_2$.

Practically, the model can be extended in three different way: 1) combining chemical concentration model across monitoring sites for each chemical 2) combining all chemical concentration models within the same monitoring site and 3) combining all chemical concentration models across monitoring sites. Thus the dimension of the Bayesian multivariate normal model can be varied depending on a combination of chemical concentration models. In this study, we investigate chemical concentration models individually as **univariate modelling case**, in comparison with combining models as **multivariate modelling case**.

We apply the Bayesian multivariate normal model with datasets from Christchurch Harbour and Hubbard Brook study. We combines the data from two monitoring sites from two rivers to examine chemical concentrations delivering into the Christchurch estuary. Some related variables are collected as independent variables or causes of variation using regression models. In the case of Hubbard Brook, we study solute levels based on

time series models. In addition, we consider more on combining tiers which are monitoring sites, solutes, or both. Examples of application are detailed in Chapter 4 and 5 respectively.

Chapter 4

Macronutrient modelling: the Christchurch Harbour

4.1 Introduction

Ecosystems in rivers and estuaries play a very important role in human health and welfare, for example, supply of drinking water from the rivers in England and Wales.

The primary macronutrients in river - nitrogen (N) and phosphate (P)- are vital substances for organism functioning. However, these may become water pollutants when these are found in excess concentrations. A great concern of ecological effects from excessive nutrient supply called Eutrophication, is intensively discussed as a global topic. A typical example is an algal bloom or rapid growth of algae in water resources. It causes water quality degradation and depletion of dissolved oxygen in water leading to elimination of fish and other aquatic life in freshwater and coastal ecology (Smith et al. 1999).

Human activities such as agriculture and emptying of sewage system are common sources of excessive riverine macronutrients (Chapman et al. 1996, Withers & Lord 2002). In particular, 59 percentages of nitrate level in UK rivers are from agriculture ADAS (2007). Moreover, under climate change, severe weather is more likely increasing in number and intensity of storms that causes higher mean river flow and higher flood frequency. These seasonal and storm-driven events contribute to macronutrient dynamics in rivers and also inputs to estuaries. (Withers & Lord 2002, Sigleo & Frick 2007, Cai et al. 2008).

The fact that increase of plant population (algae in particular) depends on the high level of nutrient concentrations. Many researchers believe that plant growths are also sensitive to proportion of nutrients, e.g. N:P, calling (co)limiting nutrients as ecological balance. Moreover, water qualities such as temperature also affect the density of aquatic organic biomass even though water light transmission and short hydraulic retention time

have been proved that they are not always strongly related to algae responses. (Smith et al. 1999).

In order to quantify macronutrients fluxes from inputstream to an estuary and its responses to external disturbances such as floods and local storms. Stochastic models can be developed to describe nutrient dynamics and seasonal using Bayesian modelling that allow us to assess and handle uncertainty variations that arise from multiple sources.

4.1.1 Purpose of analysis

To understand the macronutrients cycle, the relationship between nutrient concentrations and water qualities are analysed using statistical modelling. The change in nutrient levels and other water characteristics will be tracked to estimate the nutrient fluxes into the Christchurch Harbour estuary.

The current analysis is focusing on weekly data from two input streams: the Avon and Stour river to describe macronutritent concentrations i.e. nitrate (NO3) and phosphate (PO4) which are the most common form of nitrogen and phosphorus. In addition, the effects from seasonal and storms will also be investigated in changing of macronutrient levels and loads from rivers to estuary.

Firstly, we are modelling each macronutrients each rivers separately based on exist models and data characteristics. It will then be assessing model adequacy and performance.

Consequently, we are constructing an collaborative Bayesian modelling due to the geological pathway of the study area which macronutrients are transported along the two rivers before loading to the estuary in order to quantify annual macronutrient fluxes. It will be applied and examined with a simulation study and a real data example.

This may improve our understanding of the macronutrient dynamics and provide more accurate estimation. As a benefit, we can construct a confidence interval of that total annual fluxes as well.

4.1.2 Literature reviews

Increase of (artificial) nutrients is harmful to aquatic ecosystem such as algal bloom causes depletion of dissolved oxygen required by fish and shellfish. Nutrient levels vary not only as a result of human activities but also with the natural occurrence of storms attributed to climate change. Understanding the dynamics of nutrient cycles will enable us to maintain and restore the natural behaviour.

Water data are usually obtained from sequential water sampling for certain periods. The features of data themselves are positive, containing outliers, below detection limit or

censored, presenting seasonal pattern, exhibiting serial dependence, outlining a positive skewed and non-normal distribution (Helsel & Hirsch 2002).

A number of techniques have been used to track water quality changing and perform water quality model. Based on normality and constant variance assumptions of residuals, simple linear regression models can be constructed for the data with and without some data transformation to maintain linearity. However, strong pattern of seasonality requires a special effect consideration. We may apply multiple regression models with additional terms, for example, periodic or sinusoidal function. Furthermore, water characteristics of high skewness, missing values, and serial dependence require the more appropriate non-parametric techniques which do not make the normality assumption.

Natural variation such as rainfall and hydroflow often affect on interesting water qualities and macronutrient levels. It needs to identify more on important factors in order to achieve more accurate models. However, expert suggestion and simpler models are more acceptable for small samples (Loucks et al. 2005).

In addition, water quality models can be constructed using process-driven (mechanistic) models and/or data-driven (statistical) models. It is necessary to understand natural model chain for setting a process model using simulation. This is a more effective model, but may not respond to uncertainties as well as empirical model. Therefore, it is suggested to construct a combined model based on process-based and empirical models (Loucks et al. 2005).

Hirsch (2012) studied the effect of Tropical Storm Lee on Nitrogen, Phosphorus and Suspended sediment levels and fluxes behaviour in river. He found that the storm contributed input stream flow about 1.8 percent for the total from River Susquehanna and might increase nutrients release speed of sediment over time. It conducted about 5, 22, and 39 percent of increment in fluxes of Nitrogen, Phosphorus, and Sediment respectively in River Susquehanna and stored in Conowingo reservoir throughout the Chesapeake Bay in the USA.

Schoch et al. (2009) studied time series of nitrate concentrations in River Des Moines entering the Saylorville reservoir in Iowa, USA. Inflow nitrate concentrations were modelled with long memory function AR(20) with monthly concentration prediction almost all fell within 95% prediction interval of actual values and there is low effect of flow variation. A square root transformation was required to achieve normality and maximum likelihood estimation of nitrate concentration.

Stenback et al. (2011) evaluated multiple regression models for nitrate and total phosphorus load in River Iowa and stream using discharge of nutrients and time (decimal year) as explanatory variables. The models were constructed on the log scale. It concluded that the overestimation of nitrate transmission could be adjusted with some factors, for example, fertilizer applications.

Kadlec & Hammer (1988) developed a dynamic simulation of nutrients effects on the wetland ecology. The nutrients mathematical model of Nitrogen, Phosphorus and Chloride described variation from surface water flow, biomass growth patterns, soil, and mineral balancing to understand the wetland ecosystem.

Pirani et al. (2016) proposed Bayesian dynamic models of macronutrients in the Hampshire Avon river which account for storm-driven events based on daily nutrients (Nitrate and Phosphate) and water qualities taken between 22 November 2013 and 19 December 2014. LASSO method has been applied as a variable selection technique. The proposed model estimates uncertainty and dynamic components using nonlinear regression with penalised splines and change points structure. As a result, estimates of nutrient inputs and river streams have been used in macronutrient load calculation to the Christchurch Harbour estuary, UK.

4.2 Statistical modelling macronutrient dynamics

In this section, we are exploring the relationship between riverine macronutrients and water characteristics based on studied models by Pirani et al. (2016) using one year time series weekly data as described in Chapter 2.1.3. It appears the seasonal pattern and temporal climate change effects on macronutrients level fluctuation by a number of storms and higher river flow rate. We will also estimate macronutrient annual load to the Christchurch harbour estuary.

Nonlinear regression is a popular statistical method applied to nutrient models with relevant covariates. However, it may not be flexible to uncertainty effects, for instance, floods and storms, as a typical feature of environmental data. To cope with this event, Bayesian is a considerable approach to improve modelling. Moreover, modern statistical techniques such as change-point analysis and penalised spline regression can also be applied for tracking changes within measurements over time. (Qian et al. 2005, Alameddine et al. 2011, Pirani et al. 2016).

A similar work by Pirani et al. (2016) examined modified regression models as described below. A proposed model, penalised spline regression with change-point structure, captures well in macronutrients variation (nitrate and phosphate). The model was modified to focus more on interaction between river flow and time to cope with uncertainty events. Using daily data for a year from a monitoring site at Knapp Mill on the Hampshire Avon river, the model was constructed and investigated in prediction ability. This work also provided macronutrient annual load estimation compared with previous research.

This work is a study to evaluate those models performance to less frequency or weekly data. Water samples were weekly manual collected at a monitoring site from each of the two major rivers (the Avon and Stour) flowing through the Christchurch harbour

estuary during late April 2013 to early April 2014. These samples would be measured and analysed for macronutrients including with levels of physical and chemical properties. The daily mean river flow obtained by the UK Environment Agency is included to the study in particular river flow rate on sampling dates. Preliminary analysis of the data is described above and the data comparison is shown in Table 4.1.

Another important aim of this study is to estimate the annual load of macronutrients (sometimes called the annual fluxes/budget) to the estuary based on the weekly data. To account for a whole year of estimates with 51-week data, a replicate last record (week 51) will be attached as an extra last week (week 52) with corresponding daily mean river flow. The annual load will then be calculated with a chosen model.

TABLE 4.1: Data collection comparisons

Aspects	**Current study**	**Pirani et al. (2016)**
Study site	Knapmill on Hampshire Avon and Throop on Stour	Knapmill on Hampshire Avon only
Sampling method	manual (spot)	automatic
Sampling frequency	once a week (at low-tide time)	10 minutes (water qualities) 8-15 hours (macronutrients)
Sampling interval	26/04/2013 to 10/04/2014 (51 weeks)	22/11/2013 to 19/12/2014 (393 days)
Data processing	average on laboratory replicate analysis	summation on-site measurements over a day
Macronutrients	Nitrate:NO3 (mg/l) Phosphate:PO4 (mg/l)	Nitrate:NO3 (mg/l) Phosphate:PO4 (mg/l)
River flow and water qualities	Daily mean river flow (m^3/s) Temperature (°C) Turbidity:SPM (g/l) Conductivity (mS/cm) Dissolved Oxygen (mg/l)	Daily mean river flow (m^3/s) Temperature (°C) Turbidity (NTU) Conductivity (μS/cm) Dissolved Oxygen (%)

However, Pirani et al. (2016) worked on macronutrient data from a single river, the Avon. The current study is working with data from the two rivers delivering macronutrients as different sources to the same estuary. Beyond the line, a literature review suggested a possible way to combine river models using the multivariate normal distribution with covariate errors to describe the total macronutrient fluxes of an estuary.

Next, each macronutrient each river will be constructed models independently using various statistical methods as univariate model. We also investigate the efficiency of combining macronutrient models from both rivers, for each macronutrient separately as multivariate model. The modelling will be described further under the Bayesian approach.

4.2.1 Macronutrient Univariate Model

Variations of riverine macronutrient levels are commonly driven by river flow and biological activities varying throughout the year (Alameddine et al. 2011). Hence, the variation may be explained with daily mean river flow corresponding to the most wash-off land nutrients and ecological activities in river. These activities may be characterised through water qualities which mainly correspond to macronutrients (nitrogen and phosphorus) stimulated riverine primary production. As a result, we then construct univariate models to describe the relationship between macronutrients and hydrological characteristics for each macronutrient each river individually.

According to the exploratory analysis, the data are transformed by taking natural logarithm (on macronutrient concentrations and daily mean river flow) and standardisation (on water qualities) in order to stabilize variance, to adjust scales, to treat data skewness, and to preserve linearity. Time series plots of macronutrients data appear seasonal and uncertainty events, e.g., local storms which may be affected by climate changes. Furthermore, scatter plots show nonlinear relationship between macronutrient levels and water qualities. In addition, using logarithm of macronutrient allows to construct additive models for convenience investigation.

Regarding the literature review and the data attributes as described, we are modelling macronutrient concentrations using statistical methods, for example, regression, nonparametric smoothing, change detection, and Bayesian approach. Particularly, we are studying models reported by Pirani et al. (2016) with the weekly data from two monitoring sites on two rivers: Knapp Mill (the Hampshire Avon river) and Throop (the Stour river). Both are two major rivers budgeting nutrients to the Christchurch harbour estuary. As a result, we may model the macronutrient levels and estimate the annual loads to the estuary.

Given the dataset $\{y_{st}, f_{st}, x_{st1}, x_{st2}, ..., x_{stp}\}$ where y_{st} denotes the observed natural logarithm macronutrient (nitrate or phosphate) concentration from monitoring site $s \in \{1 \text{ for Knapp Mill, } 2 \text{ for Throop}\}$ at week $t \in \{1, 2, ..., T = 52\}$. The f_{st} is the natural logarithm of daily mean river flow. The corresponding $(p = 4)$ standardised water qualities are denoted as x_{sit}, for $i = 1, 2, ..., p$. Therefore, macronutrient levels y_{st} will be modelled with daily mean river flow and water qualities to capture its variation. The same notation will be applied to each macronutrients individually.

The model is developed using Bayesian hierarchical approach, a hierarchical model for macronutrient y_{st} is first structured with independent Gaussian model assumption. We then consider the mean of macronutrient concentrations under the first hierarchy described as

$$y_{st} \sim N(\mu_{st}, \sigma_s^2),$$

where μ_{st} denotes the macronutrient response mean for site s at week t. The variance σ_s^2 is constant over time for each site because there are no replicates enough with one year of data to estimate it each week.

The second hierarchy, we assume homoscedastic errors ϵ_{st} be the independent normally distributed random error with mean zero and constant variance σ_s^2. Hence, the model for μ_{st} at week $t = 1, 2, ..., 52$ can be functioned with all p standardised water quality variables $x_{st\cdot}$ and daily mean river flow f_{st} to describe relationship and capture nonlinear, seasonal and temporal effects as

$$\begin{aligned} \mu_{st} &= g_s(x_{st\cdot}, f_{st}) + \epsilon_{st}, \\ \epsilon_{st} &\sim N(0, \sigma_s^2) \end{aligned}$$

where $g_s(.)$ is a plug-in function describing mean of macronutrient concentration μ_{st}. Choices of $g_s(.)$ and prior setting will be discussed further.

4.2.1.1 Modelling techniques

As discussed, macronutrient data experiences seasonal variation and temporal changes due to climate change event such as local storms. Hence, the mean of macronutrient concentration changes across seasons and storm periods. In order to keep track of changing, several statistical methods are applied to describe the macronutrient level fluctuation as functions of the average of macronutrient concentration μ_{st} over time; for instance, linear regression, penalised splines, and change point detection.

Linear regression Linear regression aims to model the relationship between a response variable (y) and p explanatory variables $(x_1, x_2, ..., x_p)$ in linearity using the observed data $\{y_t, x_{t1}, x_{t2}, ..., x_{tp}\}$ where $t = 1, ..., n$. It can be written in a general form as

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + ... + \beta_p x_{tp} + \epsilon_t \\ &= \beta_0 + \sum_{i=1}^{p} \beta_i x_{ti} + \epsilon_t \end{aligned} \tag{4.1}$$

where ϵ_t is a noise or error term of an unobserved effect.

Penalised splines regression To deal with nonlinear relationship between macronutrients and p water qualities, a penalized spline regression is implemented to model using low-rank thin-plate splines. Radial basis functions are also implemented to the spline following Crainiceanu et al. (2005) for Bayesian analysis in order to avoid under- or

over-fitting the data and optimize the fit than linear splines. Over the range of each covariates, it will be evenly distributed by a set of K knots where $k_1 < k_2 < ... < k_K$. Then a general basis function of a x covariate can be written as

$$g(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} b_k (x - k_k)_+^d \tag{4.2}$$

where intercept β_0 and regression coefficient β_1 is assumed to be fixed and unknown. A vecter $\mathbf{b} = (b_1, ..., b_K)'$ is a set of random weight parameters for linear basis function $(x - k_k)_+^d$ which is equal to $(x - k_k)^d$ if $(x - k_k)^d > 0$ and zero otherwise, with d degree of the spline. The $\mathbf{b}$ components are defined with independent normal prior distribution wit zero mean and σ_b^2 unknown variance and will be estimated from the model.

Model 4.2 will be applied to the ith covariates, $i = 1, ..., p$ at time $t = 1, ..., T$ as

$$g_i(x_{ti}) = \beta_{0i} + \beta_{i1} x_{ti} + \sum_{k=1}^{K} b_{ik} (x_{ti} - k_{ik})_+^d \tag{4.3}$$

Considering the same set of knots and the spline degree d on all covariates, the total contribution for all covariates based on an additive model is given by

$$\sum_{i=1}^{p} g_i(x_{ti}) = \sum_{i=1}^{p} \beta_{0i} + \sum_{i=1}^{p} \beta_{1i} x_{ti} + \sum_{i=1}^{p} \sum_{k=1}^{K} b_{ik} (x_{ti} - k_{ik})_+^d \tag{4.4}$$

All covariate intercept terms β_{0i} will be summed and considered as a global intercept β_0. The notation of β_{1i} will be replaced with β_i for convenience.

Change point detection Data investigation shows that macronutrients exhibit the seasonal effect with changes in the mean of concentrations. In statistical analysis, change detection or change point detection is identifying times when the probability distribution of a stochastic process or time series changes. In particular, step detection, is concerning with changes in mean of the process.

Moreover, macronutrients concentration changes is also altered by river flow. The model will then include the interaction between temporal seasonal changes and river flow rate using Bayesian changepoint-threshold model, following Crainiceanu et al. (2005) and Pirani et al. (2016). The term changepoint determines changes over time and the term threshold defines changes over the river flow range.

Denoting τ_s as the switch-point in time, and the two flow threshold parameter φ_{s1} and φ_{s2}. Thus, the interaction ν_{sth} can be defined as the product of the two indicator functions and the incremental river flow corresponding to δ_{sh} for $h = 1, ..., 4$ defining as follow.

1. $\delta_{s1} I(t < \tau) I(f_{st} < \varphi_{s1})(f_{st} - \varphi_{s1})$ defining the effect of incremental flow less than φ_1 before the switch-point in time.

2. $\delta_{s2} I(t < \tau) I(f_{st} \geq \varphi_{s1})(f_{st} - \varphi_{s1})$ defining the effect of incremental flow greater than φ_{s1} before the switch-point in time.

3. $\delta_{s3} I(t \geq \tau) I(f_{st} < \varphi_{s2})(f_{st} - \varphi_{s2})$ defining the effect of incremental flow less than φ_2 after the switch-point in time.

4. $\delta_{s4} I(t \geq \tau) I(f_{st} \geq \varphi_{s1})(f_{st} - \varphi_{s2})$ defining the effect of incremental flow greater than φ_{s2} after the switch-point in time.

where $I(A) = 1$ if A is true and 0 otherwise.

4.2.1.2 Model structure

This part is mainly focusing on defining plug-in average function of macronutrient concentration using various modelling techniques as described from simple to complex function. It especially introduces a change-point structure to accommodate macronutrient dynamic changes over time and river flow simultaneously.

The model in logarithm will be constructed with additive models. Independent variables will be separated into two parts: water qualities and daily mean river flow. The first part can be considered as regression analysis describing the relationship among water qualities to the macronutrient concentrations. This relationship will be modelled with two approaches: multiple linear regression and nonparametric penalised spline. The latter or daily mean river flow is recognised as a crucial variable which significantly relate with nutrient concentration and may reflect seasonal effect (Helsel & Hirsch 2002). In addition, it can be seen in Figure 2.9, the scatter plots show nonlinear relationship between macronutrient levels and water qualities in general. Thus, it is likely to construct the model using a nonparametric smoothing function $g_i(x_{sti})$ for each water qualities to describe the mean of macronutrient concentrations from each monitoring sites s (a site a river).

