UNIVERSITY OF GLASGOW

Temporal and Spatial Modelling of Extreme River Flow Values in Scotland

by

Maria Franco Villoria

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

in the

College of Science and Engineering

School of Mathematics and Statistics

February 2013

Declaration of Authorship

I, Maria Franco Villoria, declare that this thesis titled, 'Temporal and Spatial Modelling of Extreme River Flow Values in Scotland' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where 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.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- 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.
- Part of the work presented in Chapter 3 has been published in the *Journal of Environmental Statistics* (Franco-Villoria et al. (2012)).

Signed: _____

Date: _____

Abstract

Extreme river flows can lead to inundation of floodplains, with consequent impacts for society, the environment and the economy. Flood risk estimates rely on river flow records, hence a good understanding of the patterns in river flow, and, in particular, in extreme river flow, is important to improve estimation of risk. In Scotland, a number of studies suggest a West to East rainfall gradient and increased variability in rainfall and river flow. This thesis presents and develops a number of statistical methods for analysis of different aspects of extreme river flows, namely the variability, temporal trend, seasonality and spatial dependence. The methods are applied to a large data set, provided by SEPA, of daily river flow records from 119 gauging stations across Scotland. The records range in length from 10 up to 80 years and are characterized by non-stationarity and long-range dependence.

Examination of non-stationarity is done using wavelets. The results revealed significant changes in the variability of the seasonal pattern over the last 40 years, with periods of high and low variability associated with flood-rich and flood-poor periods respectively. Results from a wavelet coherency analysis suggest significant influence of large scale climatic indices (NAO, AMO) on river flow.

A quantile regression model is then developed based on an additive regression framework using P-splines, where the parameters are fitted via weighted least squares. The proposed model includes a trend and seasonal component, estimated using the back-fitting algorithm. Incorporation of covariates and extension to higher dimension data sets is straightforward. The model is applied to a set of eight Scottish rivers to estimate the trend and seasonality in the $95th$ quantile of river flow. The results suggest differences in the long term trend between the East and the West and a more variable seasonal pattern in the East.

Two different approaches are then considered for modelling spatial extremes. The first approach consists of a conditional probability model and concentrates on small subsets of rivers. Then a spatial quantile regression model is developed, extending the temporal quantile model above to estimate a spatial surface using the tensor product of the marginal B-spline bases. Residual spatial correlation using a Gaussian correlation function is incorporated into standard error estimation. Results from the $95th$ quantile fitted for individual months suggest changes in the spatial pattern of extreme river flow over time. The extension of the spatial quantile model to build a fully spatio-temporal model is briefly outlined and the main statistical issues identified.

Acknowledgements

First and foremost I would like to thank my supervisors Marian Scott and Trevor Hoey for their valuable experience, knowledge, patience, support and enthusiasm over the last four years. None of this work would have been possible without their help and motivation. I am also very grateful to my supervisor Denis Fischbacher-Smith for his help and input in this project. I gratefully acknowledge the financial support provided by the Kelvin-Smith scheme from the University of Glasgow and SEPA for providing much of the data as well as an external co-supervisor, Alastair Cargill, whose contribution in the initial stage of the project was very helpful.

I am deeply grateful to all the people in the department, they have really made me feel welcome over the last 5 years I've spent here. My office and next-door-office mates have been great, thank you for all your help, company and affection, and for putting up with my crazy knocking! All the laughs, chats and 'creative moments' made all the hard work much easier to cope with. In special thanks to Joanna, for all the fun times outside the department.

During my time in Glasgow I have made some really good friends who have made me feel 'at home' (or almost!), thanks to Belgin and Mariana, you always know how to make me forget about the 'old friend'! Thanks also to Ross, Francesco, Alfre and Gabi. And to Paul, for all the 'inner peace' moments followed by spaghetti and ice-cream! Many thanks to all my friends back home. In particular, thanks to Lauri, my 'friendita', you are always there at the other end of the phone ready to cheer me up instantly, no matter what.

I have a great family and I can't thank you enough, you have always supported and encouraged my ideas and my dreams, even if that means being far away from you. Thanks to my sisters and my 'wee' brother, you always bring out the best in me. In particular, to my parents, you've taught me that with hard work, dedication and enthusiasm, I can achieve anything I dream of.

Thanks to Massi, for coming back, this time to stay. Glasgow would have been very different without you, *grazie per regalarmi tanti di quei momenti che hanno veramente quel gran tenore di scomparsa totale del mondo.*

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The study of extreme values in environmental data, for example air pollution levels, temperatures, sea level and river flows, has received increased interest recently (Alexander et al. (2009); Heffernan and Tawn (2004); Reich (2012)). It is now recognized that methods specifically developed to analyze extreme values, i.e. those that deviate considerably from expected average levels, are needed. In particular, river flow extremes might lead to the occurrence of floods. Even though extreme river flow values are not very likely to occur, their consequences can be devastating and entail risks at every level: human, environmental, social and economic.

Appropriate assessment of river flow levels, and, in particular, extreme levels, is essential for planning purposes. The earliest example is probably found in the annual floods of the River Nile in Egypt, that conditioned the development and organization of the population settled on the banks of the Nile in a decisive manner (Bell (1970)).

The modelling of river flow values has been the subject of study for a long time. In Scotland, part of the importance of studying river flow levels is due to the fact that rivers are widely used for public water abstraction, electricity production and recreational activities such as fishing, and have ecological significance. On the other hand, river flow levels are the basis of flood risk assessment. In the recent past, a small number of large floods have had a profound impact (Black and Burns (2002)), especially in the West, with significant economic and social impacts (Werritty and Chatterton (2004)). The

average annual cost of flooding in Scotland is estimated to be about 31.5 million for inland flooding, and 19.1 million for coastal flooding (Werritty and Chatterton (2004)). The consequences for people who have been directly affected by floods include increased levels of stress, anxiety, fear of future floods and lost of items of sentimental value (Werritty et al. (2007)). Identification of patterns in extreme river flow behaviour, mainly in the form of seasonality and long term trends, is of importance so that changes can be identified and decisions appropriately made to avoid, when possible, or alleviate any negative impacts.

1.1 Flood Risk Assessment

Three aspects need to be put together for tackling flood risk: modelling and forecasting, building infrastructure and land-use management (Kidson and Richards (2005)). The bulk of the work presented in this thesis contributes to the modelling aspect and concentrates on fluvial flooding, as distinct from pluvial flooding where high intensity rainfall overwhelms drainage system capacity. River flow records tend to be rather short (usually less than 50 years), with the consequence of methods having to rely strongly upon extrapolation, for which flood frequency analysis methods can be used, by fitting an appropriate model to the data. This fitting process involves making assumptions *a priori* about the underlying distribution of the process that generates floods, a distribution that is unknown for extreme values beyond the observed record. In the UK the first comprehensive approach to flood risk assessment was the Flood Studies Report (FSR) (Natural Environment Research Council (1975)), later substituted by the Flood Estimation Handbook (FEH) (Institute of Hydrology (1999)). The FEH includes updated annual maxima (AM) and peak over threshold (POT) databases as well as catchment descriptors for 943 gauged sites in the UK, along with methodology for dealing with both kinds of data. These can be divided into two main blocks; statistical methods, which make use of well known probability distributions, and rainfall-runoff models. The latter can be used to create synthetic POT series when data are not readily available. These kind of models usually assume the runoff response to rainfall, snowfall and temperature inputs to be time invariant and have many limitations (Black and Burns (2002)), mainly due to the combination of various sources of uncertainty and the need of constant

updating (Werritty (2002)). For further details the reader is referred to Institute of Hydrology (1999). The FEH statistical methods are based on the estimation of the median flow Q_{MED} and the growth curve (ratio between the T-year event Q_T and Q_{MED} for different values of return period T (Institute of Hydrology (1999)). For gauged rivers, the Q_{MED} is estimated using annual maxima series, while for ungauged rivers, it can be estimated from catchment descriptors and adjusted using gauged data from a catchment of similar characteristics. Regional frequency analysis (RFA) (Fowler and Wilby (2010)) assumes that data from the same region have the same frequency distribution, but differ in magnitude, so that they can be pooled together (with the resulting series assumed to be stationary) and a single distribution can be fitted. Prior to pooling, each seasonal maximum time series is centred by its median value. Then a three parameter generalized extreme value (GEV) distribution is fitted to each regional pooled series using maximum likelihood estimation (MLE) (Coles (2004)) and return levels (plus 95% associated confidence intervals) can be estimated (re-scaled to the regional average R_{MED}).