For data preparation before modelling, macronutrients (Y) and river flow rate (flow) are transformed with natural logarithm to stabilize the variance and rescale the large value. Water quality properties such as water temperature are standardised, to have mean zero and variance 1.

We assume prior distribution for global intercept $\beta_s 0$ and other regression coefficients β is normal distribution with mean zero and variance 10^4. The parameter δ for change point structures are given the same normal prior distribution with zero mean and variance 10^2. We assume the switch-point in time τ is uniformly distributed on [1,TT] where TT denotes the number of observations. We adopt uniform prior distribution for the

two river flow threshold parameter φ_{s1} and φ_{s2} on range of minimum and maximum of river flow in natural logarithm scale as [1.734,4.631] for Knapmill monitoring site and [0.888,4.938] for Throop monitoring site. For random weight parameter $\mathbf{b} = (b_1, ..., b_K)$ associated with regressors for linear basis function. Each component of $\mathbf{b}$ is denoted independent normal prior distribution with mean zero and variance σ_B^2. We assume the prior distribution of the precision or the inverse of variance σ_B^2 for each component is $Gamma(a, b)$ distribution with shape parameter $a = 1$ and scale parameter $b = 0.001$. For the precision matrix Σ^{-1} is assumed the least informative proper prior Wishart distribution with $d \times I_d$ where d is the number of chemical concentration sources and the degree of freedoom, $k = d + 1$ as discussed in section 3.4.

The mean function of macronutrient concentration modelling will be constructed as below:

Model M1 : a Penalised spline regression for the water quality properties and change-point structure on river flow This is the most flexible model coping with nonlinear, seasonal and temporal effects.

$$\mu_{st} = \beta_{s0} + \sum_{i=1}^{p} \beta_{si} x_{sti} + \sum_{i=1}^{p} \sum_{k=1}^{K} b_{ik} (x_{sti} - k_{sik})_+^d + \sum_{h=1}^{4} \delta_{sh} \nu_{sth} \tag{4.5}$$

where k_{sik} indicates equally spaced knot kth over $x_s, k_{s1} < k_{s2} < ... < k_{sK}$ for each water qualities. The last term is dealing with temporal changes and discontinuities of river flow as described.

Model M2 : a multiple linear regression for the water quality properties This is the simplest model which macronutrient levels are described with only water quality properties in linearity.

$$\mu_{st} = \beta_{s0} + \sum_{i=1}^{p} \beta_{si} x_{sti} \tag{4.6}$$

Model M3 : a multiple linear regression for the water quality properties, with a switch-point in time This model is a reduce version of equation 4.5 to deal with only step changing in the daily mean river flow over time using change-point analysis. This change is mostly due to seasonal effect and temporal effects such as local storm events. It can be seen in Figure 2.4.

$$\mu_{st} = \beta_{s0} + \sum_{i=1}^{p} \beta_{si} x_{sti} + \delta_{s1} I(t \geq \tau_s) \tag{4.7}$$

Model M4 : a multiple linear regression for the water quality properties, with a change threshold This model is a multiple linear regression with additional parts that only focus on changing step over river flow as the most factor dominating the macronutrient level or the response variable by defining φ_s.

$$\mu_{st} = \beta_{s0} + \beta_{sl} x_{slt} + \delta_{s1} f_{st} + \delta_{s2} I(f_{st} \geq \varphi_s)(f_{st} - \varphi_s) \tag{4.8}$$

Model M5 : a multiple linear regression for the water quality properties, with change-point structure on river flow This model is a multiple linear regression modified with change-point structure on river flow or a simple linear version of model (4.5).

$$\mu_{st} = \beta_{s0} + \beta_{sl} x_{slt} + \sum_{h=1}^{4} \delta_{sh} \nu_{sht} \tag{4.9}$$

Model M6 : a Penalised spline regression for the water quality properties This is a penalised spline regression version to examine the effective of fitting a smooth curve which more flexible to the data than linear relationship.

$$\mu_{st} = \beta_{s0} + \beta_{sl} x_{slt} + \sum_{l=1}^{p} \sum_{k=1}^{K} b_{kl} (x_{slt} - k_{skl})_+^d \tag{4.10}$$

The above models will be constructed to define the time varying mean μ_{st} from each site $s \in \{1 \text{ for Knapp Mill}, 2 \text{ for Throop}\}$ individually, as univariate model.

4.2.2 Macronutrient Bivariate Model

Cha et al. (2016) proposed a model to study cross-scale view of Nitrates and Phosphates limitation (N-P limiting) in river. Nitrate is believed to be the primary control factor of algae and phytoplankton growth, while Phosphate is the primary control factor of marine plant growth. The model approach is based on natural interaction of multiple chemical elements, or nutrient balancing in ecology. A bivariate Bayesian hierarchical model was applied to existing approaches by modelling multiple N-P relationship and allowing to vary by seasonal and spatial components.

Regarding the most of macronutrient in an estuary flowing along the rivers, in our case, the Hampshire Avon and the Stour are two major rivers delivering macronutrients to the Christchurch Harbour estuary from geography. In fact that the appearance of ecological data is the most induced by seasonal and surrounding environment. It can be seen similarities in time series plots on both rivers in Chapter 2. It also showed a moderate correlation of macronutrient levels between both rivers. Due to the data limitations or small data, we may benefit from combining information from related sources to improve

data modelling. Moreover, it is possible to quantify the point estimates of the annual total macronutrient loads from both or separate rivers; including with credible intervals of that estimates.

As discussed, we are interested in Bayesian bivariate normal modelling to combine data from different sources. Therefore, we are applying a bivariate normal structure as an umbrella to combine multiple models (refers to section 4.3.1) based on correlation of response variables across sources under Bayesian approach. The model will be investigated the advantage of combining method through a real problem study and a simulation.

The bivariate normal model constructs a model of two dependent response variables with covariance error term. The error distribution, being bivariate, is a 2×2 variance-covariance matrix. Let the vector of observations at time t, $\boldsymbol{y_t} = (y_{1t}, y_{2t})'$ which 1 and 2 represent Knapp Mill and Throop monitoring site respectively. If we assume macronutrient concentrations are bivariate normal distributed with mean vector $\boldsymbol{\mu_t}$ and variance-covariance matrix Σ. Bayesian hierarchy model for multivariate regression model can be written as

$$\begin{aligned}
&\boldsymbol{y_t} \overset{iid}{\sim} N_2\left(\boldsymbol{\mu_t}, \Sigma\right), \\
&\boldsymbol{\mu_t} = \begin{pmatrix} \mu_{1t} \\ \mu_{2t} \end{pmatrix}, \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \\
&\epsilon_t \sim N(0, \Sigma)
\end{aligned}$$

where $\boldsymbol{\mu_t}$ follow choices of μ describing in Section 4.3.1. σ_1^2 and σ_2^2 are the variance of y_{1t} and y_{2t} respectively. The covariance between y_{1t} and y_{2t} is $\sigma_{12} = \sigma_{21} = \rho\sigma_1\sigma_2$, $|\rho| < 1$. We are considering this bivariate normal with different Σ which allow to access an advantage of fitting Bayesian bivariate normal model.

- independent error model: a bivariate normal model with

$$\sigma_{12} = 0 \tag{4.11}$$

- dependent error model: a bivariate normal model with

$$\sigma_{12} \neq 0 \tag{4.12}$$

We specify prior distributions of regression coefficients β similar to the multiple regression model. For errors specification, the conjugate prior distribution for the precision matrix $\Phi = \Sigma^{-1}$ is Wishart distributed, $\Phi \sim Wishart_d(\mathbf{V}, k)$ where positive symmetric matrix $\mathbf{V}_{d\times d}$, d is the dimension of the distribution, and k is the degree of freedom, $k \geq d$. Note that the lower k, the less informative distribution and the density is proper if $k \geq d + 1$ (Gelman et al. 2014). A suggested choice of $\mathbf{V}$ is $d\Sigma_0$, where Σ_0 is a prior guess of the covariance (Lunn et al. 2013). In this work, we let $d = 2$, $k = 3$ for the least informative proper prior, and, $\mathbf{V} = d\Sigma_0$ where Σ_0 is $0.001 \times I_2$.

4.2.3 Macronutrient load calculation

In order to estimate the annual load, it can be calculated as the product of macronutrient concentration and mean daily river flow. However, water samples were taken from the river once a week at low tide time, we assumed this is a daily amount of macronutrients which will be delivered with the same amount everyday till the next sampling day. Hence the nutrient concentration at time t can be considered as loading rate (e.g. kg/day). Based on the select bivariate data model, we are calculating the load using the exponential of posterior men of macronutrient concentration ($mg/l/day$) denoted as c_i. On the other hand the daily mean river flow is measured in cube metre per second (m^3/s), then it will be calculated in cube metre per day denoted as q_i. As a result, the load estimate can be calculated for week load by multiple day load with 7 , then summed up for a year, equivalently with 52 weeks. These calculation should be done with MCMC samples of a select bivariate model to provide 95% credible interval (CI) of total annual load estimates.

$$Load = \sum_{i=1}^{52} (7 * (c_i * q_i) \tag{4.13}$$

In comparison among different areas, we may change the annual load unit to $Kg/year$ and normalised the annual load or divide the annual load by its catchment area which is 1706 Km^2 for Hampshire Avon (at Knappmill) and 1703 Km^2 for Dorset Stour (at Throop). This will present the normalised annual load in Kg/Km^2

4.2.4 Bayesian computation

The above models will be applied to macronutrients data under Bayesian approach. In order to derive corresponding posterior distribution of model parameters, we use Markov chain Monte Carlo (MCMC) methods, Gibbs sampler in particular, to obtain MCMC samples providing posterior statistics e.g. mean, median, quantiles, and 95% credible intervals.

Models and facilities are coded in the language R (version 3.3.0), then scripting with R2WinBUGS package (version 2.1.21; Sturtz et al. (2005)) before executing using free Windows-based software WinBUGS (version 1.4.3; Lunn et al. (2000)) to obtain Bayesian analysis and modelling. MCMC procedure is setting with the chain thinning for 10. A run length control diagnostic, Raftery and Lewis's diagnostic (Raffery & Lewis 1992, Cowles & Carlin 1996), is used to suggest a number of burn-in iterations, then the first 5,000 iterations at the beginning of the chain will be discarded as burn-ins. A longest 30,000 iterations of MCMC chain is set up under PC based resources and examined by convergence diagnostics of model parameter estimates.

4.3 Results

4.3.1 Results for Fitting Univariate Models

Pirani et al. (2016) models the daily data to describe nitrate and phosphate concentrations based on Bayesian approach, model M1-M6, with different components of linear regression (based model), penalised splined regression (flexible model), switch-point in time (change over time), and change-point structures (change interaction between river flow and time). Among those models, model M1, the proposed model of penalised spline regression with change-point structure has been claimed the most performance for describing those macronutrient concentrations with the least value both PMCC and the statistics R_B^2. It indicates the spline technique may adjust model flexibility and the change-point structure be able to capture dynamic changes over time and river flow.

Turning to our data, a year of weekly data with similar water qualities, it shows some common characteristics of water resources data: outliers on high side, seasonal patterns, and strong correlation with uncontrolled (natural) variables e.g. water flow (see Helsel & Hirsch (2002)) as discussed in Chapter 2. Particularly, presence of local floods and storms altered river flow dramatically, temporarily and uncertainly as seen in winter months. These behaviours reflect on both data from the Hampshire Avon and the Stour river.

To model such data, we examine Pirani et al. (2016) models in describing macronutrient concentrations in the two major rivers delivering to the Christchurch Harbour estuary. Consequently, the annual load budget of nitrate and phosphate to the estuary each river will be estimated.

Based on pilot runs on all models with specified prior distributions and MCMC setting, it noted some primary settings. The first for spline regression, the number of knots is set to 2 as there are no significant advantages with higher knots over complexity.

Next, some monitored parameters are dealing with identifiability and convergence problems; also inconsistency in parameter estimates. This may cause by fitting improper components to the data especially with change-point extensions e.g. change-point threshold and change-point structure. In detail, it showed an almost complete related between parameter estimates, for example between fraction of change-point threshold or two knots of penalised spline regression on some covariates or water qualities. The parameter cannot be estimated properly while showing a good convergence or vice versa. To obtain an equilibrium posterior distribution, these models would then be reduced; result in Markov chain convergence improvement and provide more sensible parameter estimates in the mean of daily river flow change points (changing in mean of daily mean river flow over time) for change-point component. However, 95% Credible Interval (CI) of daily mean river flow levels before and after the switch-point in time show an overlap.

It also diagnosed large variation of switch-point over time and daily mean river flow on MCMC estimates. It seems unlikely the different on that partitioning. This introduces the problem when there exists an embedded lower dimension model within a full model. (Gelfand & Sahu 1999)

The problem occurs in three models: the most complex model (M1) with the penalised spline and change-point structures, linear model for water quality data with a change threshold in river flow (M4), and linear model for water quality data with change-point structures (M5). It seems the data does not support well those complex models or the sample size is not enough to take advantage of the complexity. Furthermore, a linear model with a change threshold (M4) will not be studied as the reduced model becomes the simplest linear model M2.

Table 4.2 displays the model criteria PMCC and the statistics R_B^2 for which it is convenient to consider the goodness of fit and model complexity simultaneously. For all nitrate models compared with the simplest model M2, integrated components e.g. spline regression seems perform well while change detection techniques such as change threshold and change-point structures does not improve the goodness of fit as expected. Considering model M1 and M6 which applied the spline regression, the goodness of fit is much lower than the model M2. However, it is unable to against its complexity penalty (P). Based on the PMCC, it suggests the simplest model M2 with the lowest PMCC in overall. However, it seems not to be significant different among nitrate models. Contrasting with the PMCC, the traditional statistic R_B^2 obviously indicates the model M1 and M6. However, the statistic R_B^2 are agree with goodness of fit score. In summary, it seems not to be different among nitrate models, similarity in PMCC with obvious difference in R_B^2.

For phosphate models, the change detection performs a better goodness of fit and improved the model efficiency with significant decreasing in PMCC while there is not much effect of the penalised spline regression; it can be seen from model M1 and M6 which both use spline techniques but M1 also integrated the model with change point structure; in contrast with nitrate models. The statistic R_B^2 also indicates in the same direction. The PMCC and R_B^2 suggest the model M1.

Table 4.3 shows results of a validation technique, the k-fold cross-validation with k=10. In order to assess the model in predictive performance, the less value of statistic: the root mean square error (RMSE),the mean absolute error (MAE), and the continuous ranked probability score (CRPS); the more model predictive accuracy is obtained. The statistic for all models are similar, the predictive ability does not show significant difference among the candidate models. However, it appears that the nitrate models outperform the phosphate models, which have lower statistics.

On both rivers, the overall results are consistent. A model that performs the best in terms of overall performance will be chosen to examine the advantage of multivariate

modelling. We may choose the model M1, the penalised spline regression for water quality data with change-point structures to describe macronutrient concentrations for both rivers. Table 4.9 and 4.10 present parameter estimations of the chosen model, M1.

TABLE 4.2: Predictive model choice criterion (G+P) of nitrate and phosphate concentration models; partitioning into goodness of fit (G), penalty (P), along with the statistic R_B^2

	Nitrate				**Phosphate**			
Models	G	P	PMCC (G+P)	R_B^2	G	P	PMCC (G+P)	R_B^2
Knapp Mill gauging station on the River Hampshire Avon								
M1	0.21	3.14	3.35	0.82	12.14	24.92	37.06	0.73
M2	0.38	2.76	3.15	0.66	31.39	38.72	70.11	0.29
M3	0.29	2.83	3.11	0.75	21.79	32.33	54.12	0.51
M5	0.34	2.84	3.19	0.70	17.93	30.26	48.19	0.59
M6	0.22	3.07	3.29	0.81	27.95	38.56	66.51	0.37
Throop gauging station on the River Stour								
M1	0.46	3.32	3.78	0.53	2.47	7.54	10.00	0.90
M2	0.54	2.95	3.50	0.44	11.92	16.19	28.12	0.52
M3	0.51	3.06	3.56	0.48	7.73	13.13	20.86	0.69
M5	0.53	3.03	3.56	0.46	3.12	6.90	10.02	0.88
M6	0.47	3.23	3.70	0.52	10.90	16.51	27.40	0.56

Note: M1: Penalised spline for water quality data with change-point structures

M2: Linear model for water quality data without change-point structures

M3: Linear model for water quality data with a switch-point in time

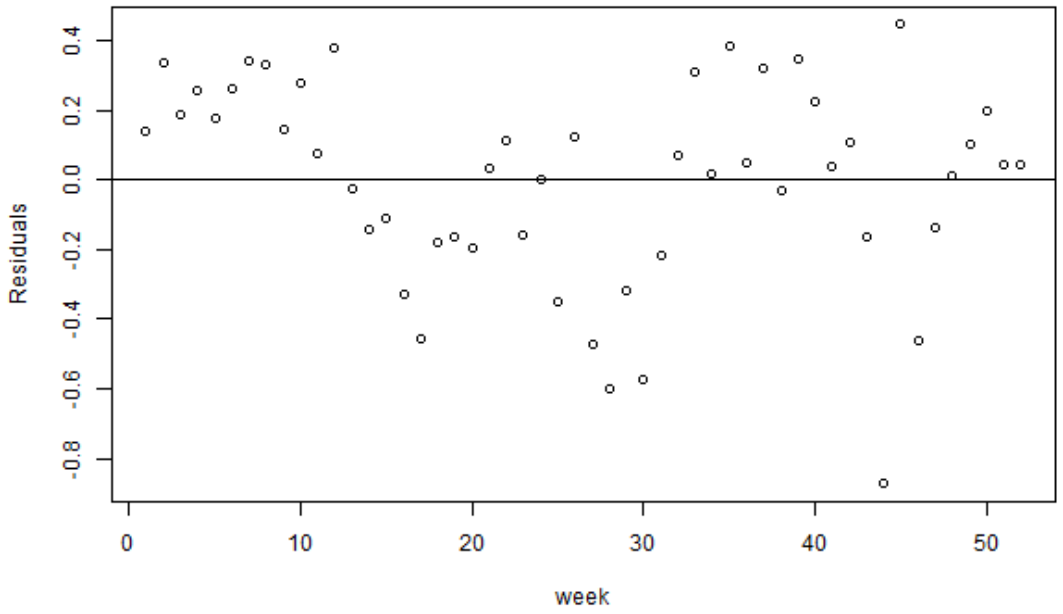
M5: Linear model for water quality data with change-point structures

M6: Penalised spline for water quality data without change-point structures

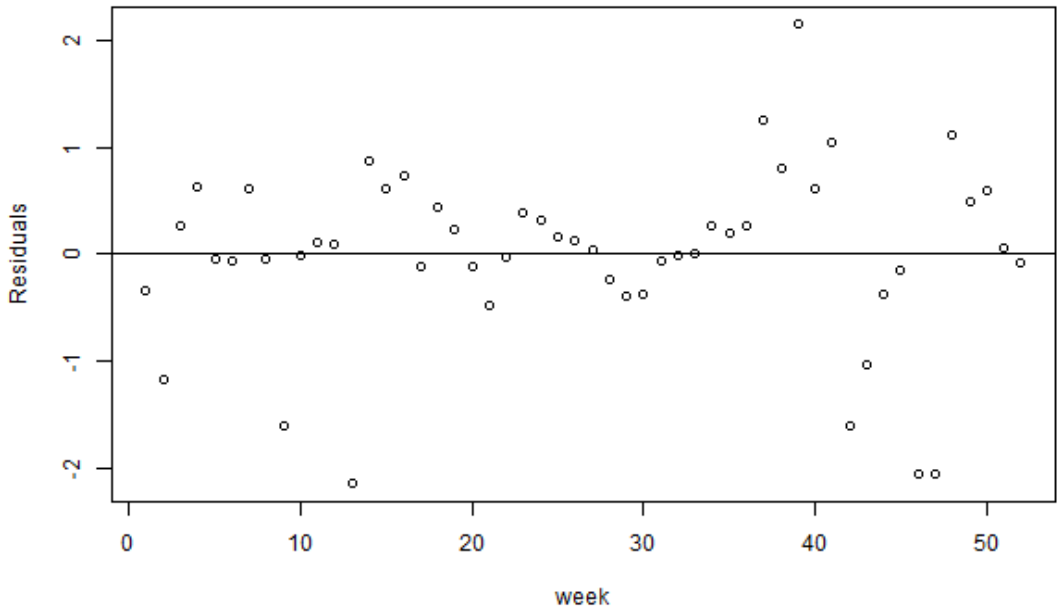
TABLE 4.3: The average 10-fold cross-validation results for nitrate and phosphate models describing with the root mean square error (RMSE), the mean absolute error (MAE), and the continuous ranked probability score (CRPS)

	Nitrate			**Phosphate**		
Models	**RMSE**	**MAE**	**CRPS**	**RMSE**	**MAE**	**CRPS**
Knapp Mill gauging station on the River Hampshire Avon						
M1	0.082	0.060	0.072	0.729	0.520	0.400
M2	0.098	0.080	0.071	0.811	0.637	0.456
M3	0.088	0.070	0.081	0.782	0.605	0.438
M5	0.086	0.069	0.070	0.731	0.525	0.402
M6	0.085	0.066	0.071	0.849	0.646	0.471
Throop gauging station on the River Stour						
M1	0.120	0.087	0.085	0.493	0.308	0.234
M2	0.116	0.091	0.078	0.504	0.420	0.288
M3	0.117	0.091	0.093	0.391	0.308	0.223
M5	0.116	0.089	0.079	0.333	0.235	0.184
M6	0.111	0.085	0.081	0.580	0.448	0.311

To assess the adequacy of the chosen model M1 for the macronutrients data, the standardised residual plots for both rivers are provided in Figure 4.1 and 4.2. It plots the median of the posterior distributions of the standardised residual against the time period for nitrate and phosphate. There is no apparent pattern. It scatters around zero random with few large values. This result supports an overall adequacy of the model for the data.