The methods contained in the FSR and the FEH, once considered as the 'gold standard' for flood risk estimation, are built upon 3 main assumptions: stationarity, homogeneity and random occurrence of extreme events. However, whether these assumptions are reliable is not so clear anymore and new and improved methods might be needed. "The new challenge of addressing climatic non-stationarity in flood risk estimation, which certainly appears to be a valid concern in Scotland, will present a substantial test for the hydrological community to provide risk estimates in the near future which serve well the needs of more distant times ahead" (Black and Burns (2002)). There seems to be spatial heterogeneity in the incidence of maximum recorded flood peaks; in Scotland, peak flows registered over the period 1970-1996 exceed previous records only in certain locations (Black and Burns (2002)), suggesting that not every method might be appropriate for every river, but a more individualized approach should be taken, or rather a general one that allows for spatial differences. Temporal clustering of large floods appears to be another issue (Black and Burns (2002)). All these changes make the improvement of flood risk estimation a difficult task and highlight the importance of carrying out studies on a regional level.

Understanding the pattern of river flows and its relationship to flooding is critical to

flood planning and risk management. To do so, new and improved estimates of flood risk which take into account the impact of climate change and possible spatial heterogeneity are needed. Records of river flows are widely used to predict flood and low flow levels, in water resource allocation, and form an important basis for assessing the impacts of climate change. Data records are often short (a few decades or less) and a range of classical time-series and extreme value methods have been used to describe the data and as a basis for making predictions.

The occurrence of severe floods since the late 80s (Black and Burns (2002); Black (1996)), regarded as a shift towards a 'flood rich' period, and the pronounced spatial pattern, with differences between the North-West and South-East across the whole of Britain (Marsh (1995)), raised awareness of the need for revising and updating the existing methodology and legislation.

1.1.1 Environmental Policy: the Flood Management Act (2009)

In 2009 the Scottish Flood Risk Management Act (The Scottish Government (2010)) was introduced to update the previous legislation by taking a more sustainable approach and was designed to meet the requirements of the European Union Directive on Flood Risk Management (2007/60/EC). The document summarizes the main responsibilities for the different government bodies, with The Scottish Environment Protection Agency (SEPA) as the leading body. The new act promotes close collaboration between SEPA and the relevant local authorities for improving the assessment and mapping of flood risk and dealing with flood management planning. Part of SEPA's responsibilities include delivering flood warnings and flood hazard maps (for a probability of 0.5% in any given given year). The methodology underpinning SEPA's flood hazard maps for estimating flow levels follows the Flood Estimation Handbook (Institute of Hydrology (1999)).

1.2 Observed and predicted changes in extreme values

Changes in extreme rainfall and river flow events are reported in a number of studies during the last thirty years, which in turn suggest an increase in flood risk. An increasing trend in both magnitude and frequency of high flows (Prudhomme et al. (2003))

and extreme rainfall events (Fowler and Wilby (2010)) in the last 30-50 years in the UK has been identified. The highest extremes are found in the West of the UK, especially in the Scottish Highlands (Fowler and Ekström (2009)). Changes in extreme events seem to be seasonally dependent. Autumn 2000 was recorded as the wettest in England and Wales since 1766 (Fowler and Kilsby (2003)). Analysis of monthly total rainfall over 1766-2000 for England and Wales reported no significant trend for the whole series, but significant results for the seasonal series, in the form of an increasing trend in winter (October-March) and a decreasing trend in summer (April-September) (Robson (2002)). Similar findings are reported for extreme river flow (in England and Wales), with no significant trends detected at a national level but found locally, especially for Scottish rivers (Robson (2002)).

In particular, in Scotland, no general trend seems to be valid across the country and observed changes are not homogeneous, neither in frequency, with estimates of changes in return periods changing with location, nor in time, with seasonal changes in extreme rainfall being more pronounced than overall ones, especially in autumn (Fowler and Wilby (2010)). Spatially, two "micro-climates" have been identified over the period 1980-2000: wetter in the North-West, with significant spring and autumn increases in the West and winter increases in the North, and drier in the South-East, especially in the summer months (Jenkins et al. (2009); Werritty (2002)). Differences in frequency and magnitude of river flow extreme events have been found between the East and the West (Black and Burns (2002); Black (1996)), as well as in trends in annual maxima series (Black (1996)). Downscaled projections from global circulation models (GCMs) for 2050s suggest that this tendency is to continue in the near future (Fowler and Wilby (2010); Werritty (2002)). Rainfall (and hence river flow) is expected to increase in the winter, with evidence of this happening in the West of Scotland. Changes are expected not only in runoff, especially in wet catchments, but also evapotranspiration (Mansell (1997)).

Long-term trends in Scottish rainfall and runoff during 1970-1996 were investigated by Werritty (2002), who identified a dry period in the 1960s-70s followed by the wettest period on record in the late 1980s-early 1990s, when frequencies of extreme events exceeded earlier ones and new peak values were registered in western rivers (Black and

Burns (2002)). Similarly, Black and Burns (2002) identify the periods 1946-1955 and 1976-1995 as periods with particularly high frequency of extremes, mainly in the West, while the East presents higher event frequencies during 1950-1960s. These findings are in agreement with the so called 'flood poor' (1964-1973) and 'flood rich' (late 1980s-early 1990s) periods (Grew and Werritty (1995)).

Changes in river flow in Scotland do not exactly mirror changes in precipitation and seem to reflect an increase in variability over the last 50-60 years. Werritty (2002) found significant increases in mean flows during 1970-late 1980s in the South-West of Scotland, especially in winter, and decreases in the North and East, while summer flows decreased overall (although not significantly). Low flows rose in 1970-1985 to then decline steeply up to 1996, while high flows increased in frequency from the mid 1980s. Werritty (2002) claims that it is very likely that Scotland will be wetter as a whole, with higher river flows (especially in winter and autumn) which in turn would impact flood risk.

Climate variability has a high influence on flood behaviour. Changes observed in the UK have been partially attributed to an increase in the frequency of westerly airflows since the 1970s (Mayes (1996)). These airflows are associated with high annual precipitations and increases in POT frequencies (Black and Burns (2002)). The observed data suggest appreciable variability in time and space across Scotland. Motivated by these observations, much work has been put into predictive studies that try to assess whether the changes are to continue in the near future.

The increase of global temperature is probably the most well known effect derived from climate change. The yearly average near-surface temperature is expected to increase over the next 100 years by between 1.5 and 6 C^o, depending on the model used to make the predictions (Hunt (2002)). In maritime areas, such as the UK, this increase is likely to cause increased flooding risk. Further, there are other consequences of climate change, such as changes in precipitation frequency and storm activity (Werritty (2002)) that may have an influence on flooding (Robson (2002)). An expected consequence of climate change is an increase in frequency and intensity of extreme rainfall events (Fowler and Kilsby (2003); Fowler and Wilby (2010)), leading to increases in flooding events and

changes in seasonality (Fowler and Kilsby (2003)). Extreme events impacts are usually felt on a local or regional level; the problem is that changes in climate extremes are difficult to identify at this level and therefore require a global approach (Alexander et al. (2009)). Climate change models are now available to simulate climatic scenarios, so that predictions can be made on what to expect under certain conditions. The UK Climate Projections (Jenkins et al. (2009)) provides a summary of the expected projections for the UK under different scenarios.

Global circulation models (GCMs), also known as general circulation models, have been developed to simulate time series of climatic variables in the world, the most common ones being temperature and precipitation (Alexander et al. (2009)). These, combined with CO_2 emission scenarios, provide estimates of future changes in the climate under certain conditions (Prudhomme et al. (2003)). These models operate on a global scale, but regional models (RCMs) are also available. The range of CO_2 emission scenarios considered includes low, medium-low, medium-high and high (Jenkins et al. (2009)). However, it is not possible to assign probabilities to scenarios, as it is still unknown which scenario is more likely to happen. This results in most climate change studies being done to provide estimates under all four possible scenarios. The type of emission scenario might have a different impact on the flood regime; higher CO_2 emissions would result in larger changes that lower ones. However, variation is much larger for the high emission scenario (Prudhomme et al. (2003)). To assess model performance, the period 1961-1990 has been established as the control period (Prudhomme et al. (2003)). The spatial resolution of global and regional climate models is too coarse when compared with hydrological processes, for which predictions are needed on a local level. Models for downscaling are available, although they are not very reliable at predicting extreme precipitation on a local scale except for winter (Fowler and Ekström (2009); Kingston et al. (2009)). Their performance is usually poor in terms of spatial distribution of the predictions (Fowler and Ekström (2009)).

It is easier to detect changes in regional rainfall data than in flood data; this is because in a flood event more components than just the climate, such as reservoirs, land management, etc (Fowler and Kilsby (2003); Prudhomme et al. (2003); Fowler and Wilby (2010)) intervene. Hence climate model projections are made for precipitation extremes,

which have then to be 'translated' into flood risk terms. Overall, climate models project rainfall increases in the UK, in particular in winter on the West of Britain (Jenkins et al. (2009); Fowler and Ekström (2009); Fowler and Wilby (2010)). Predictions of increases in heavy rainfall from climate models agree with actual observed tendencies in both the UK and worldwide (Fowler and Kilsby (2003)).

RCM projections for 2071-2100 extreme rainfall in the UK vary considerably, seasonally and regionally (Fowler and Kilsby (2003); Fowler and Ekström (2009)). While the results are confident in terms of that there will be an increase, there is much uncertainty when it comes to predicting the magnitude of these increases. Overall, RCMs underestimate regional precipitation extremes for all seasons in the UK (Fowler and Ekström (2009)), with larger differences in winter and spring in those regions where extreme precipitations tend to be large. These include the North and South of Scotland and the West of England (Fowler and Ekström (2009); Prudhomme et al. (2003)). In particular, extreme precipitation for the North of Scotland is underestimated in winter and autumn. Another issue is that the consistency amongst different models is usually very poor; different models tend to disagree largely in the summer, and agree in the winter (Fowler and Ekström (2009)) in regions with low precipitation values and low variability (Fowler and Ekström (2009)).

2071-2100 extreme rainfall predictions for the UK suggest increases for autumn, spring and winter. The largest changes are predicted for autumn and spring (Fowler and Kilsby (2003)). Predictions for winter extreme rainfall are much greater in the West than in the East of the UK (Fowler and Ekström (2009); Fowler and Wilby (2010)). Climate models predict a decreasing trend in summer rainfall, especially in the South and East of the UK (Fowler and Kilsby (2003)). However, the predictions are highly dependent on the model used and the uncertainty ranges are too wide for the predictions to be meaningful, suggesting that models might not be very reliable for the summer. Another reason for which the models are not reliable for summer predictions is that they are poor for convective events, which usually take place in summer (Fowler and Wilby (2010)). Prudhomme et al. (2003) investigated climate model projections for extreme river flow in five small catchments in the UK. Their results suggest an increase in magnitude and frequency of flood events. Very large floods are expected to increase in magnitude, while

smaller events are expected to become more frequent, especially in Northern England and Scotland.

In Scotland, RCMs predict increases in winter extreme rainfall (Fowler and Kilsby (2003)) and an increase of annual runoff of 5-15% across the country for the 2050s but that could locally exceed 25% (Werritty (2002)), highlighting the importance of working on a local level to avoid underestimation. Arnell (1996) (reported by Werritty (2002)) also reports increases in Q_{95} (low flows) of 5% or less on the rivers Don, Almond and Nith, and increases in Q_5 (high flows) of up to 24%. In Scotland, the projected changes in precipitation suggest rainfall increases that differ amongst the North, South and East of the country for the four seasons, especially in winter (Fowler and Wilby (2010)). Fowler and Wilby (2010) suggest that the projected changes are more likely to be noticeable in the near future in the East and South than in the North of Scotland.

To conclude, the gathered evidence suggests statistically significant changes in annual peak-over-threshold magnitude and frequency and annual maxima trends in Scotland during 1956-1995. In particular, dry (1960s-1970s) and wet periods (late 1980s-early 1990s) have been identified. The observed changes are not homogeneous over time or space, with no consistent increase in the size of extreme river flow values (Werritty (2002)) and increased variability. Climate model predictions suggest that these differences are likely to continue and/or increase in the near future, with fairly reliable estimates over the winter months but great uncertainty over the summer.

1.3 Structure and Objectives

This thesis explores and develops a range of statistical models for analyzing extreme river flow over time and space. The main issues identified in the previous section can be divided in three topics which form the bulk of the work presented here. First, there seems to be an increase in variability in extreme river flow and precipitation, which in turn might result in a much more variable flood regime. This is an important issue as it affects the assumption of stationarity on which most methods are built on. The first piece of work concentrates on assessing the variability of extreme river flow series by using

wavelet analysis. Second, a number of studies have aimed to identify trends in extreme river flow values. These are usually based on annual maxima or peak-over-threshold series, and the significance of trends assessed by means of linear regression or Spearman's correlation coefficient in most cases. However, trends seem to be more complex than can be expressed by a simple linear regression, with increasing and decreasing periods over the time frame investigated and more flexible models might be needed. On the other hand, changes seem to be seasonally dependent. Flexible models that accommodate non-monotonic trends and seasonal changes are proposed using quantile regression. Third, observed changes in frequency and magnitude of extreme river flow are not homogeneous over space. Predictions are poor in terms of the spatial distribution, and a more informed knowledge of the current spatial distribution of extremes and how the spatial pattern is changing over time might prove helpful for improving predictions in the near future. For this, two different ways of assessing spatial dependence in extreme river flow are investigated. The remainder of the thesis is organized in 5 chapters.

Chapter 2 describes the available data. River flow is highly influenced by precipitation, weather and catchment characteristics, so the chapter starts with a brief description of those. In particular, an exploratory analysis is presented for eight Scottish rivers and the main features and issues of the data are identified. These eight rivers will be followed through the remainder of the thesis. One of the main issues is the correlation structure of the data, whose sample autocorrelation function is characterized by a slowly decaying exponential function, suggesting long-range dependence. As this thesis concentrates on extreme river flows, extreme value analysis is introduced and a peak-over-threshold model fitted for each individual river. The chapter finishes with a short description of two large scale climatic indices, the North Atlantic Oscillation (NAO) and the Atlantic Multidecadal Oscillation (AMO), whose influence on river flow is explored in Chapter 3.

Another issue identified in Chapter 2 is the nonstationarity of the river flow series, mainly due to changes in the variability. Chapter 3 introduces wavelet analysis as a way of analyzing nonstationary series simultaneously in the time and frequency domains. Not only does it provide a decomposition of the series, allowing the trend and seasonality to be isolated, but also any other possible cyclic components in the data, it also identifies what is the main source of variability and whether there are significant changes in the

variability over time. Results from a wavelet analysis applied to monthly maxima for the eight rivers are presented. To finish the chapter, the relationship between monthly maxima and large scale climatic indices is explored using wavelet coherency.

Chapter 4 concentrates on explicitly modelling the trend and seasonality of extreme river flow. For this purpose, quantile regression is used. This approach allows models of high quantiles of the distribution of river flow conditioning on time to be built, so that the temporal evolution can be assessed. The model is fitted in a P-spline framework, avoiding the use of linear programming methods to fit the parameters and using a weighted least squares approximation instead. The proposed temporal quantile model is fitted to each of the eight rivers individually assuming independent observations. The dependence structure is accounted for in the inferential process to produce approximate pointwise confidence bands for the fitted model that account for the long-range dependence structure of the data identified in Chapter 2.