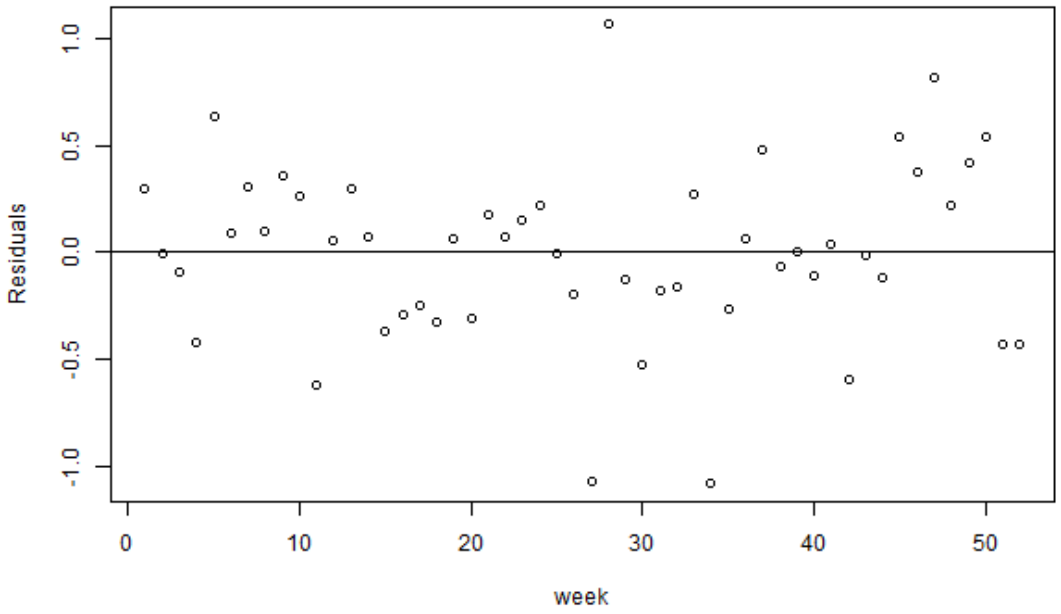


(a) Nitrate

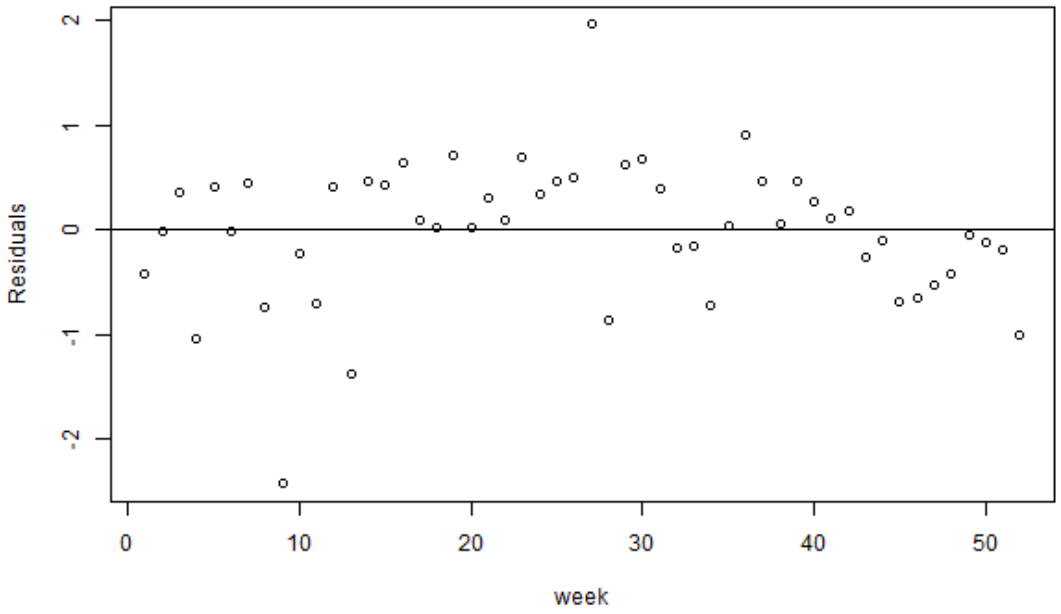


(b) Phosphate

FIGURE 4.1: Standardised residuals of model M1 of macronutrient concentration from Knapp Mill station (the Hampshire Avon river)



(a) Nitrate



(b) Phosphate

FIGURE 4.2: Standardised residuals of model M1 of macronutrient concentration from Throop station (the Stour river)

4.3.2 Results for Fitting Bivariate Models

In this section we investigate the efficiency of the Bayesian bivariate normal model by combining data from two major rivers: the Hampshire Avon and The Stour, to estimate macronutrients loads to the Christchurch Harbour estuary. We analyse the macronutrient data for 52 weeks during 26 April 2013 to 10 April 2014. The data are collected and sampled to measure from two monitoring sites: Knapp Mill on the Hampshire Avon and Throop on the Stour. The data contains macronutrient concentrations (nitrate and phosphate), daily mean river flow rate, and four water qualities (temperature, conductivity, dissolved oxygen, and turbidity). In section 4.3.1, each macronutrient each rivers are analysed and modelled separately. However, an object to study macronutrient is not only its relationship with flow and water qualities, but also to quantify macronutrient annual budget to the estuary which the most delivered from two rivers. Based on the geography of study area, data limitation and annual budget estimation, we are interesting to construct a combine data model using the Bayesian bivariate normal modelling. The model may also provide a property to assess data uncertainty which typically occurs in natural data.

To evaluate the Bayesian bivariate model, we then model and compare models with the same practised model structure in Chapter 4.3.1 under the bivariate framework stated in section 4.2.2. By replacing the variance-covariance matrix, the model becomes two different model sets: dependent models be our interested model (covariance exists) and independent models (zero covariance) represents combination of two independent models.

We start with checking our written code by comparing results between individual models as shown in section 4.3.1. Technically, the latter is similar with individual models, but only constructs under the Bayesian bivariate with zero covariance (independent). Table 4.4 and 4.5 show model choice criterion PMCC and R_B^2 for comparison. Overall, the results show similar results. A small difference may cause by randomization. For comparison, we define a random seed and apply to all modelling which each models start random at the same point. On the other hand, bivariate model is constructed with data from both rivers. Then, the same random seed will affect only once for each combinations. However, it proves our model and computer code for constructing the Bayesian bivariate normal model with a covariance matrix. Hence, we may use only model 4.11 instead of individuals in Section 4.3.1 for convenience to represent a case of modelling without bivariate normal construction.

TABLE 4.4: Predictive model choice criterion (G+P) of nitrate concentration models; partitioning into goodness of fit (G), penalty (P), along with the statistic R_B^2: comparison of individual models and independent models with the same model structure under the Bayesian bivariate normal framework (zero covariance)

	Individual				**Independent**			
Models	G	P	PMCC (G+P)	R_B^2	G	P	PMCC (G+P)	R_B^2
Knapp Mill gauging station on the River Hampshire Avon								
M1	0.21	3.14	3.35	0.82	0.21	3.14	3.35	0.82
M2	0.38	2.76	3.14	0.66	0.38	2.76	3.15	0.66
M3	0.29	2.83	3.11	0.75	0.29	2.83	3.12	0.75
M5	0.34	2.84	3.19	0.70	0.35	2.84	3.19	0.69
M6	0.22	3.07	3.29	0.81	0.21	3.07	3.28	0.81
Throop gauging station on the River Stour								
M1	0.46	3.32	3.78	0.53	0.45	3.32	3.78	0.53
M2	0.54	2.95	3.50	0.44	0.54	2.96	3.50	0.44
M3	0.51	3.06	3.56	0.48	0.51	3.06	3.56	0.48
M5	0.53	3.03	3.56	0.46	0.53	3.03	3.56	0.46
M6	0.47	3.23	3.70	0.52	0.46	3.24	3.71	0.52

TABLE 4.5: Predictive model choice criterion (G+P) of phosphate concentration models; partitioning into goodness of fit (G), penalty (P), along with the statistic R_B^2: comparison of individual models and independent models with the same model structure under the Bayesian bivariate normal framework (zero covariance)

	Individual				**Independent**			
Models	G	P	PMCC (G+P)	R_B^2	G	P	PMCC (G+P)	R_B^2
Knapp Mill gauging station on the River Hampshire Avon								
M1	12.14	24.92	37.06	0.73	12.10	24.83	36.93	0.73
M2	31.39	38.72	70.11	0.29	31.37	38.79	70.16	0.29
M3	21.79	32.33	54.12	0.51	21.65	32.42	54.07	0.51
M5	17.93	30.26	48.19	0.59	18.03	30.13	48.16	0.59
M6	27.95	38.56	66.51	0.37	27.90	38.73	66.63	0.37
Throop gauging station on the River Stour								
M1	2.47	7.54	10.00	0.90	2.50	7.49	10.00	0.90
M2	11.92	16.19	28.12	0.52	11.93	16.23	28.16	0.52
M3	7.73	13.13	20.86	0.69	7.80	12.96	20.75	0.69
M5	3.12	6.90	10.02	0.88	3.11	6.91	10.02	0.88
M6	10.90	16.51	27.40	0.56	10.90	16.52	27.42	0.56

Next, we compare the efficiency of macronutrient models with and without the bivariate property on the two rivers by specifying a covariance matrix. Table 4.6 and 4.7 display model choice criterion PMCC and R_B^2 to compare predictive quality of model under the Bayesian bivariate framework between independent models with zero covariance (model 4.11) and dependent models with covariance (model 4.12).

For nitrate, there is a considerably drop in PMCC for model with the bivariate property model (dependent), especially with complex models. The most decreasing on PMCC happen in penalty term. This refers to variance of prediction on observed over MCMC iterations is reduced. Thus, combining data, it may provide more information to better capture relationship between macronutrients and water characteristics.

Surprisingly for phosphate, PMCC slightly increase for Knapp Mill bivariate model and decrease for Throop bivariate model. It seems having more data does not improve modelling as expected. It might be because model or selected covariates capture only a part of macronutrients variation.

TABLE 4.6: Predictive model choice criterion (G+P) of nitrate concentration models; partitioning into goodness of fit (G), penalty (P), along with the statistic R_B^2: comparison of dependent models (covariance exists) and independent models (zero covariance) with the same model structure under the Bayesian bivariate normal framework

	Dependent				**Independent**			
Models	G	P	PMCC (G+P)	R_B^2	G	P	PMCC (G+P)	R_B^2
Knapp Mill gauging station on the River Hampshire Avon								
M1	0.12	0.24	0.36	0.89	0.21	3.14	3.35	0.82
M2	0.38	0.46	0.85	0.66	0.38	2.76	3.15	0.66
M3	0.23	0.32	0.56	0.79	0.29	2.83	3.12	0.75
M5	0.17	0.32	0.49	0.85	0.35	2.84	3.19	0.69
M6	0.15	0.25	0.40	0.86	0.21	3.07	3.28	0.81
Throop gauging station on the River Stour								
M1	0.44	0.66	1.10	0.55	0.45	3.32	3.78	0.53
M2	0.55	0.68	1.23	0.43	0.54	2.96	3.50	0.44
M3	0.48	0.66	1.15	0.50	0.51	3.06	3.56	0.48
M5	0.51	0.68	1.19	0.47	0.53	3.03	3.56	0.46
M6	0.45	0.65	1.10	0.54	0.46	3.24	3.71	0.52
Total								
M1	0.56	0.90	1.46	0.86	0.66	6.47	7.12	0.83
M2	0.93	1.14	2.08	0.76	0.92	5.72	6.64	0.77
M3	0.72	0.99	1.70	0.82	0.79	5.89	6.69	0.80
M5	0.68	1.00	1.68	0.83	0.88	5.87	6.75	0.78
M6	0.60	0.90	1.50	0.85	0.68	6.31	6.99	0.83

TABLE 4.7: Predictive model choice criterion (G+P) of phosphate concentration models; partitioning into goodness of fit (G), penalty (P), along with the statistic R_B^2: comparison of dependent models (covariance exists) and independent models (zero covariance) with the same model structure under the Bayesian bivariate normal framework

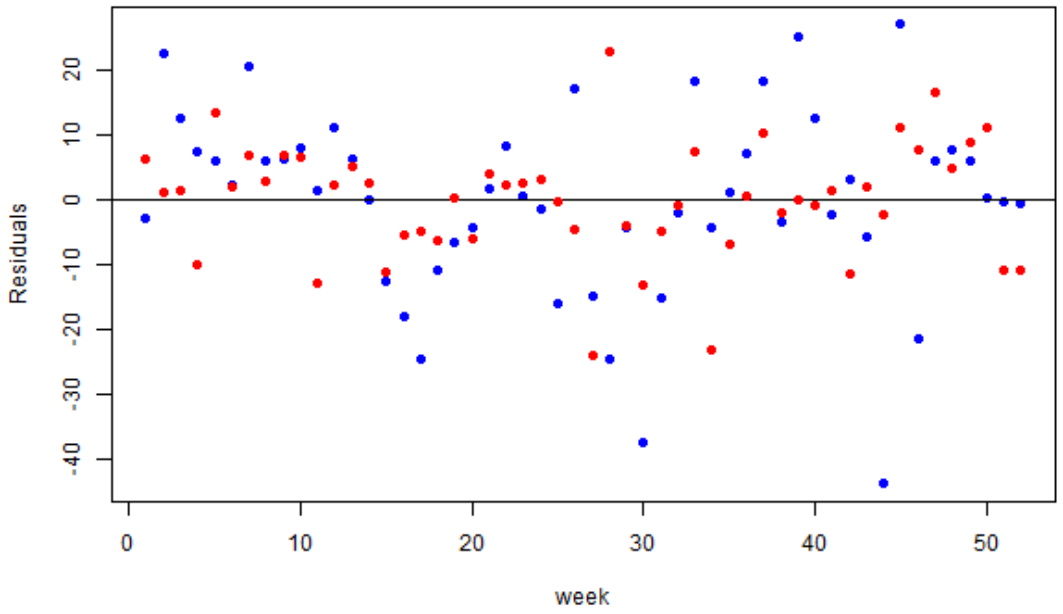
	Dependent				**Independent**			
Models	G	P	PMCC (G+P)	R_B^2	G	P	PMCC (G+P)	R_B^2
Knapp Mill gauging station on the River Hampshire Avon								
M1	13.55	23.95	37.51	0.69	12.10	24.83	36.93	0.73
M2	34.40	39.85	74.25	0.22	31.37	38.79	70.16	0.29
M3	26.39	34.16	60.55	0.40	21.65	32.42	54.07	0.51
M5	20.63	28.62	49.25	0.53	18.03	30.13	48.16	0.59
M6	32.38	40.45	72.83	0.27	27.90	38.73	66.63	0.37
Throop gauging station on the River Stour								
M1	2.65	4.59	7.24	0.89	2.50	7.49	10.00	0.90
M2	12.27	14.14	26.41	0.51	11.93	16.23	28.16	0.52
M3	8.33	10.47	18.80	0.67	7.80	12.96	20.75	0.69
M5	4.25	6.04	10.29	0.83	3.11	6.91	10.02	0.88
M6	11.58	14.23	25.81	0.54	10.90	16.52	27.42	0.56
Total								
M1	16.20	28.54	44.74	0.90	14.60	32.32	46.93	0.91
M2	46.67	53.99	100.66	0.71	43.30	55.02	98.32	0.73
M3	34.72	44.62	79.34	0.78	29.45	45.38	74.82	0.81
M5	24.88	34.66	59.53	0.84	21.14	37.03	58.18	0.87
M6	43.96	54.68	98.64	0.72	38.81	55.25	94.05	0.76

Based on small data, we use a 10-fold cross-validation procedure to evaluate the bivariate model in prediction. Data from both rivers are combined to construct macronutrient model. Hence, data (52+52 records) is divided into 10 folds randomly To obtain statistics RMSE, MAE, and CRPS. Each fold becomes a test set to validate a result model trained with the remaining data. However, data in week 1, 35, 36, and 52 are assigned only as training data due to lack of predictive property on the starting point, unusual events, and lost sampling weeks. Table 4.8 shows statistics to evaluate the predictive quality of the bivariate model. Overall, decreasing in statistics indicates that the bivariate model or combining data may improve the efficiency of prediction.