Chapter 5 explores the spatial dependence in extreme river flow by means of two different approaches. The first one consists of a conditional probability model that was proposed by Keef et al. (2009). An example of its application on two different areas of Scotland is shown, as well as on the eight rivers selected in Chapter 2. The second approach is an extension of the temporal quantile model presented in Chapter 4. The spatial model is built in a regression-like context using bi-variate P-splines. The chapter ends with an outline of how a fully spatio-temporal model could be fitted.

Chapter 6 summarizes the results and main findings, comparing the results obtained from the different methods used to analyzed river flow, in particular extreme river flow. It also points out the limitations in the methodology used and suggests areas of study which should be considered in the future.

All the analysis was carried out using the software package R (R Development Core Team (2001)).

Chapter 2

Environmental Context of River Flow

This chapter describes the main characteristics of Scottish rivers, and how average river flow patterns have changed over the last forty years. These changes vary depending on a river's geographical location and catchment characteristics (Smithson (1969); Acreman and Sinclair (1986)). The geographical location is important as it will determine the weather patterns that influence the corresponding catchment and hence river flow. Further, the variability in precipitation type, amount and seasonality across geographical locations has a direct impact on river flow. To understand better the differences in flow regimes of these rivers, as well as their spatial interdependence, we need to understand first the different weather patterns that affect Scotland, as well as how catchment characteristics vary. A summary of these is followed by a description of the river flow data available. A detailed exploratory analysis is performed to identify the main features of the data for eight rivers, covering a range of geographical locations and catchment sizes. Then a possible model for the correlation structure in the data, namely long range dependence, is introduced. Since this thesis is concerned with better understanding extreme events in river flow, extreme value theory is summarized, along with an application to the eight rivers used for the exploratory analysis. Finally, two large scale climatic indices that will be investigated as possible covariates, the Atlantic Multidecadal Oscillation (AMO) and the North Atlantic Oscillation (NAO), are introduced. The chapter concludes with a summary of the main characteristics of the data and the points that need to be addressed in the following chapters.

2.1 Rainfall regions and weather patterns

River flow is directly related to the amount of precipitation that falls within a catchment. Hence, flow differences between rivers reflect to some extent the different precipitation and weather patterns that affect Scotland. A widely used weather type classification for the UK is that of Lamb (1972). His classification, consisting of seven categories, is category A (anticyclonic type), C (cyclonic type), W (westerly type), NW (north-westerly type), N (northerly type), E (easterly type) and S (southerly type) (Lamb (1972); reported in Steel (1999)). In Scotland, three categories, namely the westerly, south-westerly and cyclonic types appear to be the main influences on weather (Steel (1999)), with the westerly being the predominant one (Steel (1999)). Scotland is a small country (78.789 km^2) but with very varied topography. Topography interacts with air flow, so influencing local climate, resulting in great variability in weather and precipitation patterns over space (Mayes (1991)). That, coupled with the South-West to North-East UK prevailing wind gradient, which is accentuated by topography, means that a global classification like that of Lamb (1972) might be too coarse to describe the complex weather systems that affect Scotland. Instead, a regionalized classification scheme might be more appropriate (Mayes (1991)).

Of special interest are those weather patterns that can directly be linked to rainfall. An illustration of rainfall patterns over the period 1981-2010 can be seen in Figure 2.1. In general, low pressure systems passing through Scotland are associated with wet and windy weather, while high pressure systems are related to more stable conditions and usually bring drier and less variable weather (Barnett et al. (2006)). Pressure systems themselves change from year to year, and seem to be more variable in winter than in summer. Barnett et al. (2006) report winter decreases in pressure in the North of Scotland, an area of mainly low pressures, from 1961 to 2004 and no or little change in the South, resulting in a marked increase in the North to South gradient in pressure (Barnett et al. (2006)). Rainfall increases and their relationship to weather systems for the period 1959-1963 were investigated by Smithson (1969), who calculated the percentage of total annual rainfall related to each of nine weather patterns using data from seven stations across Scotland. His findings are summarized in Table 2.1. These nine weather patterns can be classified into three groups (Smithson (1969)):

a) Frontal situations (warm, cold and occluded fronts)

   1. **Warm fronts** are responsible for a high percentage of the total rainfall in the West coast of Scotland. As the front approaches the western mountains, rainfall intensity, but not duration, usually increases. In the East, rainfall decreases in amount, intensity and duration.

   2. **Cold fronts** are more frequent than warm fronts. They are related to greater intensity but shorter duration rainfall than warm fronts, which translates into short heavy showers rather than long-duration rainfall events. As the cold front makes its way from the western islands to the mainland, mean intensity increases while duration decreases (due to topography). In the East, rainfall associated with cold fronts is fairly stable.

   3. **Occluded fronts** have a similar effect to warm fronts in the West, but the effect of topography is stronger. However, their contribution to rainfall in the East coast increases. In the East, the number of occluded fronts that generate rainfall is smaller than in the West (where most of them generate rainfall) but when they do, the amount of rainfall that the front generates does not decrease (with respect to the West) and the events are longer in time. This translates into greater variability in the East than in the West; i.e. they either do not cause rainfall, or they do and then it is intense and of long duration.

b) Airstream situations (maritime polar, continental polar, Artic and warm sectors) are related to a progressive decrease in rainfall from windward to leeward sides. The pressure gradient is highly influential; showers are more frequent near low pressures and less near high pressures. Another important factor is the direction; airstreams coming from the North West of Scotland (uplands) subside on the leeward sides, increasing stability and decreasing the chance of showers. In particular:

   1. **Maritime polar** airstreams are associated with high intensity rainfall (heavy showers).

   2. **Continental polar** airstreams are associated with stable conditions and light rain.

3. **Artic** airstreams account for a smaller rainfall contribution in the West coast than in the East coast.

4. **Warm sectors** are responsible for the most variable rainfall spatial distribution over Scotland. They are related to high intensity rainfall over the mountains (Steel (1999)). On the East, their effects are unimportant.

c) Cyclonic (low pressure) situations (non-frontal depressions and non-frontal thunderstorms) are often causes of flooding in eastern areas of Scotland (Steel (1999)):

1. Rainfall distribution related to **non-frontal depressions** is variable, but affects mainly the East, due to the large number of depressions over the East, especially in summer. Topography does not increase rainfall intensity but prolongs its duration slightly.

2. **Non-frontal thunderstorms** are isolated by nature and make a small contribution to rainfall totals. However, their contribution is significant as they produce very high intensity rainfall. They tend to happen during the summer in the East (and be of convective origin) and in autumn in the West, where they have a frontal origin.

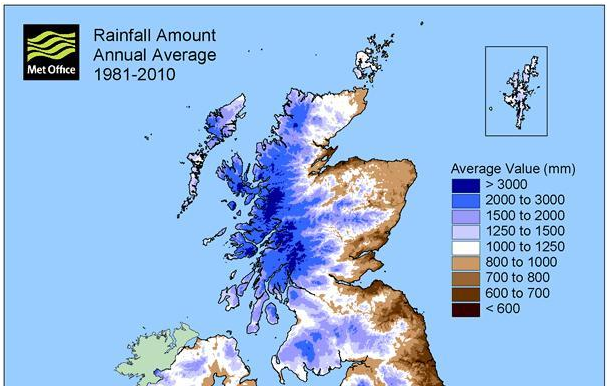

FIGURE 2.1: Annual average rainfall amount in Scotland over the period 1981-2010. Source: MetOffice (2012)

Smithson (1969) shows considerable spatial variation as well as some temporal variation in rainfall distribution over Scotland. Westerly events seem to be more influential on upland stations, while cyclonic situations are more influential in the lowlands (Steel (1999)); this results in the South being drier than the North (Barnett et al. (2006)). The main difference in the origin of rainfall appears to be found between the western mountain areas and the East, rather than just between the West and the East (Barnett

Weather type	West	Mountains	East	Overall
Warm front	main contributor	intensity increases same duration	decrease in intensity, amount and duration	
Cold front	high intensity, short duration	increase intensity, decrease duration	stable	more frequent than warm fronts
Occluded front	similar to warm front	stronger topography effect	highly variable	
Maritime				high intensity rainfall
Continental				stable conditions, light rain
Artic	small contribution		greater contribution	
Warm sector		high intensity	unimportant	most variable spatial distribution
Non frontal depression		small increase in duration	main effects	variable distribution
Thunderstorm	autumn		summer	isolated, small contribution