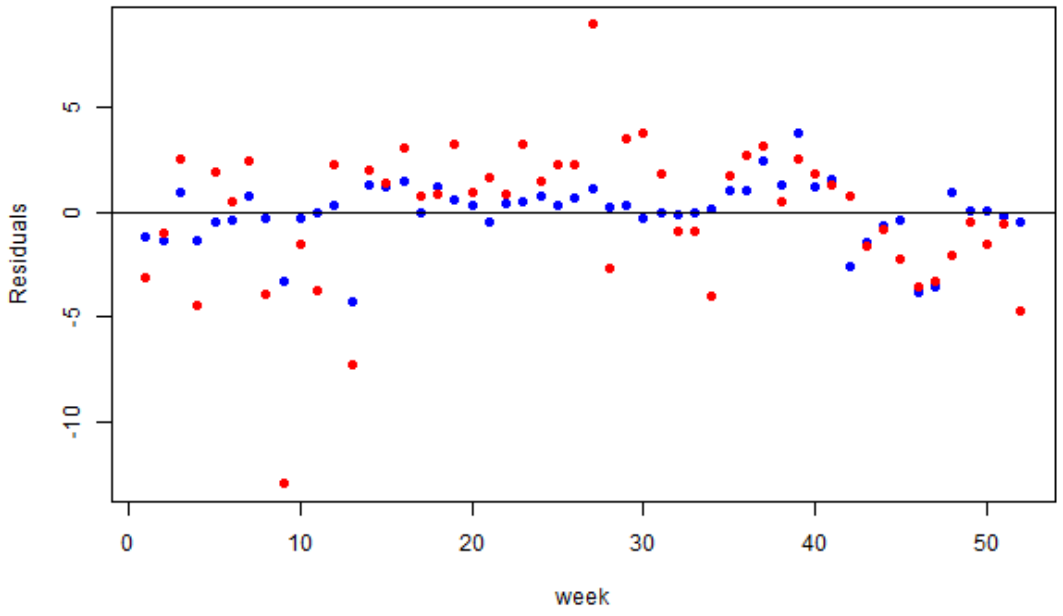
TABLE 4.8: The average 10-fold cross-validation results for nitrate and phosphate models describing with the root mean square error (RMSE), the mean absolute error (MAE), and the continuous ranked probability score (CRPS)

	Dependent			**Independent**		
Models	**RMSE**	**MAE**	**CRPS**	**RMSE**	**MAE**	**CRPS**
Nitrate						
M1	0.109	0.076	0.057	0.103	0.077	0.078
M2	0.109	0.081	0.059	0.105	0.082	0.074
M3	0.102	0.074	0.054	0.100	0.077	0.073
M5	0.110	0.080	0.058	0.103	0.080	0.075
M6	0.093	0.068	0.051	0.103	0.077	0.077
Phosphate						
M1	0.564	0.352	0.276	0.561	0.355	0.284
M2	0.542	0.415	0.297	0.683	0.529	0.375
M3	0.468	0.353	0.253	0.618	0.459	0.334
M5	0.436	0.323	0.231	0.499	0.338	0.264
M6	0.599	0.423	0.304	0.733	0.540	0.383

To assess the adequacy of model for the macronutrients data, the standardised residual plots of a sample model M1 for both nutrients are provided in Figure 4.3. It plots the median of the posterior distributions of the standardised residual against the time period for nitrate and phosphate. There is no apparent pattern. It scatters around zero random with few large values. This result supports an overall adequacy of the model for the data.



(a) Nitrate



(b) Phosphate

FIGURE 4.3: Standardised residuals of model M1 of macronutrients concentration from Knapp Mill station (blue) and Throop station (red)

4.3.3 Parameter and flux estimates

4.3.3.1 Macronutrient Univariate Model

Parameter estimates for the chosen model M1 are presented in Table 4.9 and 4.10 for Knapp Mill and Throop monitoring sites respectively. For nitrate, the similar switch-point in time for both rivers indicate the occurrence about week 12, the early July 2013. However, the almost full-time range of 95%CI (KM:(1.41,51.48), TR:(1.37,51.36)) refers to not dividing it clearly. This can also be seen in a wide range of change threshold in flow. The overlap of change thresholds in flow before and after that switch-point shows a lack of significance in different of the flow levels, though the median indicates that. This may suggest an incompatibility of change-point structure to the data corresponding with the model selection result on Nitrate concentration as discussed. However, at the fist stage to investigate the bivariate model performance, all combining modes will be constructed with the same model.

TABLE 4.9: Parameter estimations for model M1 of Nitrate and Phosphate concentrations (Knapp Mill)

	Nitrate		**Phosphate**	
Parameters	Median	95% CI	Median	95% CI
Change-point structures				
τ (Switch-point in time)	12.01	(1.406,51.48)	7.088	(2.225,40.0)
φ_1 (Change threshold in flow before τ)	2.042	(1.741,3.819)	2.684	(1.9,4.021)
φ_2 (Change threshold in flow after τ)	4.492	(2.628,4.624)	3.582	(1.756,4.612)
δ_1 (slope for low flow before τ)	0.115	(-17.43,17.82)	-0.057	(-15.67,16.18)
δ_2 (slope for high flow before τ)	-	-	-5.649	(-13.45,3.685)
δ_3 (slope for low flow after τ)	-	-	-0.389	(-9.319,9.547)
δ_4 (slope for high flow after τ)	-0.075	(-17.44,17.41)	-0.370	(-14.3,9.967)
Penalised splines				
β_0 (Global intercept)	1.724	(1.604,1.855)	-2.967	(-4.194,-2.136)
β_1 (Fixed effect for temperature)	-0.178	(-0.409,0.070)	0.340	(-0.278,0.974)
β_2 (Fixed effect for conductivity)	0.274	(0.029,0.551)	-0.319	(-0.910,0.371)
β_3 (Fixed effect for dissolved oxygen)	0.006	(-0.198,0.188)	-0.078	(-0.432,0.298)
β_4 (Fixed effect for turbidity)	0.090	(-0.099,0.294)	0.366	(-0.142,0.891)
σ_{b1} (SD for spline on temperature)	0.030	(0.015,0.087)	0.035	(0.016,0.116)
σ_{b2} (SD for spline on conductivity)	0.030	(0.015,0.090)	0.035	(0.016,0.134)
σ_{b3} (SD for spline on dissolved oxygen)	0.027	(0.014,0.075)	0.033	(0.016,0.098)
σ_{b4} (SD for spline on turbidity)	0.029	(0.015,0.086)	0.031	(0.016,0.104)
Other				
σ^2 (Measurement error variance)	0.049	(0.033,0.076)	0.357	(0.237,0.570)

TABLE 4.10: Parameter estimations for model M1 of Nitrate and Phosphate concentrations (Throop)

	Nitrate		**Phosphate**	
Parameters	Median	95% CI	Median	95% CI
Change-point structures				
τ (Switch-point in time)	11.65	(1.368,51.36)	2.685	(2.007,11.73)
φ_1 (Change threshold in flow before τ)	1.295	(0.894,3.924)	2.221	(1.057,4.557)
φ_2 (Change threshold in flow after τ)	4.728	(2.155,4.928)	3.803	(0.975,4.885)
δ_1 (slope for low flow before τ)	0.063	(-16.91,17.24)	-0.056	(-17.04,16.01)
δ_2 (slope for high flow before τ)	-	-	-1.586	(-16.11,15.07)
δ_3 (slope for low flow after τ)	-	-	-0.681	(-4.137,-0.433)
δ_4 (slope for high flow after τ)	-0.011	(-17.99,17.74)	0.056	(-4.655,13.94)
Penalised splines				
β_0 (Global intercept)	1.978	(1.863,2.104)	-2.278	(-3.057 ,-0.549)
β_1 (Fixed effect for temperature)	-0.120	(-0.411,0.157)	0.103	(-0.299,0.531)
β_2 (Fixed effect for conductivity)	0.180	(0.120,0.482)	-0.243	(-0.688,0.202)
β_3 (Fixed effect for dissolved oxygen)	0.065	(-0.128,0.269)	-0.157	(-0.510,0.142)
β_4 (Fixed effect for turbidity)	0.039	(-0.051,0.130)	0.075	(-0.056,0.208)
σ_{b1} (SD for spline on temperature)	0.030	(0.015,0.087)	0.032	(0.015,0.098)
σ_{b2} (SD for spline on conductivity)	0.030	(0.015,0.088)	0.032	(0.016,0.096)
σ_{b3} (SD for spline on dissolved oxygen)	0.027	(0.014,0.076)	0.030	(0.015,0.095)
σ_{b4} (SD for spline on turbidity)	0.038	(0.016,0.200)	0.038	(0.016,0.200)
Other				
σ^2 (Measurement error variance)	0.053	(0.036,0.082)	0.110	(0.074,0.177)

Table 4.11 shows the posterior median estimates and 95%CI for the catchment area normalised annual total nitrate and phosphate fluxes in Kg/Km^2 for the year from the last week of April 2013 to the early week of April 2014. These estimates are similar on all considering models. It indicate the most macronutrient loads are from the river Stour with almost twice time for nitrate and 5 times for phosphate. The last four rows each sections of the table 4.11, provide estimate of annual budget for the Hampshire Avon at Knapp Mill reported by Pirani et al. (2016) using the same structure of model M1 and M2, the estimate of mean annual fluxes for the Hampshire Avon at Knapp Mill reported by Jarvie et al. (2005), and the UK wide average reported by Nedwell et al. (2002). However, for macronutrients at Knapp Mill, the estimates is lower than Pirani et al. (2016) which estimated for a very unusual year with exceptionally high rainfall. Moreover, the level of nitrate keep raising with addition effect of uncertainty storm events

while phosphate is decreasing. This may indicate the efficiency of water management plan for the last two decades. Unfortunately, the more reviews of phosphate study is required for better understand its annual load.

TABLE 4.11: Posterior median and 95% credible interval (CI) for the total annual macronutrient fluxes during 26 April 2013 and 10 April 2014. Values are catchment area standardised with Kg/Km^2 including comparable estimates from the literature

Models	Nitrate		Phosphate	
	Annual budget	**95% CI**	**Annual budget**	**95% CI**
Knapp Mill gauging station on the River Hampshire Avon				
M1	2853.0	(2640.0,3096.0)	19.12	(15.26,24.82)
M2	2865.0	(2656.0,3094.0)	19.21	(14.61,25.35)
M3	2864.0	(2652.0,3094.0)	19.66	(15.42,24.96)
M5	2870.0	(2635.0,3120.0)	19.16	(14.24,27.42)
M6	2854.0	(2639.0,3084.0)	20.6	(15.35,28.0)
Pirani et al. (2016) (M1) 2013-14	2978.9	(2937.9,3016.4)	31.6	(30.2,33.1)
Pirani et al. (2016) (M2) 2013-14	2936.7	(2890.4,2981.8)	29.8	(28.3,31.3)
Jarvie et al. (2005): 1993-2000	2050	-	71	-
Nedwell et al. (2002): 1995-96	1400	-	152	-
Throop gauging station on the River Stour				
M1	4342.0	(3979.0,4753.0)	109.9	(97.39,125.2)
M2	4373.0	(4020.0,4755.0)	124.5	(103.4,149.8)
M3	4374.0	(4016.0,4766.0)	116.8	(97.68,139.6)
M5	4405.0	(4007.0,4840.0)	109.9	(96.91,124.5)
M6	4351.0	(3987.0,4751.0)	124.0	(103.0,149.8)

4.3.3.2 Macronutrient Bivariate Model

An our aim is to estimate the annual budget of nitrate and phosphate into the Christchurch Harbour estuary. The bivariate model offers a chance to calculate such macronutrient loads from the Hampshire Avon and the Stour river; also construct 95% CI of estimates. It also provides each river estimation. Parameter estimates of a sample model M1 are presented in Table 4.12 and 4.13 for Knapp Mill and Throop monitoring sites respectively. For nitrate, the similar switch-point in time for both rivers indicate the occurrence about week 31.91, the early July 2013. However, the almost full-time range of 95%CI (KM:(1.76,51.51), TR:(1.308,51.35)) refers to not dividing it clearly. This can also be seen in a wide range of change threshold in flow. The overlap of change thresholds in flow before and after that switch-point shows a lack of significance in different of

the flow levels, though the median indicates that. This may suggest an incompatibility of change-point structure to the data corresponding with the model selection result on Nitrate concentration as discussed. However, at the fist stage to investigate the bivariate model performance, all combining modes will be constructed with the same model.

TABLE 4.12: Parameter estimations for model M1 of Nitrate concentrations

	Knapp Mill		**Throop**	
Parameters	Median	95% CI	Median	95% CI
Change-point structures				
τ (Switch-point in time)	31.91	(1.76,51.51)	10.85	(1.38,51.38)
φ_1 (Change threshold in flow before τ)	2.227	(1.743,4.313)	1.341	(0.892,4.024)
φ_2 (Change threshold in flow after τ)	4.47	(2.898,4.625)	4.779	(1.957,4.93)
δ_1 (slope for low flow before τ)	0.128	(-14.95,14.73)	0.050	(-16.66,17.24)
δ_2 (slope for high flow before τ in time)	-	-		(,)
δ_3 (slope for low flow after τ in time)	-	-		(,)
δ_4 (slope for high flow after τ)	-0.257	(-18.25,17.49)	-0.004	(-18.37,18.25)
Penalised splines				
β_0 (Global intercept)	1.702	(1.762,1.884)	1.987	(1.932,2.07)
β_1 (Fixed effect for temperature)	-0.136	(-0.249,0.008)	-0.145	(-0.325,0.013)
β_2 (Fixed effect for conductivity)	0.356	(0.211,0.481)	0.173	(0.013,0.336)
β_3 (Fixed effect for dissolved oxygen)	-0.054	(-0.128,0.023)	0.084	(-0.026,0.194)
β_4 (Fixed effect for turbidity)	0.211	(0.094,0.330)	0.038	(0.0016,0.078)
σ_{b1} (SD for spline on temperature)	0.0289	(0.0149,0.080)	0.0303	(0.0153,0.0854)
σ_{b2} (SD for spline on conductivity)	0.0365	(0.0176,0.1011)	0.0281	(0.0147,0.0778)
σ_{b3} (SD for spline on dissolved oxygen)	0.0267	(0.0144,0.0708)	0.0269	(0.0145,0.0736)
σ_{b4} (SD for spline on turbidity)	0.0417	(0.0189,0.1197)	0.0382	(.0165,0.2085)
Bivariate covariance matrix				
$\sigma^2_{11}, \sigma^2_{22}$ (within river)	0.002	(0,0.003)	0.010	(0.007,0.016)
$\sigma^2_{12}, \sigma^2_{21}$ (between river)	0.001 (0,0.004)			

TABLE 4.13: Parameter estimations for model M1 of Nitrate and Phosphate concentrations (Throop)

	Nitrate		**Phosphate**	
Parameters	Median	95% CI	Median	95% CI
Change-point structures				
τ (Switch-point in time)	3.708	(2.162,39.37)	2.628	(2.017,6.976)
φ_1 (Change threshold in flow before τ)	2.629	(1.826,4.389)	2.248	(1.158,4.279)
φ_2 (Change threshold in flow after τ)	3.795	(1.756,4.615)	4.059	(1.183,4.812)
δ_1 (slope for low flow before τ)	0.603	(-17.4,17.55)	-1.095	(-20.04,14.62)
δ_2 (slope for high flow before τ)	-4.011	(-13.56,10.26)	1.396	(-15.53,15.3)
δ_3 (slope for low flow after τ)	-0.329	(-8.906,8.548)	-0.612	(-1.522,-0.453)
δ_4 (slope for high flow after τ)	-0.505	(-13.91,11.29)	0.383	(-0.573,13.82)
Penalised splines				
β_0 (Global intercept)	-3.027	(-4.048,-2.192)	-2.362	(-2.919,-0.792)
β_1 (Fixed effect for temperature)	0.050	(-0.56,0.682,)	0.001	(-0.317,0.324)
β_2 (Fixed effect for conductivity)	0.074	(-0.602,0.978)	-0.134	(-0.462,0.201)
β_3 (Fixed effect for dissolved oxygen)	-0.138	(-0.513,0.185)	-0.215	(-0.478,0.006)
β_4 (Fixed effect for turbidity)	0.365	(-0.186,0.837)	0.081	(-0.015,0.174)
σ_{b1} (SD for spline on temperature)	0.035	(0.016,0.116)	0.031	(0.015,0.092)
σ_{b2} (SD for spline on conductivity)	0.049	(0.018,0.22)	0.031	(0.015,0.093)
σ_{b3} (SD for spline on dissolved oxygen)	0.032	(0.016, 0.098)	0.030	(0.015,0.087)
σ_{b4} (SD for spline on turbidity)	0.032	(0.016,0.112)	0.038	(0.016,0.200)
Bivariate covariance matrix				
$\sigma^2_{11}, \sigma^2_{22}$ (within river)	0.347	(0.225,0.575)	0.067	(0.044,0.111)
$\sigma^2_{12}, \sigma^2_{21}$ (between river)	0.093 (0.041,0.179)			

Table 4.14 shows the posterior median estimates and 95%CI of annual total nitrate and phosphate loads in the Christchurch Harbour estuary for the year from the last week of April 2013 to the early week of April 2014. Values are normalised by the catchment area Kg/Km^2 for comparison. It also provides load estimates by Pirani et al. (2016).

TABLE 4.14: Posterior median and 95% credible interval (CI) for the annual total macronutrient loads based on the bivariate normal model during 26 April 2013 and 10 April 2014 (normalised by the catchment area Kg/Km^2)

Models	**Nitrate**		**Phosphate**	
	Annual budget	**95% CI**	**Annual budget**	**95% CI**
Knapp Mill gauging station on the River Hampshire Avon				
M1	2835.0	(2772.0,2897.0)	18.68	(15.07,24.1)
Pirani et al. (2016) (M1) 2013-14	2978.9	(2937.9,3016.4)	31.6	(30.2,33.1)
Pirani et al. (2016) (M2) 2013-14	2936.7	(2890.4,2981.8)	29.8	(28.3,31.3)
Throop gauging station on the River Stour				
M1	4327.0	(4161.0,4506.0)	108.4	(98.66,119.5)
Total				
M1	7162.0	(6971.0,7365.0)	127.2	(114.8,141.9)

Note: M1: Penalised spline for water quality data with change-point structures

Chapter 5

Precipitation solute modelling: the Hubbard Brook

5.1 Introduction

Hubbard brook experimental forest was established in 1955 to study the effect of forests on the water cycle and water quality. Under the collaboration between the Forest Service and external researchers. The study was conducted in the small watersheds since the early 1960s for hydrological research. Several monitoring stations were deployed to collect data about precipitation, weather and stream flow in watersheds. The data was continually collected more than 50 years since 1963 as long-term data on watershed study.

Raising of acid rain in the US before the 1970s, air pollution from industry is a major source of acid compounds as sulfuric and nitric acid in precipitation. The fallen acid rain on forests and freshwater directly disturbs habitat, water quality and ecosystem functioning. It leads to illness and premature death of plants, trees, fishes, and birds; following by the difficulties of human living and economy from acidity water.

Hence, it is important to keep monitoring on precipitation chemistry to evaluate the acidity effects on forests and fresh water. The value of long-term observation has been claimed as a part of fact-based decision-making in environmental policy. (Sullivan et al. 2018)

However, the solute compounds are vastly varying from weather to weather, storm to storm, and region to region. Thus, there are some sources of variation of solute compounds such as season and storms in modelling consideration. A statistical method, Bayesian approach, will be applied to handle this uncertainty and complexity through Markov chain Monte Carlo (MCMC) methods. Moreover, it is also handle some data problems such as missing values with probability statement.

5.1.1 Purpose of analysis

We aim to construct a model of precipitation solute concentration and estimate the annual solute loads from precipitation within the Hubbard Brook experimental forest. We use the 11-water year of weekly data from the three rain gauges: RG-11, RG-22, and RG-23 during June 2002 to May 2013. We are interested in quantifying the concentration of three solutes in precipitation: Calcium (Ca), Sulphate (SO4), and Nitrate (NO3). In particular, sulphate and nitrate are indicators of acidity rainfalls as pollutants.