TABLE 2.1: Summary of rainfall characteristics associated with weather patterns in Scotland (Smithson (1969))

et al. (2006); Steel (1999); Smithson (1969)). As Barnett et al. (2006) point out, the difference between the East and the West is not clear cut, but the 'line' that divides the two regions follows the drainage divide. As a result of these differences, three distinct areas can be defined, the western coast, the (western) mountains and the East (coast). Rainfall in the western coastal area is mainly related to frontal situations. Rainfall in the mountains is mainly related to categories in which topography has an important effect (Smithson (1969)), namely fronts, warm sectors and maritime polar airstreams, all with a predominantly westerly origin. On the west coast these categories still contribute to rainfall, but to a lesser extent than over the mountains. Rainfall towards the East coast is mainly related to non-frontal depressions and occluded fronts. However, the variability in the East is greater than in the West (Smithson (1969)). Barnett et al. (2006) also found a difference between the East and the West in the number of days of intense rainfall (defined as any day with rainfall over 10mm) during the period 1961-1990. Annual averages in the West (1631-4495mm) were higher than in the East (700-1000mm), where the maximum number of consecutive dry days had increased. Precipitation totals and days of heavy rainfall over the period 1961-2004 also show a spatial pattern that is seasonally dependent (Barnett et al. (2006)). The Met Office (MetOffice (2012)) provides information on rainfall annual averages for the UK over thirty year periods. It reports a small average increase (2%) in rainfall during 1981-2010 with respect to the previous period (1971-2000), although the increase varies up to 10% when looking at seasonal changes. In Scotland, the North West has an average rainfall of between 2000 and over 3000 mm, in comparison with the East, whose average values are betwen 700 and 1000mm (Figure 2.1).

Floods in Scotland are usually triggered by either "convective storms, cyclonic/frontal precipitation (along warm, cold or occluded fronts), or precipitation resulting from the orographic uplift of airstreams" (Steel (1999)). In small Scottish catchments, the main factors responsible for floods are frontal rain (Black and Werritty (1993); reported in Steel (1999)) and thunderstorms (Black and Werritty (1997)). In mountain areas, snowmelt can be an important flood generating mechanism (Black and Werritty (1997)). Flooding in eastern areas of Scotland is usually caused by cyclonic (and southerly) conditions while in the North and West they are related to westerly conditions (Steel (1999)).

2.2 Catchment Characteristics

The UK is divided into river basin districts, each of which comprises a set of river basins and is regulated by the corresponding authority (SEPA in Scotland). Scotland (apart from a small part in the South) forms a river basin district on its own that comprises 2005 rivers and a number of other water bodies such as lakes, coasts and groundwaters (SEPA (2005)). In addition, 580 small (with a catchment area <10km^2) further rivers have been identified (SEPA (2005)). A classification based on altitude, catchment size and geology divides these 2005 rivers into fifteen classes, with a clear differentiation between the East and West.

Catchment characteristics play an important role in the complicated relationship between precipitation, runoff and river flow. The size and location of the catchment, along with the gradient, type of soil and land use can help explain why under similar climatic conditions two rivers may respond in a different manner. Acreman and Sinclair (1986) highlight the importance of catchment characteristics saying that adjacent basins can have very different physical and hydrological characteristics and hence a regionalisation based solely on geographical location might not be appropriate. As Werritty and Hoey (2004) point out, nearby catchments, specially small catchments, can display different flood regime trends.

In Scotland, not only is there a weather gradient between the East and the West, but also physical characteristics of the catchments differ. This means that both factors coupled together have an influence on the occurrence of flood events. Overall, eastern facing catchments are less steep and larger than those western facing. In particular, the highest (altitude-wise) and steepest catchments are located in the North-West, and the lowest gradients in the South and East (Werritty and Hoey (2004)). The combination of high altitude steep catchments and high rainfall totals in the North West results in drainage densities (average length of river channels in a 1km^2 area) being much higher in the North West than in the rest of Scotland (Werritty and Hoey (2004)).

Acreman and Sinclair (1986) proposed a hydrological regionalization of Scotland based solely on representative catchment characteristics, including basin area, slope, lake storage, mainstream length and base flow index. Using information from 168 gauging stations, Acreman and Sinclair (1986) identified five regions (Figure 2.2) using clustering analysis, with the catchments included in each region not necessarily being geographically contiguous. Table 2.2 provides a summary of the clusters, as well as the main hydrometric areas (Figure 2.3) they include.

- **Cluster 1**: eastern catchments characterized by better drained soils and more variable soil moisture deficit, with a greater proportion of summer floods (Acreman and Sinclair (1986)). These catchments are mainly located in the North-East.
- **Cluster 2**: large basins with small lake storage. Small-scale events tend to result from isolated floods on sub-basins or from low-intensity rainfall over a larger area, while larger floods are related to an increasing number of sub-basins responding together (Acreman and Sinclair (1986)). This is the largest cluster by far. It includes nearly all the lowlands (apart from a couple of small areas included in Cluster 5), big river catchments and a couple of areas up North.
- **Cluster 3**: this cluster cannot be easily defined as its attributes are highly variable. It is very small and includes two basins, one in the far North and another one in the North-East.
- **Cluster 4**: generally small catchments in the North-West.
- **Cluster 5**: a small cluster, mainly formed by small basins, where rainfall tends to be more uniform that on large basins, and relatively far from each other geographically.

2.3 River data

Daily river flow data have been selected for 119 gauging stations across Scotland (Figure 2.4) on the basis of geographic location, quality and length of the records. Data were provided by SEPA and the National River Flow Archive (NRFA). The length of the records is variable, ranging from 20 up to 79 years. For all 119 rivers, data are

Cluster	Main Location	Characteristics	Floods	Hydrometric Areas
1	East	well drained soils, variable soil moisture deficit	summer	9, 10, 11, 14 parts of 7 and 16
2	lowlands big river catchments North	large basins with small lakes storage	small events from floods on sub-basins or low-intensity rainfall over a large area large floods from a set of sub-basins responding together	2, 4, 7, 8, 12, 15, 17, 19, 20, 21, 77, 78, 79, 83, 84, 96, 97 parts of 3, 9, 11, 18
3		very small		small part of 3 and 15
4	North-West			5, 6, 16, 94 parts of 3, 15, 18
5		small basins		86, 89, parts of 21

TABLE 2.2: Main characteristics of the catchment clusters identified by Acreman and Sinclair (1986)

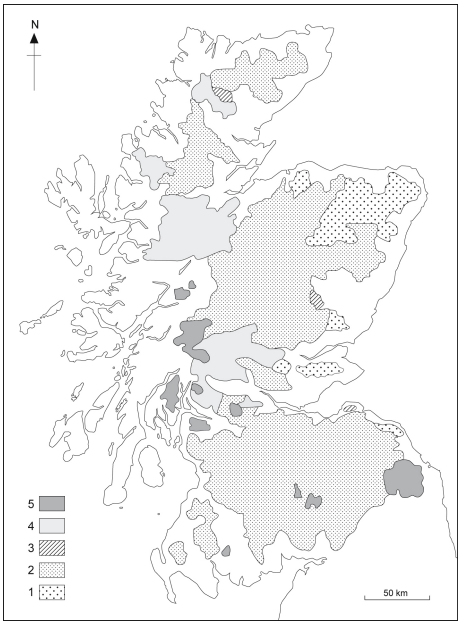

FIGURE 2.2: Clusters based on catchment characteristics. Modified from Acreman and Sinclair (1986)