A preliminary study of precipitation data will be performed to understand the data characteristics. Data treatment and preparation are also provided for modelling purpose. Under Bayesian approach, we are first considering a regression model of each solute concentrations with precipitation volume as an only regressor. Alternatively, methods of time series model can also be applied to solute concentrations. In addition, a combined information model using the multivariate normal distribution is constructing and evaluating its performance as well.

In addition, the annual solute loads will be estimated and will be calculated the 95% credible interval (CI) of estimates using the selected model.

5.2 Statistical Modelling Precipitation Solutes

In this chapter, we are examining the pattern of precipitation chemistry time series. Weekly samples were collected in HBES project and analysed for solute quantities. The recent 574-week time series of precipitation chemistry, from June 2002 to May 2016 are used in our study. We will construct precipitation solute concentrations using Bayesian approach in univariate and multivariate viewpoints. Then the annual solute loads will be calculated based on the selected Bayesian model.

5.2.1 Precipitation Solute Univariate Model

Variations of precipitation solute concentrations are naturally driven by acid rain and snow. As discussed, we then study univariate models to describe precipitation solute concentrations using time series models under Bayesian framework to cover uncertainties event.

Given y_{kjt} denote the quantity of modelling precipitation solute at rain gauge $k \in \{1{=}\text{RG-11}, 2{=}\text{RG-22}, 3{=}\text{RG-23}\}$, particular solute $j \in \{1{=}\text{Calcium}, 2{=}\text{Sulphate}, 3{=}\text{Nitrate}, 4{=}\text{Ammonium}, 5{=}\text{Potassium}, 6{=}\text{Sodium}, 7{=}\text{Chloride}\}$, at week $t \in \{1, 2, ..., T = 574\}$. We are considering y_{kjt} in natural logarithm scale as discussed.

A Bayesian hierarchical model for solute concentration y_{kjt} is used to construct the model with independent Gaussian model assumption. Then, the mean of precipitation solute concentrations will be described. The first hierarchical described by

$$y_{kjt} \sim N(\mu_{kjt}, \sigma^2_{kj})$$

where μ_{kjt} denotes the mean concentration for rain gauge k and solute j at week t. The variance σ^2_{kj} of rain gauge k and solute j are assumed to be constant over time. The second hierarchy can be written in terms of a time series function and the error term ϵ_{kjt} as

$$\mu_{kjt} = g_{kj}(\mathbf{y}) + \epsilon_{kjt},$$

$$\epsilon_{kjt} \sim N(0, \sigma_k j^2)$$

where the mean function $g_{kj}(\mathbf{y})$ will be described as follows. The errors of unobserved effect are assumed to be the independent identically normally distributed random error with mean zero and constant variance $\sigma_k j^2$

To model solute concentrations, some time-varying techniques are applying to model the time series. The mean function $g_{kj}(\mathbf{y})$ are described as below:

5.2.1.1 Linear regression model (LM)

This is the simplest model which describing time series by only the intercept then the predicted values at any week t are all the same as the mean model. Thus, the mean function can be written as

$$g_{kj}(\mathbf{y}) = \mu_{kj}, \quad t = 1, 2, ..., 574 \tag{5.1}$$

The concentration for each rain gauge each precipitation solute is assumed to be a constant over time as an average of that concentration.

5.2.1.2 Linear regression and autocorrelated error (LMARE)

It is common for time series regression that the error may have a time series structure or there is a correlation between the error at the different times. Thus, the model may be modified to have autocorrelated errors. In order to remove this temporal dependence such as trend and seasonality, a lag difference parameter p is defined as the lag of subtraction between consecutive errors. The more p, the more complex model is structured. It also can be defined by seasonal period. Based on investigating, the first-order autoregressive AR(1) or defining $p = 1$ is suitable for this study. Since the error term follows

an AR(1) process, then

$$\epsilon_{kjt} = \phi_{kj}\epsilon_{kj(t-1)} + u_{kjt}$$

where ϕ_{kj} is a autocorrelation parameter such that $|\phi_{kj}| < 1$ and u_{kjt} are independent $N(0, \sigma_k j^2)$

Thus, the mean function can be written as

$$g_{kj}(\mathbf{y}) = \mu_{kj} + \phi_{kj}\epsilon_{kj(t-1)}. \tag{5.2}$$

The error term still has mean zero and constant variance will be affected by the autocorrelation parameter and derived as

$$\Psi_k j^2 = \frac{\sigma_k^2 j}{1 - \phi_k j^2}$$

Hence, the error term can be rewritten as

$$\epsilon_{kjt} = y_{kjt} - \mu_{kj} - \phi_{kj}\epsilon_{kj(t-1)} \tag{5.3}$$

with the first error term $\epsilon_{kj1} = y_{kj1} - \mu_{kj}$.

5.2.1.3 The random walk time series model (RdW)

This is a particular model which is identical to the AR(1) model with $\phi_{kj} = 1$. In other words, it is equally likely to change (higher or lower) to the next value (step) regarding the previous value or the past. Each time, the variable takes a random step to the next value with independent and identically distributed on step size. Thus, the change from the last value (i.e., $y_{kjt} - y_{kj(t-1)}$) can be considered as the first difference of the sequence. Applying this to model 5.2.1.2, the mean function will be rewritten as

$$g_{kj}(\mathbf{y}) = \mu_{kj} + \eta_{kj(t-1)} \tag{5.4}$$

This model is called the random walk with "drift" or step sizes distribution has a non-zero mean (i.e., $\mu_k j$). In addition of extension to the AR(1), the model becomes the mean model if $\phi_k j$ is small, $\phi_k j \to 0$.

5.2.1.4 The first-order autoregressive model (AR1)

Alternatively to model the error, it may be considered in the level that process the model error. In the other word, we define uncertainties of response values directly using a model e.g. a linear combination of p past values or order p. This model is handling a variation of different time series patterns, while the variance of the error term of

model 5.2.1.2 handle scale of time series. Autocorrelation parameters ϕ_q; $q \in 1, 2, ..., p$ refer to coefficients or weights of that p consecutive past values, $|\phi_q| < 1$. Hence, an autoregressive model of order $p = 1$ (AR(1) model) can be written as

$$y_{kjt} = \mu_{kj} + \eta_{kjt} + \epsilon_{kjt}$$
$$\eta_{kjt} = \phi_{kj}\eta_{kj(t-1)} + \xi_{kjt}$$

where $\epsilon_{kjt} \sim N(0, \sigma_k^2 j)$ and $\xi_{kjt} \sim N(0, \sigma_{\eta_{kj}}^2)$ independently. Hence, the mean function will be defined as

$$g_{kj}(\mathbf{y}) = \mu_{kj} + \phi_{kj}\eta_{kj(t-1)} \tag{5.5}$$

This model provides a less complex structure on recursive y_{kjt} term and probable estimating on missing value (y_{kjt}).

5.2.1.5 Linear regression and monthly indexes model (SLM)

According to an evidence of seasonality and classical time series decomposition method, we may include a seasonal component that use time indices as explanatory (dummy) variables to an ordinary regression model. In this study, there seems to be a repeat seasonal pattern (monthly in consideration) each water year. In fact that climatic variation is a common behaviour of ecological data. Thus, the mean function 5.1 will be rewritten as

$$g_{kj}(\mathbf{y}) = \mu_{kj} + \sum_{m=1}^{11} \delta_{mkj} * M_{mkjt} \tag{5.6}$$

where M_{mkjt} are monthly dummy variables (0 or 1) with associated coefficients δ_{mkj}. The variable M_m takes the value 1 in month m and zero elsewhere. The dummy coefficients δ_{mkj} show the concentration in the given month relative to December as the base month.

5.2.1.6 The first-order autoregressive and seasonal indexes model (SAR1)

This model extends the AR(1) model in 5.2.1.4 with monthly indexes to handle the seasonal variation. Hence, the mean function will be defined as

$$g_{kj}(\mathbf{y}) = \mu_{kj} + \sum_{m=1}^{11} \delta_{mkj} * M_{mkjt} + \phi_{kj}\eta_{kj(t-1)} \tag{5.7}$$

5.2.1.7 The first-order autoregressive and yearly indexes model (yAR1)

This model extends the AR(1) model in 5.2.1.4 with yearly indexes to handle the yearly variation. Hence, the mean function will be defined as

$$g_{kj}(\mathbf{y}) = \mu_{kj} + \sum_{w=1}^{11} \delta_{wkj} * W_{wkjt} + \phi_{kj}\eta_{kj(t-1)} \tag{5.8}$$

5.2.1.8 The first-order autoregressive and sinusoidal model (SNAR1)

This model extends the AR(1) model in 5.2.1.4 with sinusoidal function to capture temporal oscillations in time series. Sinusoidal model can be fit using nonlinear least squares to obtain a good fit.

To deal with seasonal in time series, the fluctuation of series can be represented by a wave form, sinusoidal function. Thus a model composing with seasonal harmonic can take the form of

$$y_t = \sum_{h=0}^{H} \{\alpha_h cos(\omega_h t) + \beta_h sin(\omega_h t)\} + e_t$$

where H represents number of harmonic terms, larger number is better keeping track the fluctuation. Hence, the mean function can be defined as

$$g_{kj}(\mathbf{y}) = \mu_{kj} + \sum_{h=0}^{H} \{\alpha_h cos(\omega_h t) + \beta_h sin(\omega_h t)\} \tag{5.9}$$

5.2.2 Precipitation Solute Multivariate Model

In addition, there appears to be a moderate correlation among the concentrations of precipitation solute. Thus, we are interested in combining those data under multivariate normal distribution. It may provide the better estimates for the missing values and total annual load calculation.

As a result, we are combining each chemistry univariate models into a Bayesian multivariate normal model. In this study, we are constructing a multivariate normal model to combine the same univariate models for all gauges all solute concentrations (see section 5.2.1). In order to examine the Bayesian multivariate normal model, we are consider to combine with three view points by: rain gauge, solutes, and combination of rain gauge and solutes, due to moderate correlation among them,

First, let the vector of observations each solute j at week t, $\boldsymbol{y_{jt}} = (y_{kjt}, k = 1, ..., 3)'$ which rain gauge $k \in \{1$=RG-11, 2=RG-22, 3=RG-23$\}$. If we assume solute concentrations are multivariate normal distributed with mean vector $\boldsymbol{\mu_{jt}}$ and variance-covariance

matrix Σ. Bayesian multivariate normal model can be written as

$$\begin{aligned}
&\boldsymbol{y_{jt}} \overset{iid}{\sim} N_3\left(\boldsymbol{\mu_{jt}}, \Sigma\right), \\
&\boldsymbol{\mu_{jt}} = \begin{pmatrix} \mu_{1jt} \\ \mu_{2jt} \\ \mu_{3jt} \end{pmatrix}, \Sigma = \begin{bmatrix} \sigma_{1j}^2 & \sigma_{(12)j} & \sigma_{(13)j} \\ \sigma_{(21)j} & \sigma_{2j}^2 & \sigma_{(23)j} \\ \sigma_{(31)j} & \sigma_{(32)j} & \sigma_{3j}^2 \end{bmatrix} \\
&\epsilon_{jt} \sim N(0, \Sigma)
\end{aligned}$$

where σ_{kj}^2 is the variance of $y_{kj.}$. The covariance is $\sigma_{(uv)j} = \sigma_{(vu)j} = \rho_{(uv)j}\sigma_u j \sigma_v j$, $|\rho_{(uv)j}| < 1$ and $u, v \in k$. This model is combining all gauge models for specific solutes.

Second, combining all solutes from each gauges, let the vector of observations each rain gauge k at week t, $\boldsymbol{y_{kt}} = (y_{kjt}, j = 1, ..., 7)'$ which solute $j \in \{1$=Calcium, 2=Sulphate, 3=Nitrate, 4=Ammonium, 5=Potassium, 6=Sodium, 7=Chloride$\}$. Assume solute concentrations are multivariate normal distributed with mean vector $\boldsymbol{\mu_{kt}}$ and variance-covariance matrix Σ. Bayesian multivariate normal model can be described as

$$\begin{aligned}
&\boldsymbol{y_{kt}} \overset{iid}{\sim} N_7\left(\boldsymbol{\mu_{kt}}, \Sigma\right), \\
&\boldsymbol{\mu_{kt}} = \begin{pmatrix} \mu_{k1t} \\ \vdots \\ \mu_{k7t} \end{pmatrix}, \Sigma = \begin{bmatrix} \sigma_{k1}^2 & \cdots & \sigma_{k(17)} \\ \vdots & \ddots & \vdots \\ \sigma_{k(71)} & \cdots & \sigma_{k7}^2 \end{bmatrix} \\
&\epsilon_{kt} \sim N(0, \Sigma)
\end{aligned}$$

where σ_{kj}^2 is the variance of $y_{kj.}$. The covariance is $\sigma_{k(uv)} = \sigma_{k(vu)} = \rho_{k(uv)}\sigma_k u \sigma_k v$, $|\rho_{k(uv)}| < 1$ and $u, v \in j$.

Last, combining all gauges all solutes, let the vector of observations each rain gauge k at week t, $\boldsymbol{y_{kjtt}} = (y_{kjt}, k = 1, ...3, j = 1, ..., 7)'$ which rain gauge $k \in \{1$=RG-11, 2=RG-22, 3=RG-23$\}$ and solute $j \in \{1$=Calcium, 2=Sulphate, 3=Nitrate, 4=Ammonium, 5=Potassium, 6=Sodium, 7=Chloride$\}$. Assume solute concentrations are multivariate normal distributed with mean vector $\boldsymbol{\mu_{kjt}}$ and variance-covariance matrix Σ. Bayesian

multivariate normal model can be written as

$$\boldsymbol{y_{kjt}} \overset{iid}{\sim} N_{(3\times 7)}\left(\boldsymbol{\mu_{kjt}}, \Sigma\right),$$

$$\boldsymbol{\mu_{kjt}} = \begin{pmatrix} \mu_{k1t} \\ \vdots \\ \mu_{k7t} \end{pmatrix}, \Sigma = \begin{bmatrix} \sigma^2_{1,1} & \sigma^2_{2,1} & \cdots & \sigma_{21,1} \\ \sigma_{1,2} & \sigma^2_{2,2} & \cdots & \sigma_{21,2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,21} & \sigma_{2,21} & \cdots & \sigma^2_{21,21} \end{bmatrix}$$

$$\epsilon_t \sim N(0, \Sigma)$$

where $\sigma^2_{a,a}$ is the variance of y of combination of k rain gauge and solutes j y_{kj}. The covariance is $\sigma_{a,b} = \sigma_{b,a} = \rho_{ab}\sigma_a\sigma_b$, $|\rho_{ab)}| < 1$, $a, b = 1, ..., 3 \times 7$, $a \neq b$.

By view points, The element of $\boldsymbol{\mu}$ will be defined with mean function as described in section 5.2.1. However, we are using the best function for overall subset within the same view point (each gauge, each solute, or each combination of gauge and solute) in modelling. This is a simple approach and more practicable than using the best mean function each subset. Moreover, there will be investigate the Bayesian multivariate normal model performance with different number of component under the same multivariate model, for example, varying the number of gauges, the number of solutes, or both.

In addition, the autoregressive models (AR(1)) and its modification in 5.2.1.6, 5.2.1.7 and 5.2.1.8 will be considered as independent on $y_{j.}$ or the covariance is zero but dependent on $\eta_{j.}$ with multivariate normally distributed instead.

The exploratory result shows the moderate correlation among precipitation solute concentrations. The multivariate model may improve the estimation of missing values of solute concentrations using the relative information from other solutes.

5.2.3 Solute load calculation

The annual loads can be calculated by product of each solute concentration (mg/l) and the amount of precipitation (mm), them roll over for a period. Based on Bayesian modelling, missing solute concentrations will be estimated for load calculation. The posterior mean of solute concentration represents weekly concentrations then be multiplied by weekly precipitation, then sum over each water year. (June to May next year, e.g., a water year 2002 starts from 1 June 2002 to 31 May 2003). It also provide the 95% credible interval of the annual solute loads.

$$L_y = \sum_{w=1}^{w_y} (C_{yw} * P_{yw}) \tag{5.10}$$

where L_y denotes solute loads of water year $y \in \{2002, ..., 2012\}$, C_{yw} denotes the posterior mean of weekly solute concentration for week w of water year y where $w \in 1, 2, ..., w_y$ which w_y is the last week of water year y. P_{yw} denotes weekly precipitation for week w of water year y. The annual loads unit is $grams(g)$.

5.2.4 Bayesian computation

In order to derive corresponding posterior distribution of model parameters, we use Markov chain Monte Carlo (MCMC) methods, Gibbs sampler in particular, to obtain MCMC samples providing posterior statistics e.g. mean, median, quantiles, and 95% credible intervals.

Models and facilities are coded in the language R (version 3.4.2), then scripting with JAGS using R2jags package (version 0.5-7; https://CRAN.R-project.org/package=R2jags) that allows to run JAGS (Just Another Gibbs Sampler by Martyn Plummer) models within R providing the similar functionality as the R2WinBUGS package but faster in practical. In order to obtain Bayesian analysis and modelling. MCMC procedure is setting with the chain thinning every 10^{th} interation, 3,000 burn-ins, and 7,000 iterations that provides a good convergence on model parameter estimates.

5.3 Results

5.3.1 Results for Fitting Univariate Models

For the univariate case, we fitted models on each seven solutes each three rain gauges independently. These are evaluated on predictive performance with predictive model choice criteria (PMCC) which models with the lower PMCC are more preferable. Figure 5.1 illustrates PMCC of all fitted models for simple comparison. Table 5.1 provides PMCC and its partition of RG-11 fitted models. The lowest PMCC of the fitted models for each solute each Rain gauge are marked in bold.

Considering LM model and extened models,i.e. LMARE (with AR1 error) and SLM (with monthly indexes), the lower PMCC on SLM model implies the effect of the monthly index on capturing seasonal fluctuations. Turning into AR1 model emphasising the immediately prior value in time, a much lower PMCC value compared with LM model may verify that relevance. This interrelation is noticeable and usually founded in ecological data.

As a consequence, AR1 model is preferable to modelling, then we are considering on some modifications of the autoregressive model: Rdw (with $\phi = 1$), SAR1 (with monthly indexes), yAR1 (with yearly indexes), and SNAR1 (with h sinusoidal components). Note

that additional sinusoidal components are sets of sine and cosine functions to capture time series harmonic with known period (52 weeks a year approximately for weekly data). The number of sinusoidal terms ($h > 1$) refers the more complex model, the more harmonic captured on time series. In practice, we varied the number of sinusoidal terms from 1 to 4, but the PMCCs are similar. Therefore, we consider only SNAR1 with h=1, or AR1 with one additional term of sine and cosine function, the simplest sinusoidal model.

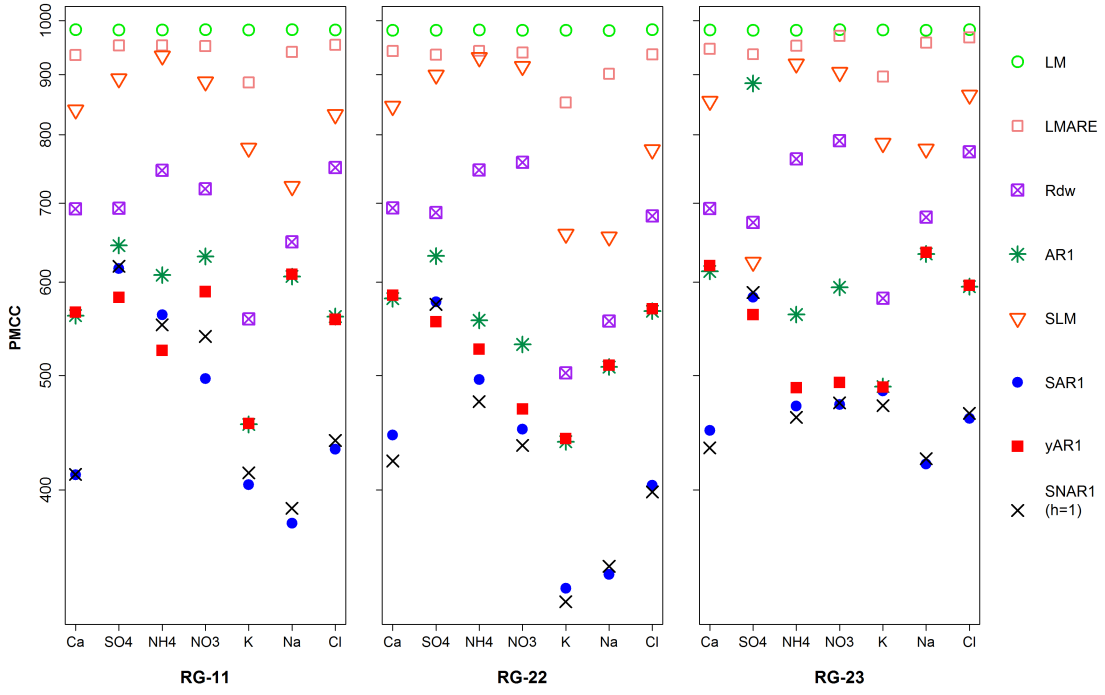


FIGURE 5.1: Predictive model choice criterion: PMCC of **univariate** models of seven solutes from all rain gauges

TABLE 5.1: Predictive model choice criterion: PMCC(G+P) of **univariate** models of seven solutes from **RG-11** ; partitioning into goodness of fit (G), penalty (P). (The lowest PMCC of fitted models for each solute is presented in bold.)

Solute/model	G	P	PMCC
Calcium (Ca)			
LM	490.49	491.94	982.43
LMARE	462.57	472.18	934.75
SLM	409.57	430.34	839.91
AR1	182.39	379.64	562.03
Rdw	280.64	411.81	692.45
SAR1	98.12	313.73	**411.85**
yAR1	185.15	380.78	565.93
SNAR1($h = 1$)	101.38	310.82	412.20

Continued on next page

Table 5.1 – continued

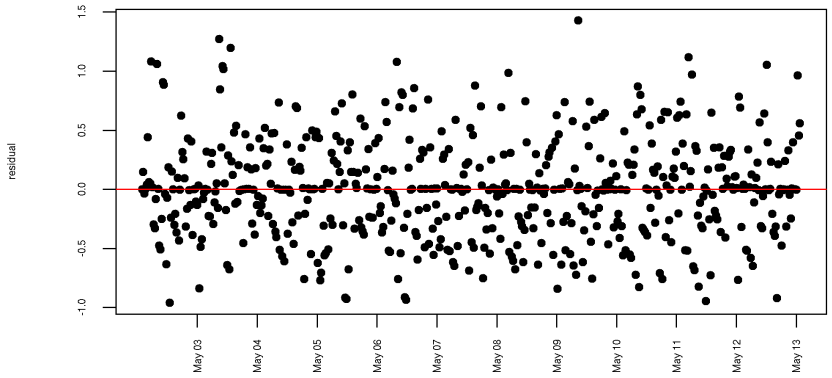
Solute/model	G	P	PMCC
Sulphate (SO4)			
LM	490.54	491.69	982.23
LMARE	473.31	479.33	952.64
SLM	434.98	457.76	892.74
AR1	244.51	400.06	644.57
Rdw	287.59	405.49	693.08
SAR1	221.01	395.08	616.09
yAR1	198.94	383.61	**582.57**
SNAR1($h = 1$)	224.30	394.74	619.04
Ammonium (NH4)			
LM	489.62	492.29	981.91
LMARE	472.84	479.60	952.44
SLM	455.39	477.91	933.30
AR1	205.23	403.20	608.43
Rdw	312.53	434.04	746.57
SAR1	173.14	389.73	562.87
yAR1	152.69	372.36	**525.05**
SNAR1($h = 1$)	168.47	383.68	552.15
Nitrate (NO3)			
LM	490.01	492.30	982.31
LMARE	471.41	480.16	951.57
SLM	432.82	454.60	887.42
AR1	225.14	405.85	630.99
Rdw	296.66	423.14	719.80
SAR1	139.66	357.29	**496.95**
yAR1	195.37	393.62	588.99
SNAR1($h = 1$)	165.26	374.49	539.75
Potassium (K)			
LM	489.62	492.16	981.78
LMARE	437.92	448.06	885.98
SLM	380.11	399.48	779.59
AR1	139.19	315.34	454.53
Rdw	211.39	346.94	558.33
SAR1	109.90	294.10	**404.00**
yAR1	139.63	315.82	455.45
SNAR1($h = 1$)	117.22	296.33	413.55
Sodium (Na)			
LM	490.50	492.26	982.76
LMARE	465.81	474.54	940.35
SLM	352.42	370.88	723.30

Continued on next page

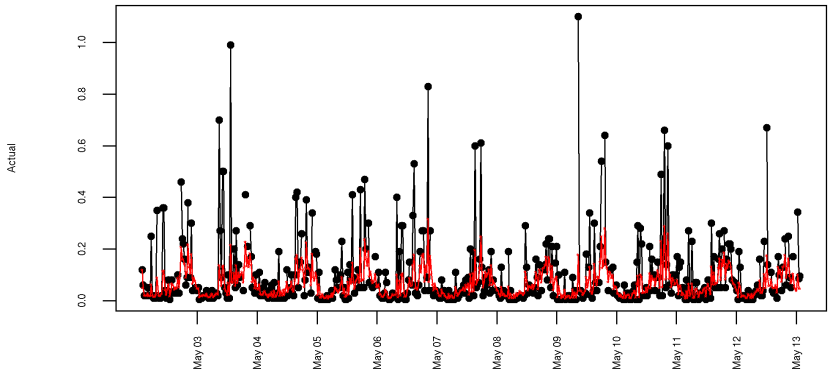
Table 5.1 – continued

Solute/model	G	P	PMCC
AR1	224.82	381.90	606.72
Rdw	262.00	387.27	649.27
SAR1	93.23	281.43	**374.66**
yAR1	226.62	382.72	609.34
SNAR1($h = 1$)	99.24	286.65	385.89
Chlorine (Cl)			
LM	489.99	491.89	981.88
LMARE	472.31	481.20	953.51
SLM	405.39	426.82	832.21
AR1	170.71	390.35	561.06
Rdw	310.08	440.60	750.68
SAR1	108.34	324.73	**433.07**
yAR1	168.79	389.14	557.93
SNAR1($h = 1$)	113.61	326.87	440.48

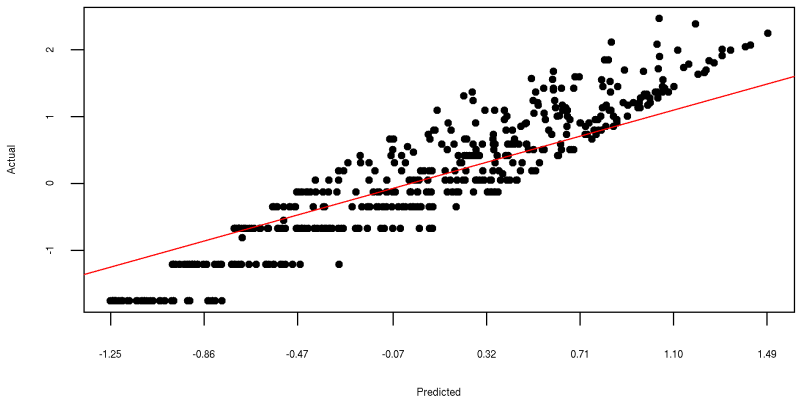
It is possible to select the best model with the lowest PMCC each solute each rain gauge. However, we are interested in a simple case of using the same model for every single time series or a global model. On the whole, SAR1 and SNAR1($h = 1$) models have very low PMCC value on the most solutes and rain gauges. Both models fit the data well, but we choose SNAR1($h = 1$) model as the best model for further analysis because it is a smaller model with fewer parameters. It also reduces resource usage, for instance, computer memory and processing time. Additionally, we also check the chosen model if it explains the data adequately.



(a) Time series plot of residuals

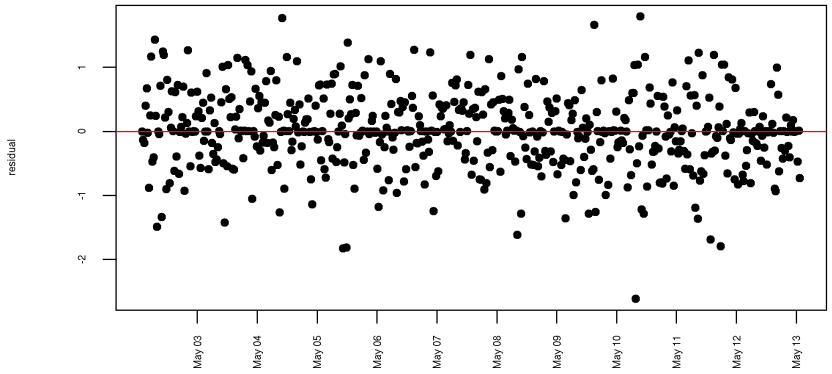


(b) Time series plot of actual data (black) and mean of predicted values (red)

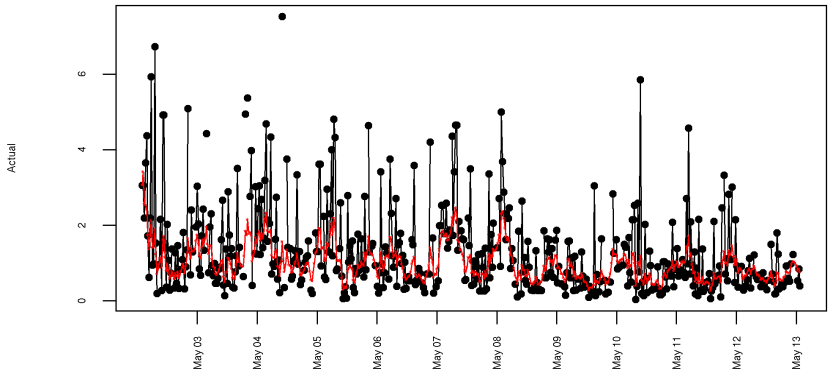


(c) Scatter plot of observed values against predicted values along with the $y = x$ line

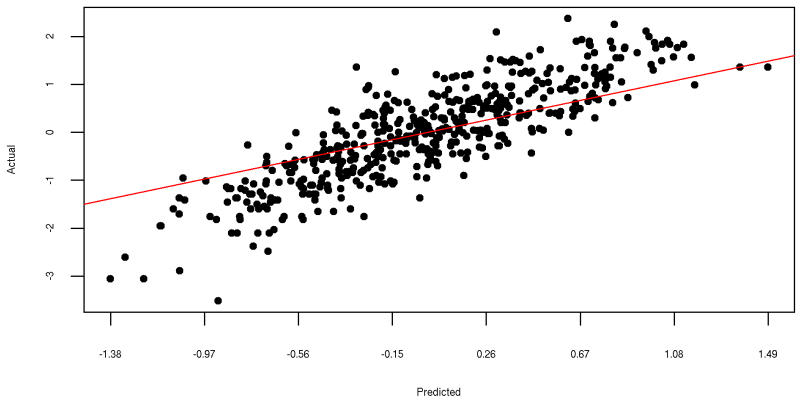
FIGURE 5.2: Diagnostic plots of SNAR1 model of Na from RG-11



(a) Time series plot of residuals



(b) Time series plot of actual data (black) and mean of predicted values (red)



(c) Scatter plot of observed values against predicted values along with the $y = x$ line

FIGURE 5.3: Diagnostic plots of SNAR1 model of SO4 from RG-22

Figure 5.2 shows diagnostic plots for SNAR1 univariate model of Na from rain gauge RG-11. The first subfigure plots residuals over time (week). The residuals are randomly dispersed around the zero line. There are not any clear patterns of residuals. This suggest that there is no serial correlation of the error terms or the error terms are independent. Also, it still shows the remaining outlier effects spiking at the same time as seen it from actual data plot in the next subfigure, e.g. the highest measurement in 2009. The second subfigure compares the actual values or measured Na concentrations in black line with mean of predicted values in red line (in logarithmic scale). It appears that the predicted value follows the observed and capture fluctuation or seasonal variation as well. The last subfigure plots actual values agaist predicted values to examine the accuracy of prediction. The ideal accuracy is all points align with the diagonal y=x line. The plot indicates overestimates on the low measurement. The stacked points may due to the precision of measurement or recordings and the lower limit detection value (a recording value if an unmeasurable solute occurs). It also indicates underestimates on the high measurement. As a result, the model fits to the data considerably and provides accurate prediction.

Another example of diagnostic plots for SNAR1 model of SO4 from rain gauge RG-22 are presented in Figure 5.3. The plot of residuals shows independent errors behaviour. The model fit to the data and has a good prediction. Similar checking on model adequacy and accuracy have been applied to each solute each rain gauges solute modelling.

Note on a candidate model, yAR1 model, it includes water year indexes as predictors of time to AR1 model. It is dealing with the long term variation in a time series or cyclical changes. we observe the lowest PMCC value of the model, especially on SO4. This may describe regular variation in monthly data of 11 water years. It need to be investigated more in detail.

5.3.2 Results for Fitting Multivariate Models

Next on multivariate case, we investigate the performance of combining set of solute models with correlated errors of the Bayesian multivariate normal. We assume that all solutes have the same fitted model or the global model which is the chosen best model from univariate case, SNAR1($h = 1$). Besides, we also conduct local SNAR1 model with a various number of sine and cosine functions ($h \in 2, 3, 4$). However, there are small changes of PMCC among SNAR1 models with the range of h compared to the model complexity. Therefore, we fit SNAR1($h = 1$) each time series combining in a multivariate model, namely mSNAR1. This makes it comparable to the univariate model results.

In practise, we investigate the performance of multivariate models according to the ways of combining solute models by varying number of rain gauges and solutes. We want to

study the effect of having: more solutes with the same gauge and the same solutes with more gauges. These are compared with Bayesian predictive approach.

Firstly, combining solute models by varying the number of solutes, Table 5.2 shows PMCC values of mSNAR1 with varied number of solutes from rain gauge RG-11 only. The lower PMCC on the maximum seven solute models illustrates a considerable improvement if the more solutes have been included into the multivariate model. In other words, the more solutes, the more information for modelling. Therefore, the multivariate models with seven solutes is preferable. Results of mSNAR1 models of RG-22 and RG-23 are also similar with RG-11.

TABLE 5.2: Predictive model choice criterion: PMCC of **multivariate** model of various solutes from Rain gauges; partitioning into goodness of fit (G), penalty (P) – Examples of RG-11

Solute	G	P	PMCC
Three solutes			
Ca	201.76	255.28	457.04
SO4	24.86	113.38	138.24
NO3	129.94	201.75	331.69
all	356.56	570.41	926.97
Five solutes			
Ca	140.42	225.36	365.78
SO4	55.78	117.38	173.16
NO3	98.54	156.95	255.49
NH4	36.13	106.62	142.75
Cl	170.68	307.30	477.98
all	501.55	913.61	1415.16
Seven solutes			
Ca	143.75	196.45	340.20
SO4	64.97	120.28	185.25
NO3	95.66	156.04	251.70
NH4	43.84	112.63	156.47
K	32.77	152.98	185.75
Na	21.31	58.60	79.91
Cl	14.99	56.07	71.06
all	417.29	853.05	1270.34

Next, we examine multivariate models of combining solute models across rain gauges. As we expect that more solutes should provide better predicted values of the model. Besides fitting the multivariate model, mSNAR1 with seven solutes, we also study the effect of including more time series from different rain gauges into the model.

Figure 5.4 illustrates the changing of PMCC of solute models from RG-11 when including more data from other gauges for two and three gauges mSNAR1 models. The details of PMCC are provided in Table 5.3. In general, it reflects declining PMCC of bigger multivariate models by including more time series for other rain gauges. Following this, we choose the biggest multivariate model mSNAR1 of seven solutes from all three gauges as the best model to describe the precipitation solute concentrations and their variation.

Although there is a benefit of having more data of the multivariate model, surprisingly, PMCC of potassium (K) models are increasing and it is less accurate in prediction. This might be dealing with the remaining intractable variation which requires further investigation.

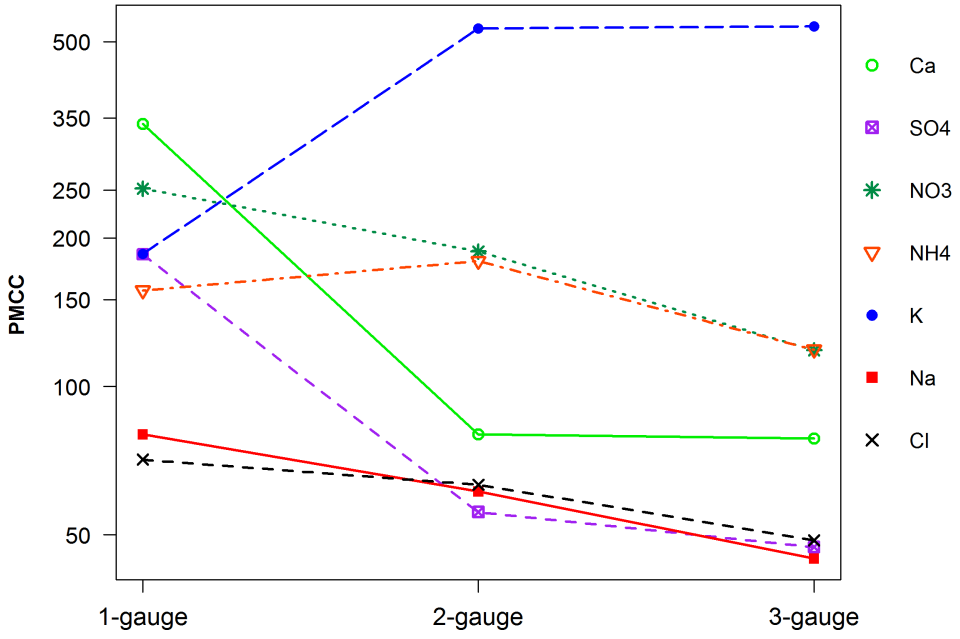


FIGURE 5.4: Predictive model choice criterion: PMCC of **multivariate** model of seven solutes from various rain gauges with RG-11 as a base gauge. Note: 2-gauge is of including RG-22 data

TABLE 5.3: Predictive model choice criterion: PMCC of **multivariate** model of seven solutes; partitioning into goodness of fit (G), penalty (P) – varying rain gauges

Solute/model	G	P	PMCC
One gauge: RG-11			
Ca	143.75	196.45	340.20
SO4	64.97	120.28	185.25
NO3	95.66	156.04	251.70
NH4	43.84	112.63	156.47