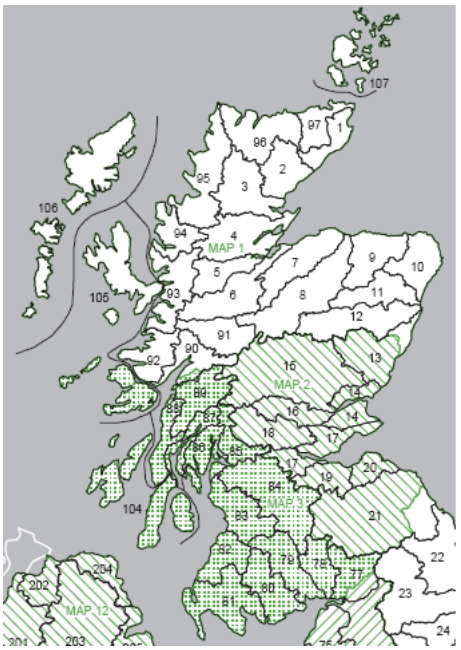

FIGURE 2.3: Scotland's hydrometric areas. The shading represents the three regions defined by SEPA: North, East and West. Source: Marsh and Hannaford (2008)

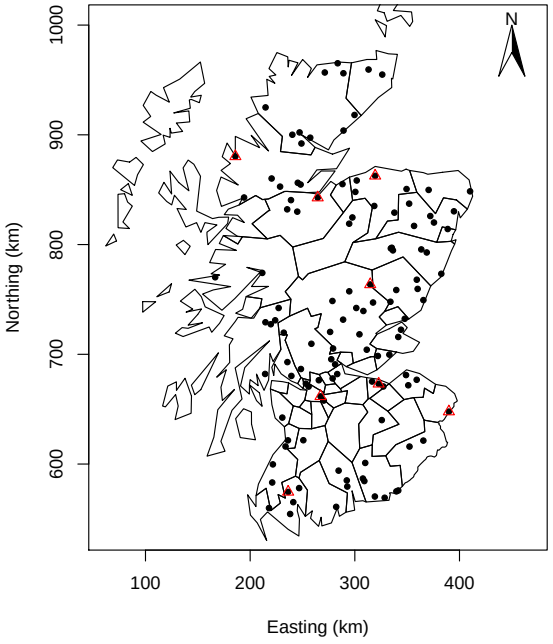

FIGURE 2.4: Location of the 119 selected gauging stations. The red triangles denote the eight river gauging sites presented in the exploratory analysis. Note the regions in the map do not correspond to hydrometric areas but to counties

available over the period 7^{th} November 1995 - 31^{st} December 2005 (3703 observations). The 29^{th} of February was removed and the resulting series were log transformed to stabilize the variance and improve normality (see Appendix A for frequency distributions). Two of the river series included zero values, station 17016 (Lochy Burn, one zero value) and station 20002 (West Peffer, ten zero values). Both are very small rivers (catchment sizes 14km^2 and 26.2km^2, respectively) that are affected by industrial and agricultural abstraction. The zero values were substituted with the value 0.0005m^3/s so that logged series could be produced. Forty-three series contained missing values. However, the missing proportions (< 0.1) were small enough not to be a concern, and missing values were imputed using linear interpolation. The interpolation was done separately for each month to better reproduce the variability of the series; i.e. missing values in January were imputed using only recorded values in January, and so on. This was done using the R function `approxfun`. This river data set is believed to be fairly representative of the population of Scottish rivers, including rivers whose catchments vary in size, characteristics and geographical location.

2.3.1 Exploratory analysis

Only a small set of rivers are presented here individually, summarized in Table 2.3 and shown in Figure 2.5. A table with the main characteristics of all 119 rivers is included in Appendix A. These were chosen to represent a range of catchment sizes and locations. Five of the rivers (Tay, Water of Leith, Tweed, Clyde and Water of Minnoch) belong to Acreman and Sinclair's cluster 2, two (Ewe and Ness) to cluster 4 and one (Lossie) to cluster 1. An illustration of the time series plot of the river Tay can be seen in the top graph of Figure 2.14. The remaining ones are included in Appendix A.

Unless stated otherwise, catchment information was obtained from the NRFA webpage (NRFA (2008)). The catchments of these rivers are a combination of lowlands and uplands, the gauging stations being placed in the lower part of the catchment. Permeability plays an important role in determining the amount of rainfall that runs into the river. Most of the catchments are a mixture of moderate-high permeable soil (lower part of the catchment) and non-permeable soils (higher part of the catchment), with the Tay and Ness being predominantly non-permeable. The catchments of River Ewe and Water

Station	River	Location	Catch.Area	Time period	Length (days)
7003	Lossie	Sheriffmills (NE)	216km^2	01/10/63-31/10/07	16152
15006	Tay	Ballathie (NE)	4587km^2	01/10/52-31/10/08	20471
19006	W. of Leith	Murrayfield (SE)	107km^2	01/01/63-31/12/05	15695
21009	Tweed	Norham (SE)	4390km^2	01/10/62-31/10/08	16821
94001	Ewe	Poolewe (NW)	441km^2	19/10/70-31/12/08	14309
6007	Ness	Ness-side(NW)	1839km^2	01/01/73-01/02/10	13537
84013	Clyde	Daldowie (SW)	1903km^2	01/10/63-04/11/08	16460
81006	W. of Minnoch	Minnoch Bridge(SW)	141km^2	01/08/86-31/12/09	8548

TABLE 2.3: Main characteristics of the eight selected rivers for exploratory analysis

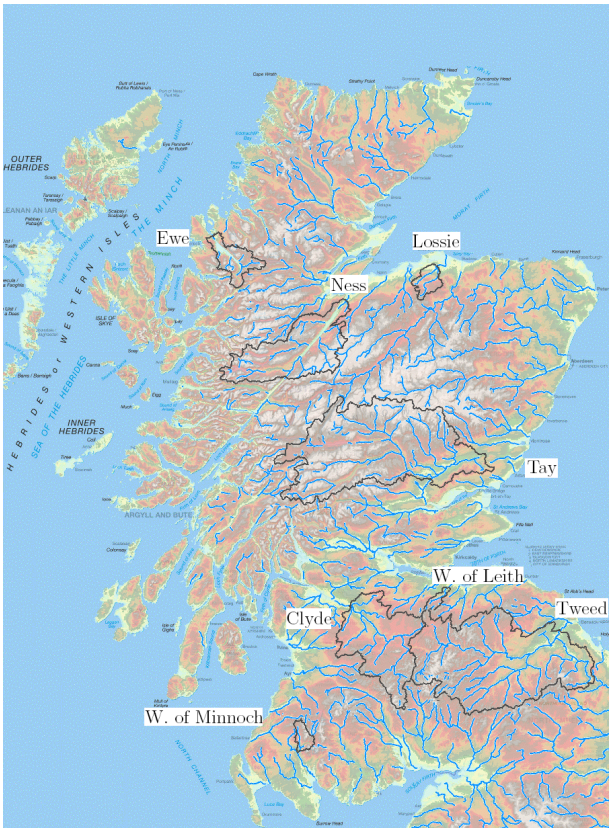

FIGURE 2.5: River catchments corresponding to the eight rivers analyzed

of Minnoch are completely non permeable, while the Water of Leith is mainly highly permeable. Another characteristic important for a catchment's response to rainfall is the presence of lakes and reservoirs. The catchments of rivers Ewe and Ness contain large lakes. The River Ewe drains through Loch Maree, which exerts a big influence on its flow regime. The Tay catchment also has several lakes that significantly reduce

the flooding impact of heavy rains and snowmelt in the higher catchment by attenuating flood peaks (Macdonald et al. (2006)); due to the effects of the Loch Tay, the daily variation on the river Tay is smaller than on any other Scottish river, making this record particularly stable. On the other hand, the construction of the "Breadalbane Hydro Scheme" in the 1950s had an important influence on flow rates (Macdonald et al. (2006)). In the Tweed catchment there are some reservoirs whose effects (less than 3% of the catchment) are not significant for major floods but might help mitigate minor ones (Fox and Johnson (1997)). Some of these rivers are regulated; the Tay and Ness are affected by hydro-power regulation (NRFA (2008); Macdonald et al. (2006)). Public water supply and industrial or agricultural abstraction reduce runoff in the Tay, Tweed and Lossie, and the Water of Leith is subject to regulation from surface water and/or ground water. Runoff on the Clyde is increased by effluent returns. Of special interest since in an urbanized area are the Lossie (Elgin), Water of Leith (Edinburgh) (this having the largest urbanized area), Ness (Inverness), Clyde (Glasgow) and Tay (Perth).