Continued on next page

Table 5.3 – continued

Solute/model	G	P	PMCC
K	32.77	152.98	185.75
Na	21.31	58.60	79.91
Cl	14.99	56.07	71.06
One gauge: RG-22			
Ca	105.92	191.08	297.01
SO4	65.43	124.49	189.93
NH4	61.66	131.86	193.52
NO3	93.90	168.23	262.13
K	69.84	155.17	225.01
Na	16.36	54.50	70.86
Cl	21.15	62.15	83.30
One gauge: RG-23			
Ca	131.90	206.43	338.33
SO4	42.82	104.28	147.11
NH4	71.85	147.40	219.25
NO3	118.54	182.63	301.18
K	56.85	194.95	251.80
Na	19.48	60.38	79.86
Cl	18.79	61.79	80.58
Two gauges: RG-11 and RG-22			
RG-11			
Ca	16.46	63.44	79.90
SO4	10.11	45.50	55.61
NH4	60.50	119.03	179.53
NO3	59.94	127.71	187.65
K	246.15	285.40	531.55
Na	14.83	46.41	61.24
Cl	13.13	50.08	63.21
RG-22			
Ca.1	29.58	78.28	107.86
SO4.1	8.79	43.04	51.83
NH4.1	92.80	142.07	234.87
NO3.1	46.85	129.14	175.99
K.1	142.98	176.15	319.13
Na.1	13.24	45.61	58.85
Cl.1	17.25	52.76	70.01
Two gauges: RG-11 and RG-23			
RG-11			
Ca	17.25	56.86	74.11
SO4	7.92	40.06	47.98

Continued on next page

Table 5.3 – continued

Solute/model	G	P	PMCC
NH4	91.26	129.56	220.82
NO3	26.25	94.65	120.90
K	245.26	278.41	523.67
Na	16.36	41.44	57.80
Cl	7.86	40.76	48.62
RG-23			
Ca.2	13.92	55.60	69.52
SO4.2	8.05	39.02	47.07
NH4.2	77.72	150.68	228.40
NO3.2	69.83	132.35	202.18
K.2	237.04	275.92	512.96
Na.2	7.30	37.30	44.60
Cl.2	24.11	55.25	79.36
Two gauges: RG-22 and RG-23			
RG-22			
Ca.1	26.08	72.28	98.36
SO4.1	9.13	44.02	53.15
NH4.1	106.59	147.80	254.39
NO3.1	35.04	110.40	145.44
K.1	143.92	175.88	319.80
Na.1	8.57	35.75	44.32
Cl.1	9.66	41.32	50.98
RG-23			
Ca.2	13.45	58.83	72.28
SO4.2	9.13	44.27	53.40
NH4.2	95.53	149.53	245.06
NO3.2	88.30	146.10	234.40
K.2	239.25	279.83	519.08
Na.2	9.20	39.10	48.30
Cl.2	8.68	40.99	49.67
All gauges: RG-11, RG-22, and RG-23			
RG-11			
Ca	22.91	55.54	78.45
SO4	9.94	37.30	47.24
NH4	32.29	86.40	118.69
NO3	29.45	88.93	118.38
K	251.44	285.76	537.20
Na	10.56	34.20	44.76
Cl	9.69	39.07	48.76
RG-22			

Continued on next page

Table 5.3 – continued

Solute/model	G	P	PMCC
Ca.1	37.17	73.18	110.35
SO4.1	15.49	43.83	59.32
NH4.1	23.90	84.12	108.02
NO3.1	53.10	114.89	167.99
K.1	150.39	180.94	331.33
Na.1	9.31	34.61	43.92
Cl.1	9.21	38.59	47.80
RG-23			
Ca.2	13.75	48.21	61.96
SO4.2	8.60	35.44	44.04
NH4.2	42.73	109.82	152.55
NO3.2	50.91	108.53	159.44
K.2	227.33	273.49	500.82
Na.2	8.01	32.91	40.92
Cl.2	10.23	38.44	48.67

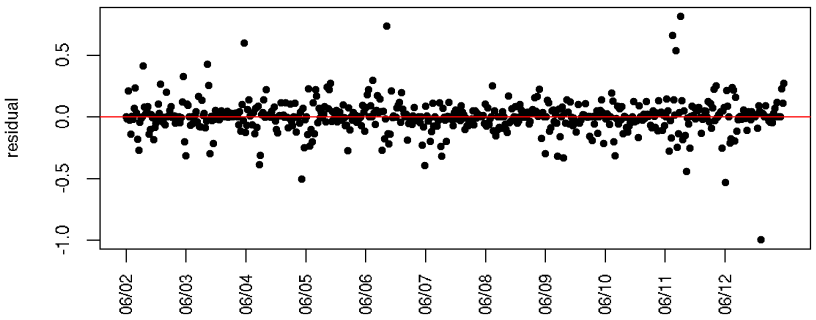
Note: one-gauge of RG-11 is from a part of Table 5.2

Figure 5.5 and 5.6 are diagnostic plots of Na from RG-11 and SO4 from RG-22 respectively. In comparison of the same selected solutes, obviously, the multivariate model provides smaller residuals and better predicted values than univariate models.

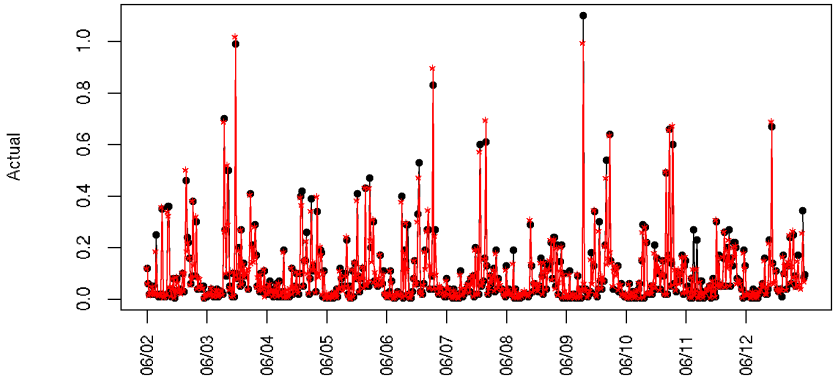
Correspondingly, the more information may induce the more reliable model by fitting a multivariate model with correlated errors instead of several individual univariate models.

5.3.3 Parameter and Load Estimates

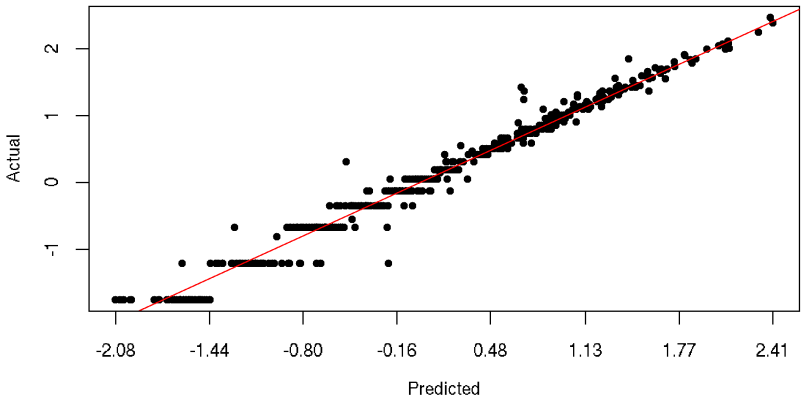
Parameter estimates of the chosen multivariate model mSNAR1 for precipitation solute concentrations from Hubbard Brook are presented in Table 5.4.



(a) Time series plot of residuals

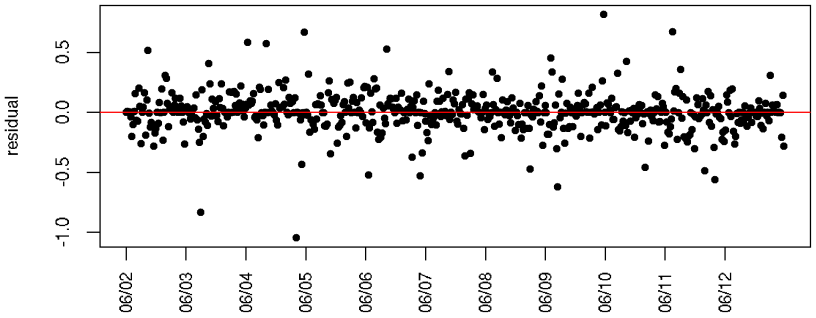


(b) Time series plot of actual data (black) and mean of predicted values (red)

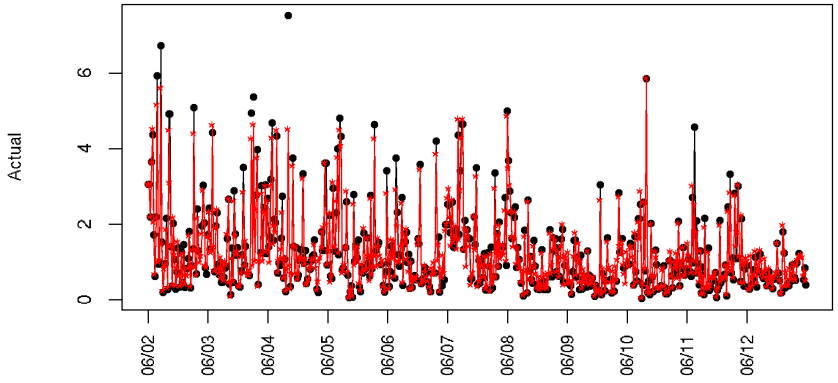


(c) Scatter plot of observed values against predicted values along with the $y = x$ line

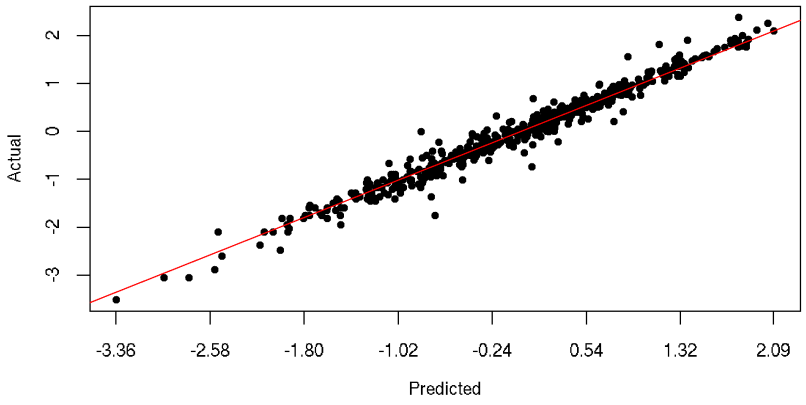
FIGURE 5.5: Diagnostic plots of mSNAR1 model of Na from RG-11



(a) Time series plot of residuals



(b) Time series plot of actual data (black) and mean of predicted values (red)



(c) Scatter plot of observed values against predicted values along with the $y = x$ line

FIGURE 5.6: Diagnostic plots of mSNAR1 model of SO4 from RG-22

TABLE 5.4: Parameter estimations of model mSNAR1 for precipitation solute concentrations from Rain gage RG-11

	Calcium (Ca)		**Sulphate (SO_4)**		**Nitrate (NO_3)**	
Parameters	Mean	95% CI	Mean	95% CI	Mean	95% CI
Regression						
μ (Global intercept)	-2.776	(-2.906,-2.647)	-0.341	(-0.475,-0.210)	-0.399	(-0.562,-0.240)
σ (SD for solutes)	0.441	(0.336,0.542)	0.313	(0.257,0.373)	0.517	(0.382,0.643)
Autoregressive term						
ϕ (autoregressive coefficient)	0.476	(0.378,0.578)	0.537	(0.449,0.629)	0.457	(0.353,0.563)
σ_η (SD for AR term)	0.886	(0.423,2.018)	0.889	(0.426,1.994)	0.887	(0.424,2.056)
Multivariate covariance matrix						
$\sigma^2_{11}, \sigma^2_{22}, \sigma^2_{33}$ (within solutes)	0.714	(0.575,0.857)	0.623	(0.529,0.722)	1.174	(0.958,1.402)
$\sigma^2_{12}, \sigma^2_{21}$ (between Ca & SO_4)			0.516	(0.439,0.602)		
$\sigma^2_{13}, \sigma^2_{31}$ (between Ca & NO_3)			0.614	(0.510,0.728)		
$\sigma^2_{23}, \sigma^2_{32}$ (between SO_4 & NO_3)			0.701	(0.598,0.813)		

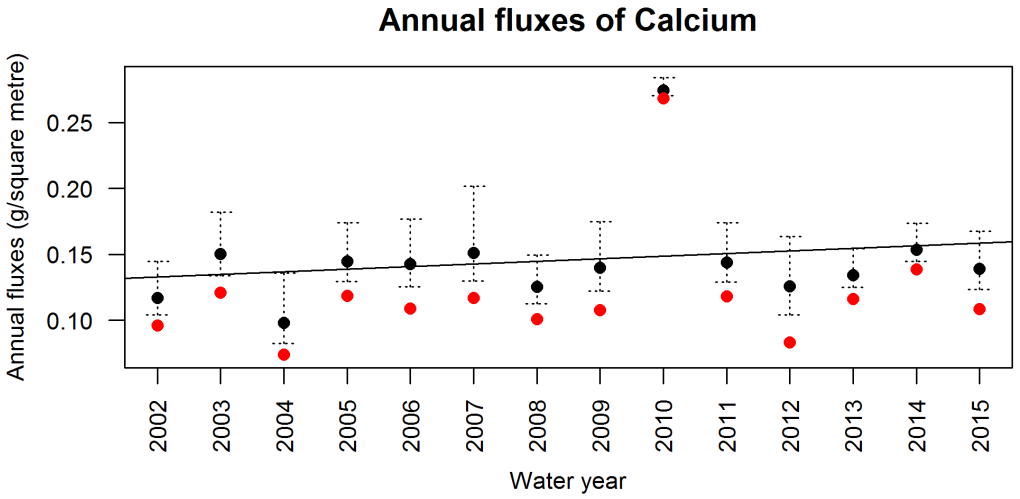
It can be seen that the autoregressive coefficient indicate a moderate correlation to the one-time previous precipitation solute concentration each week t. In the other word, the solute concentration of the current week much depends on the last week concentration. Nitrate concentration presents a larger variance than Calcium and Sulphate.

Table 5.5 and Figure 5.7 shows the posterior mean estimates and 95%CI for the annual total precipitation solute loads in mg/m^2 for 11 water years. (from June 2002 to May 2013). It is clearly seen that Calcium remains constant while there are steady decreases on Sulphate and Nitrate concentrations. The declines of the posterior mean of annual loads of solutes may indicate the progress of US policy on acid precipitation on the last decade.

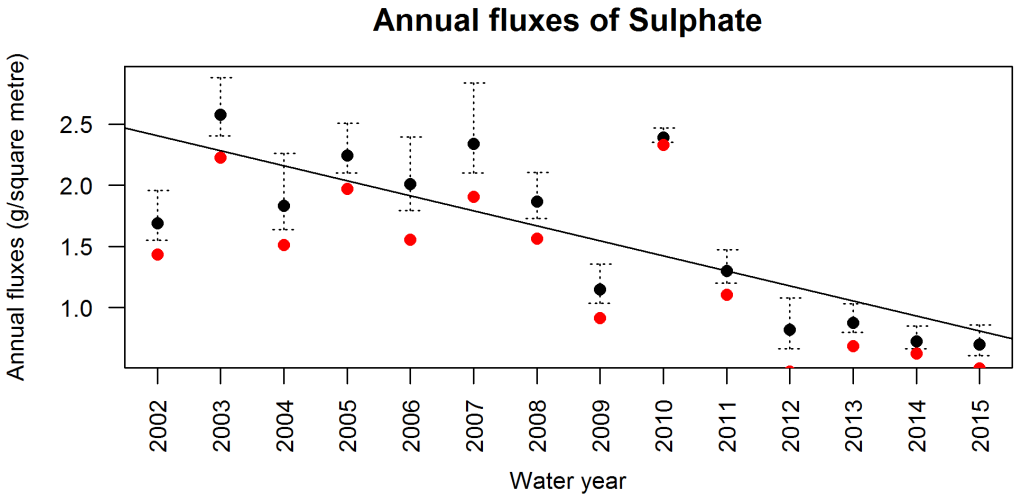
Turning into monthly view from Figure 5.8, it shows varying patterns over time indicating seasonal through the water year. There seems to be a similar pattern between Sulphate and Nitrate concentrations but Sulphate has a smaller variation than Nitrate concentrations.

TABLE 5.5: Posterior mean and 95% credible interval (CI) of annual loads of precipitation solute concentrations on the catchment area RG-11 *g*

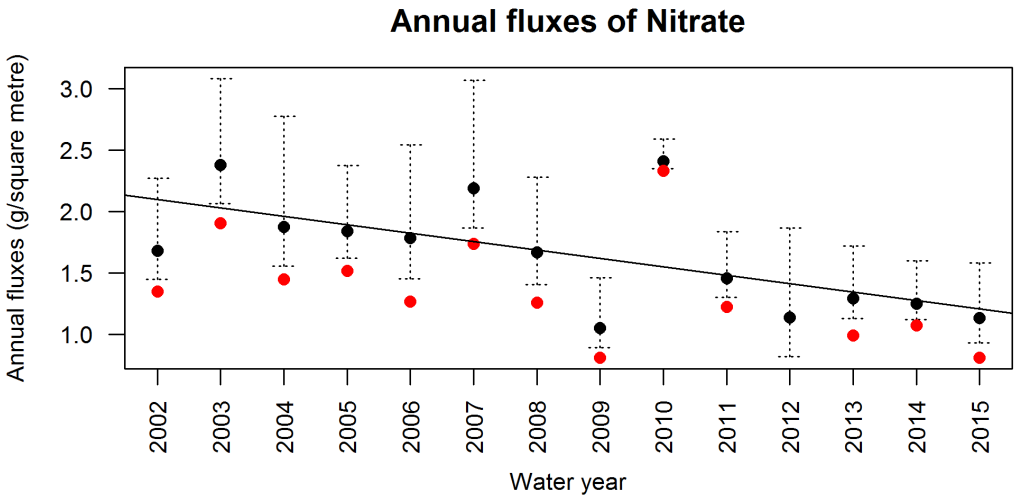
	Calcium		**Sulphate**		**Nitrate**	
Time	**Load**	**95% CI**	**Load**	**95% CI**	**Load**	**95% CI**
by water year						
2002	0.117	(0.104,0.144)	1.688	(1.55,1.959)	1.678	(1.446,2.269)
2003	0.15	(0.134,0.182)	2.574	(2.401,2.881)	2.377	(2.065,3.082)
2004	0.098	(0.082,0.136)	1.833	(1.639,2.26)	1.872	(1.554,2.774)
2005	0.145	(0.129,0.174)	2.244	(2.1,2.509)	1.838	(1.62,2.372)
2006	0.143	(0.125,0.177)	2.01	(1.793,2.396)	1.782	(1.452,2.543)
2007	0.151	(0.13,0.202)	2.339	(2.099,2.836)	2.19	(1.865,3.066)
2008	0.125	(0.113,0.149)	1.867	(1.727,2.104)	1.668	(1.404,2.277)
2009	0.14	(0.122,0.175)	1.149	(1.035,1.354)	1.052	(0.891,1.461)
2010	0.274	(0.27,0.284)	2.39	(2.353,2.469)	2.41	(2.347,2.59)
2011	0.144	(0.129,0.174)	1.298	(1.201,1.475)	1.457	(1.3,1.836)
2012	0.126	(0.104,0.163)	0.821	(0.666,1.081)	1.137	(0.818,1.863)
2013	0.134	(0.125,0.154)	0.877	(0.797,1.032)	1.291	(1.127,1.717)
By Month						
January	0.1	(0.093,0.116)	1.003	(0.946,1.115)	2.142	(1.973,2.591)
February	0.157	(0.147,0.176)	1.068	(0.975,1.219)	1.682	(1.438,2.257)
March	0.199	(0.176,0.24)	2.001	(1.809,2.319)	2.551	(2.151,3.395)
April	0.273	(0.233,0.345)	1.888	(1.526,2.531)	2.024	(1.409,3.446)
May	0.217	(0.203,0.244)	1.72	(1.619,1.909)	1.557	(1.419,1.896)
June	0.185	(0.168,0.217)	2.349	(2.211,2.602)	1.803	(1.618,2.251)
July	0.195	(0.185,0.218)	2.708	(2.565,2.989)	1.991	(1.799,2.492)
August	0.12	(0.105,0.149)	2.469	(2.289,2.795)	1.623	(1.361,2.253)
September	0.21	(0.191,0.245)	2.561	(2.391,2.832)	2.117	(1.844,2.739)
October	0.171	(0.156,0.2)	1.551	(1.411,1.822)	1.619	(1.415,2.1)
November	0.126	(0.113,0.148)	1.993	(1.873,2.2)	2.152	(1.92,2.669)
December	0.086	(0.081,0.096)	1.204	(1.158,1.308)	1.876	(1.772,2.202)