A boxplot of monthly data is presented for each river in Figures 2.6-2.13 to help identify the main features of the data. There is a clear annual seasonal cycle; however, this cycle does not seem to be constant over the years. Hence, we can find periods with a very pronounced seasonality characterized by higher flows during the winter and lower flows during the summer and periods where the variability is very small and hence there is not such a marked difference between winter and summer. There are also some years where there are two cycles at some sites.

Given the large amount of data, it is difficult to detect at first glance any trends or patterns in a time series plot of the data (Figure 2.14, top figure). To identify the overall trend as well as the seasonal part of the series the `stl` (Cleveland et al. (1990)) R function was applied to the daily series. This function decomposes the time series as the sum of trend, seasonal and residual components. Different smoothing window values for the trend and seasonality were considered, but no major differences were found. The resulting decomposition (with seasonal smoothing window=365 days and trend smoothing window=3 years) for the river Tay is shown in Figure 2.14. Results for the remaining seven rivers are included in Appendix A.

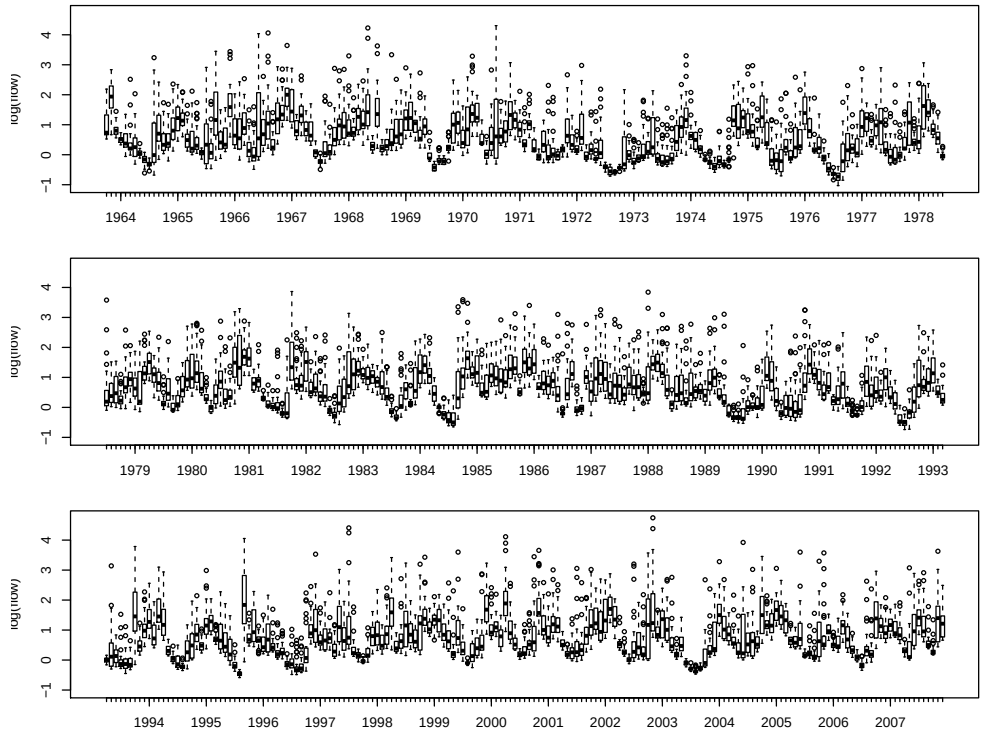

FIGURE 2.6: Monthly boxplot of river flow ($\log(m^3/s)$). River Lossie (Station 7003)

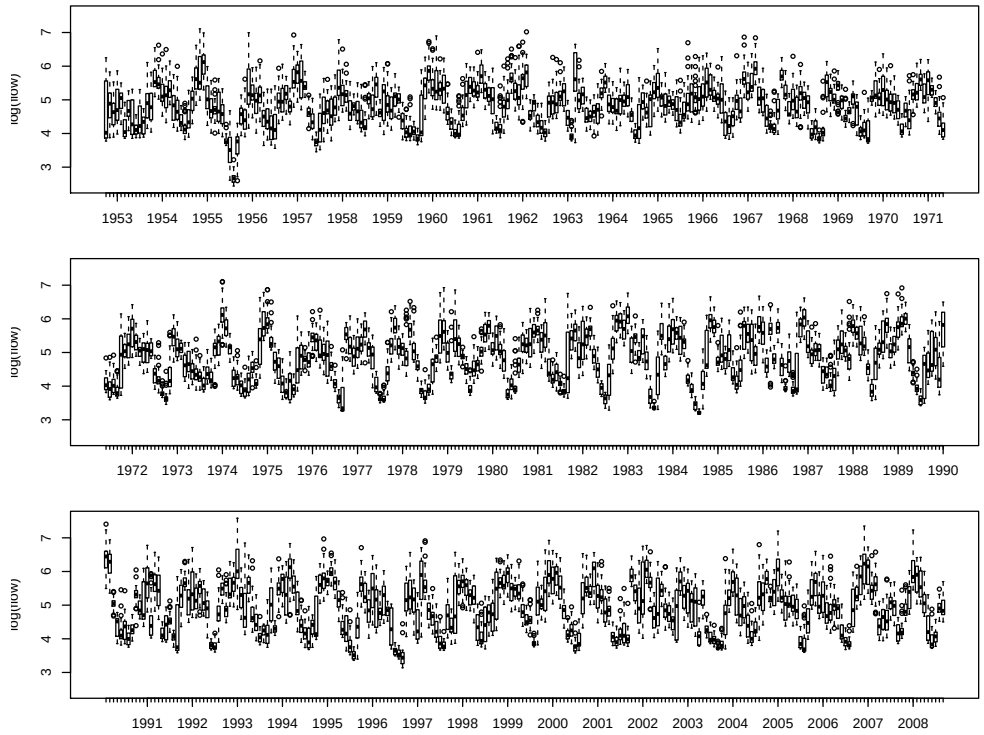

FIGURE 2.7: Monthly boxplot of river flow ($\log(m^3/s)$). River Tay (Station 15006)

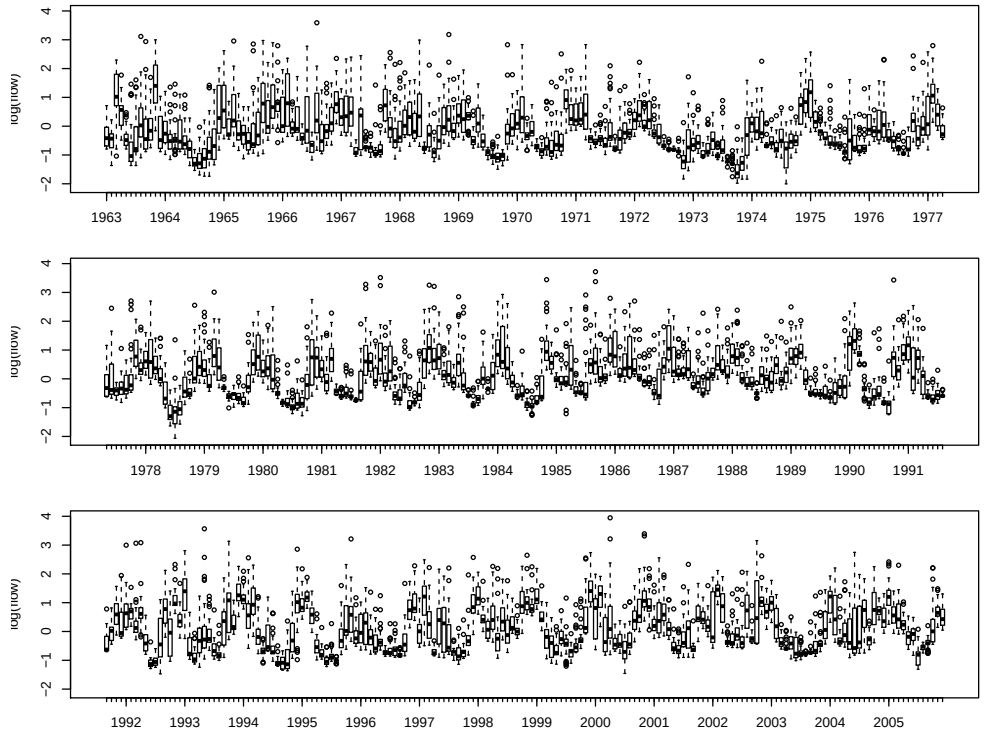

FIGURE 2.8: Monthly boxplot of river flow ($\log(m^3/s)$). Water of Leith (Station 19006)