(a) Calcium (Ca)

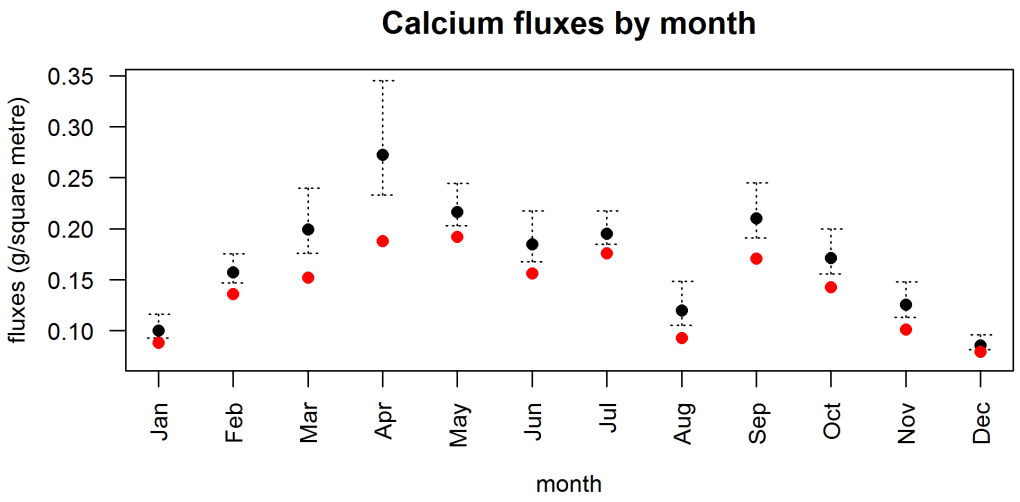


(b) Sulphate (SO4)

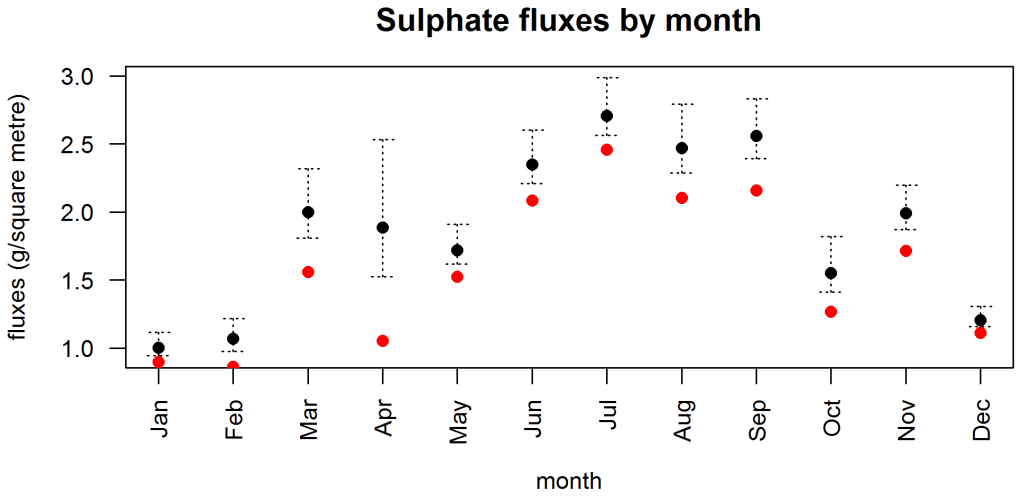


(c) Nitrate (NO3)

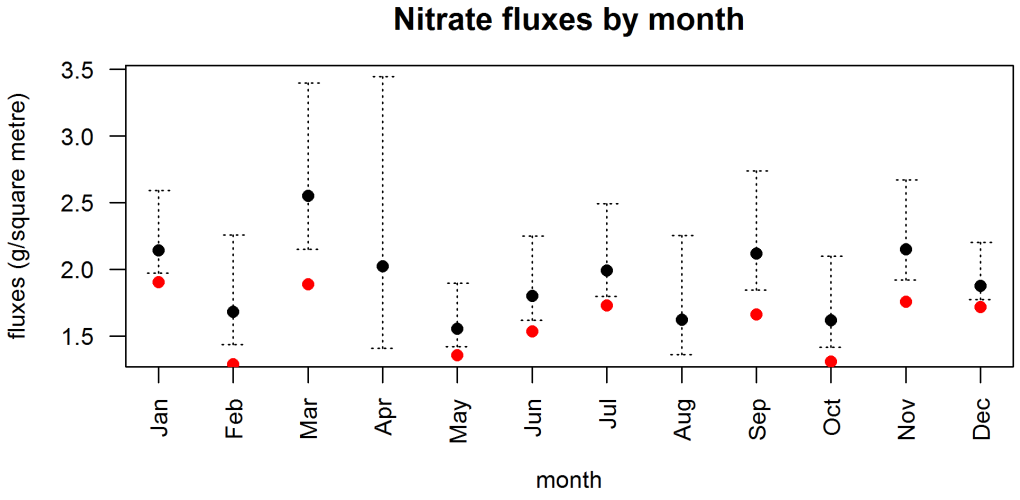
FIGURE 5.7: Plots of posterior mean and 95% credible interval (CI) of annual loads of precipitation solute concentrations on the catchment area RG-11 (by water year



(a) Calcium (Ca)



(b) Sulphate (SO4)



(c) Nitrate (NO3)

FIGURE 5.8: Plots of posterior mean and 95% credible interval (CI) of annual loads of precipitation solute concentrations on the catchment area RG-11 (By month)

Chapter 6

Conclusion

6.1 Thesis summary

Water pollution is an important global problem. Mostly, it is contaminated by excessive artificial chemicals from human activities. As a result, it will be harm to people and ecosystems. In order to prevention, we need to estimate the quantity of these pollutants which have been circulated along water cycle and end up at water resources.

However, it is not easy to estimate as these chemicals come from several sources in different forms depending on the current state in water cycle. Also, we need to collect the chemical quantities which flow through the cycle to estimate chemical load in water sources, especially in stream and precipitation.

We use the data provided by Christchurch Harbour Macronutrients Project in the UK and and the Hubbard Brook Ecosystem Study in US to construct a statistical model describing the amount of chemicals. Moreover, we have to deal with uncertainties which affect to chemical concentration such as seasonal, storm, etc. It is impossible to find the best model tracking all uncertainties, but the best possible model. In addition, examining the water sample is expensive then the most collected data are limited and ecology data itself has many specific problem such as extreme outlier, missing value, or truncated data.

As far as literatures, many kind of modelling were applied to understand the relationship between chemical concentrations and related attributes. We are focusing on constructing Bayesian multivariate Normal model to use the most information based on geography of catchment area. As a result, we may estimate annual chemical load as a target result.

Christchurch Harbour data consist of macronutrient concentations and related water characteristics data such as water temperature and sediment levels in water sample collected from two monitoring sites located on different rivers. This study offers weekly data

for one year as small data case. The data experiences some missing values and require imputation before modelling. However, imputation techniques is not our interests.

Hubbard Brook data is a kind of time series. The data was collected in long term by examining precipitation sample from rain gauges located around this experiment forest. The data was collected daily but not completely through out the years. So, we are summing the data into weekly data. Moreover, the data has been applied with many laboratory standards. This will affect to the record of very small measurement which may easy to detect.

As a result of fitting Bayesian multivariate normal model, the model performs better in capturing relationship between macronutrients and water characteristics than individual models for the Christchurch Harbour data. The similar conclusion can be drawn from the Hubbard Brook as time series example. Moreover, it shows a better performance if more sources are included into the same umbrella of Bayesian Multivariate normal model. In addition, the annual chemical loads are calculated from the selected model including with credible intervals of total loads which can not be obtained from individual models.

6.2 Future work

1. Challenges with spatial data:
   Regarding the availability of precipitation from all rain gauges stations, we may consider spatio-temporal precipitation model to estimate area precipitation base on geography. So using information from neighbour rain gauges should provide better provide precipitation estimation on that small area under uncertainties.

2. Modelling with high frequency data:
   Besides, high frequency data also provided by Christchurch Harbour Macronutrients Project under intensive data collection programme. We may applie the Bayesian Multivariate Normal model with this high frequency data.

3. Aggregated models challenge:
   Due to small samples, small changes in modelling may affect to large change in model selection. Therefore, we may reduce selection risk by model averaging or to aggregate candidate models which are likely to improve goodness of fit.

4. Bayesian modelling using other language:
   This work use, a free and convenient software, WinBUGS and JAGS as a Bayesian engine to construct Bayesian models. Also, using r2winbugs and r2jags package to reveal results for further analysis. However, the more sources combining into the model, the longer run time needed. New language such as INLA uses the different structure which claim a better run time.

Bibliography

ADAS (2007), Nitrates in water - the current status in england(2006), Technical report, Defra. [Accessd 07/12/2015].
URL: *http://webarchive.nationalarchives.gov.uk/20090731151606/http://www.defra.gov.uk/environ supportdocs/d1-nitrateswater.pdf*

Alameddine, I., Qian, S. S. & Reckhow, K. H. (2011), 'A bayesian changepoint–threshold model to examine the effect of tmdl implementation on the flow–nitrogen concentration relationship in the neuse river basin', *Water research* **45**(1), 51–62.

Bailey, A. S., Hornbeck, J. W., Campbell, J. L. & Eagar, C. (2003), *Hydrometeorological database for Hubbard Brook Experimental Forest: 1955-2000*, Vol. 305, US Department of Agriculture, Forest Service, Northeastern Research Station Newtown Square.

Buso, D. C., Likens, G. E. & Eaton, J. S. (2000), *Chemistry of precipitation, streamwater, and lakewater from the Hubbard Brook Ecosystem Study: a record of sampling protocols and analytical procedures*, US Department of Agriculture, Forest Service, Northeastern Research Station.

Cai, Y., Guo, L., Douglas, T. A. & Whitledge, T. E. (2008), 'Seasonal variations in nutrient concentrations and speciation in the chena river, alaska', *Journal of Geophysical Research: Biogeosciences (2005–2012)* **113**(G3).

Cha, Y., Alameddine, I., Qian, S. S. & Stow, C. A. (2016), 'A cross-scale view of n and p limitation using a bayesian hierarchical model', *Limnology and Oceanography* **61**(6), 2276–2285.

Chapman, D. et al. (1996), *Water quality assessments - a guide to the use of biota, sediments and water in environmental monitoring*, 2 edn, E & Fn Spon London. On Behalf of United Nations educational, Scientific and Cultural Organization, World Health Organization, United Nations Environment Programme.

Cowles, M. K. & Carlin, B. P. (1996), 'Markov chain monte carlo convergence diagnostics: a comparative review', *Journal of the American Statistical Association* **91**(434), 883–904.

Crainiceanu, C., Ruppert, D. & Wand, M. P. (2005), 'Bayesian analysis for penalized spline regression using winbugs', *Journal of Statistical Software* pp. 1–24.

Federer, A. C., Flynn, L. D., Martin, W. C., Hornbeck, J. W. & Pierce, R. S. (1990), 'Thirty years of hydrometeorologic data at the hubbard brook experiment forest, new hampshire', *Gen. Tech. Rep. NE-141. Radnor, PA: US Department of Agriculture, Forest Service, Northeastern Forest Experiment Station. 44 p.* **141**.

Galloway, J. N. & Cowling, E. B. (1978), 'The effects of precipitation on aquatic and terrestrial ecosystems: A proposed precipitation chemistry network', *Journal of the Air Pollution Control Association* **28**(3), 229–235.

Galloway, J. N. & Likens, G. E. (1978), 'The collection of precipitation for chemical analysis', *Tellus* **30**(1), 71–82.

Geisser, S. (1965), 'Bayesian estimation in multivariate analysis', *The Annals of Mathematical Statistics* **36**(1), 150–159.

Gelfand, A. E. & Ghosh, S. K. (1998), 'Model choice: A minimum posterior predictive loss approach', *Biometrika* **85**(1), 1–11.
URL: *http://biomet.oxfordjournals.org/content/85/1/1.abstract*

Gelfand, A. E. & Sahu, S. K. (1999), 'Identifiability, improper priors, and gibbs sampling for generalized linear models', *Journal of the American Statistical Association* **94**(445), 247–253.

Gelman, A., Carlin, J. B., Stern, H. S. et al. (2014), *Bayesian data analysis*, Vol. 3, Taylor & Francis.

Gorham, E. (1998), 'Acid deposition and its ecological effects: a brief history of research', *Environmental Science & Policy* **1**(3), 153–166.

Green, M. B., Campbell, J. L., Yanai, R. D., Bailey, S. W., Bailey, A. S., Grant, N., Halm, I., Kelsey, E. P. & Rustad, L. E. (2018), 'Downsizing a long-term precipitation network: Using a quantitative approach to inform difficult decisions', *PloS one* **13**(5), e0195966.

Helsel, D. R. & Hirsch, R. M. (2002), *Statistical methods in water resources*, Techniques of Water Resources Investigations, Book 4, Chapter 3A, US Geological survey Reston, VA.

Hervé, M. R., Nicolè, F. & Lê Cao, K.-A. (2018), 'Multivariate analysis of multiple datasets: a practical guide for chemical ecology', *Journal of chemical ecology* **44**(3), 215–234.

Hirsch, R. (2012), 'Flux of nitrogen, phosphorus, and suspended sediment from the susquehanna river basin to the chesapeake bay during tropical storm lee, september 2011', *U.S. Geological Survey Scientific Investigations Report 2012–5185* . as an indicator of the effects of reservoir sedimentation on water quality.

Jarvie, H. P., Neal, C., Withers, P. J., Wescott, C. & Acornley, R. M. (2005), 'Nutrient hydrochemistry for a groundwater-dominated catchment: The hampshire avon, {UK}', *Science of The Total Environment* **344**(1–3), 143–158.
URL: *http://www.sciencedirect.com/science/article/pii/S0048969705001051*

Kadlec, R. H. & Hammer, D. E. (1988), 'Modeling nutrient behavior in wetlands', *Ecological Modelling* **40**(1), 37–66.

Laud, P. W., Wisconsin, M. C. & Ibrahim, J. G. (1995), 'Predictive model selection', *Journal of the Royal Statistical Society, Ser. B* **57**, 247–262.

Likens, G., Driscoll, C., Buso, D., Siccama, T., Johnson, C., Lovett, G., Fahey, T., Reiners, W., Ryan, D., Martin, C. et al. (1998), 'The biogeochemistry of calcium at hubbard brook', *Biogeochemistry* **41**(2), 89–173.

Likens, G. E. (2013), *Biogeochemistry of a forested ecosystem*, Springer Science & Business Media.

Loucks, D., Van Beek, E., Stedinger, J., Dijkman, J. & Villars, M. (2005), *Water Resources Systems Planning and Management: An Introduction to Methods, Models and Applications*, Studies And Reports in Hydrology, Unesco.
URL: *https://books.google.co.uk/books?id=iPZRAAAAMAAJ*

Lunn, D. J., Thomas, A., Best, N. & Spiegelhalter, D. (2000), 'Winbugs – a bayesian modelling framework: Concepts, structure, and extensibility', *Statistics and Computing* **10**(4), 325–337.
URL: *http://dx.doi.org/10.1023/A:1008929526011*

Lunn, D., Jackson, C., Best, N. et al. (2013), *The BUGS book: A practical introduction to Bayesian analysis*, CRC press.

McLaughlin, S. & Wimmer, R. (1999), 'Tansley review no. 104 calcium physiology and terrestrial ecosystem processes', *The New Phytologist* **142**(3), 373–417.

Nedwell, D., Dong, L., Sage, A. & Underwood, G. (2002), 'Variations of the nutrients loads to the mainland uk estuaries: correlation with catchment areas, urbanization and coastal eutrophication', *Estuarine, Coastal and Shelf Science* **54**(6), 951–970.

Pirani, M., Panton, A., Purdie, D. A. & Sahu, S. K. (2016), 'Modelling macronutrient dynamics in the hampshire avon river: A bayesian approach to estimate seasonal variability and total flux', *Science of The Total Environment* **572**, 1449 – 1460.
URL: *http://www.sciencedirect.com/science/article/pii/S0048969716308221*

Press, S. J. (2005), *Applied multivariate analysis: using Bayesian and frequentist methods of inference*, Courier Corporation.

Qian, S. S., Reckhow, K. H., Zhai, J. & McMahon, G. (2005), 'Nonlinear regression modeling of nutrient loads in streams: A bayesian approach', *Water Resources Research* **41**(7).

Raffery, A. & Lewis, S. (1992), 'One long run with diagnostics: Implementation strategies for markov chain monte carlo', *Statist. Sci* **7**, 493–497.

Rowe, D. B. (2002), *Multivariate Bayesian statistics: models for source separation and signal unmixing*, CRC press.

Schoch, A. L., Schilling, K. E. & Chan, K.-S. (2009), 'Time-series modeling of reservoir effects on river nitrate concentrations', *Advances in Water Resources* **32**, 1197–1205.

Sigleo, A. & Frick, W. (2007), 'Seasonal variations in river discharge and nutrient export to a northeastern pacific estuary', *Estuarine, Coastal and Shelf Science* **73**(3–4), 368–378.
URL: *http://www.sciencedirect.com/science/article/pii/S0272771407000388*

Singh, A. & Agrawal, M. (2007), 'Acid rain and its ecological consequences', *Journal of Environmental Biology* **29**(1), 15.

Smith, V., Tilman, G. & Nekola, J. (1999), 'Eutrophication: impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems', *Environmental Pollution* **100**(1–3), 179–196.
URL: *http://www.sciencedirect.com/science/article/pii/S0269749199000913*

Stenback, G. A., Crumpton, W. G., Schilling, K. E. & Helmers, M. J. (2011), 'Rating curve estimation of nutrient loads in iowa rivers', *Journal of Hydrology* **396**(1), 158–169.

Sturtz, S., Ligges, U. & Gelman, A. (2005), 'R2winbugs: A package for running winbugs from r', *Journal of Statistical Software* **12**(3), 1–16.
URL: *http://www.jstatsoft.org*

Sullivan, T. J., Driscoll, C. T., Beier, C. M., Burtraw, D., Fernandez, I. J., Galloway, J. N., Gay, D. A., Goodale, C. L., Likens, G. E., Lovett, G. M. et al. (2018), 'Air pollution success stories in the united states: The value of long-term observations', *Environmental Science & Policy* **84**, 69–73.

Tiao, G. C. & Box, G. E. (1981), 'Modeling multiple time series with applications', *journal of the American Statistical Association* **76**(376), 802–816.

Tiao, G. C. & Zellner, A. (1964), 'On the bayesian estimation of multivariate regression', *Journal of the Royal Statistical Society: Series B (Methodological)* **26**(2), 277–285.

Withers, P. J. & Lord, E. I. (2002), 'Agricultural nutrient inputs to rivers and groundwaters in the uk: policy, environmental management and research needs', *Science of*

The Total Environment **282–283**, 9–24.
URL: *http://www.sciencedirect.com/science/article/pii/S0048969701009354*