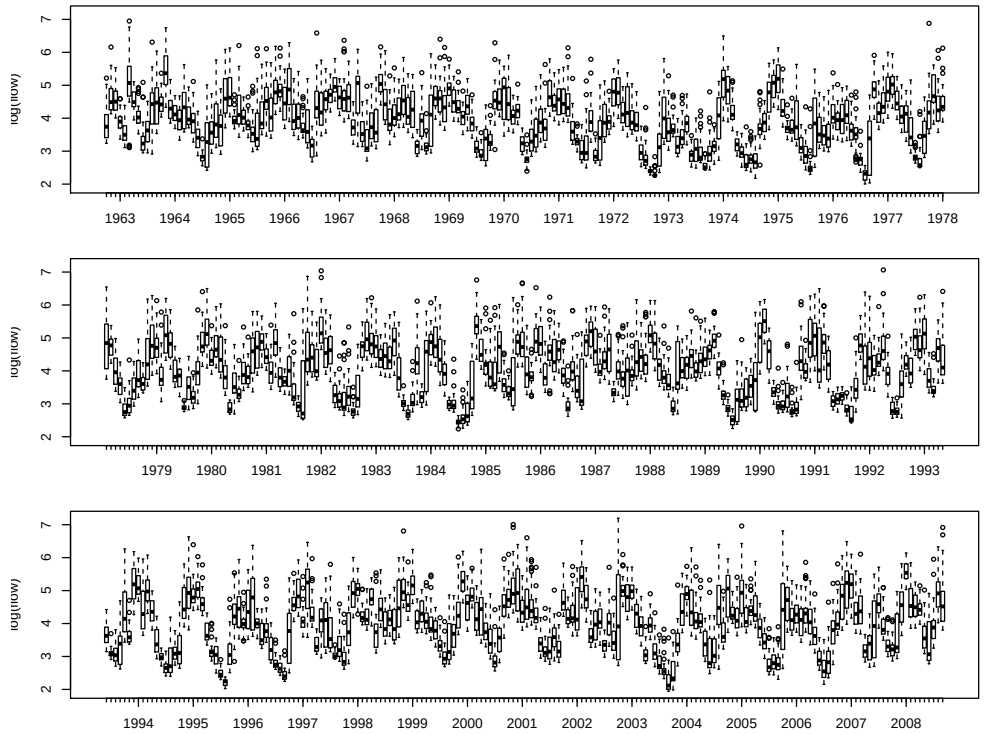

FIGURE 2.9: Monthly boxplot of river flow ($\log(m^3/s)$). River Tweed (Station 21009)

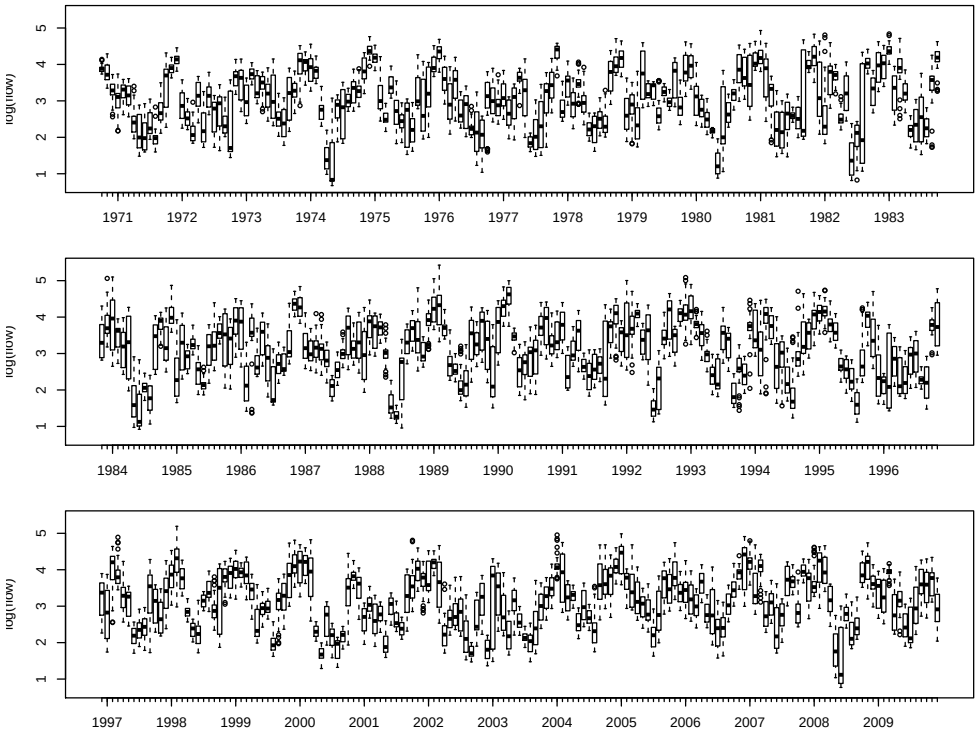

FIGURE 2.10: Monthly boxplot of river flow ($\log(m^3/s)$). River Ewe (Station 94001)

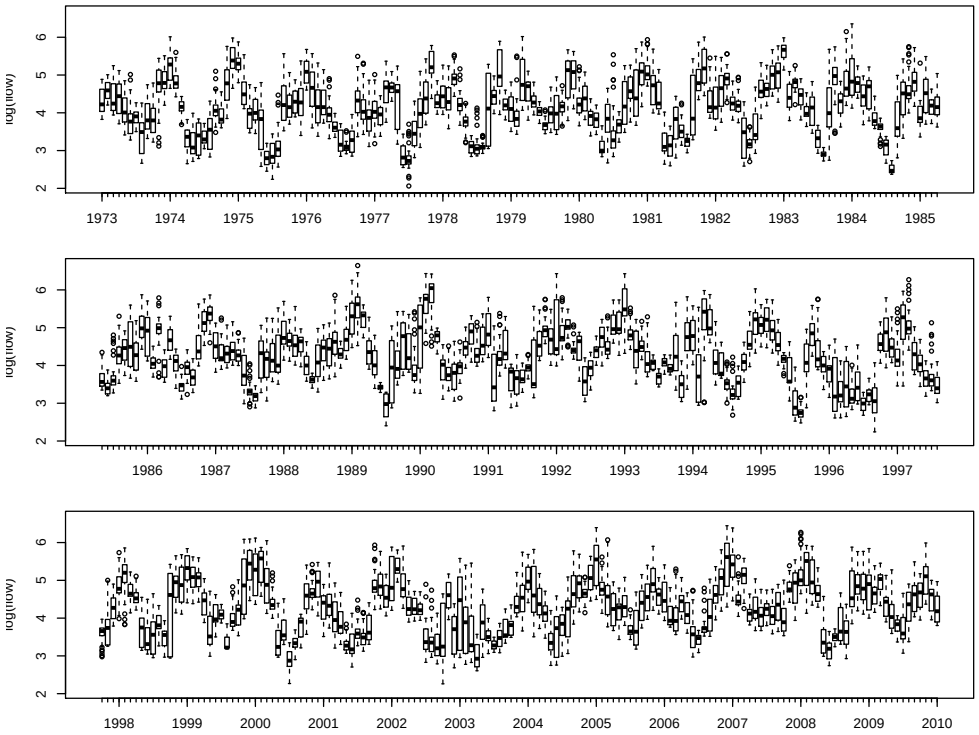

FIGURE 2.11: Monthly boxplot of river flow ($\log(m^3/s)$). River Ness (Station 6007)

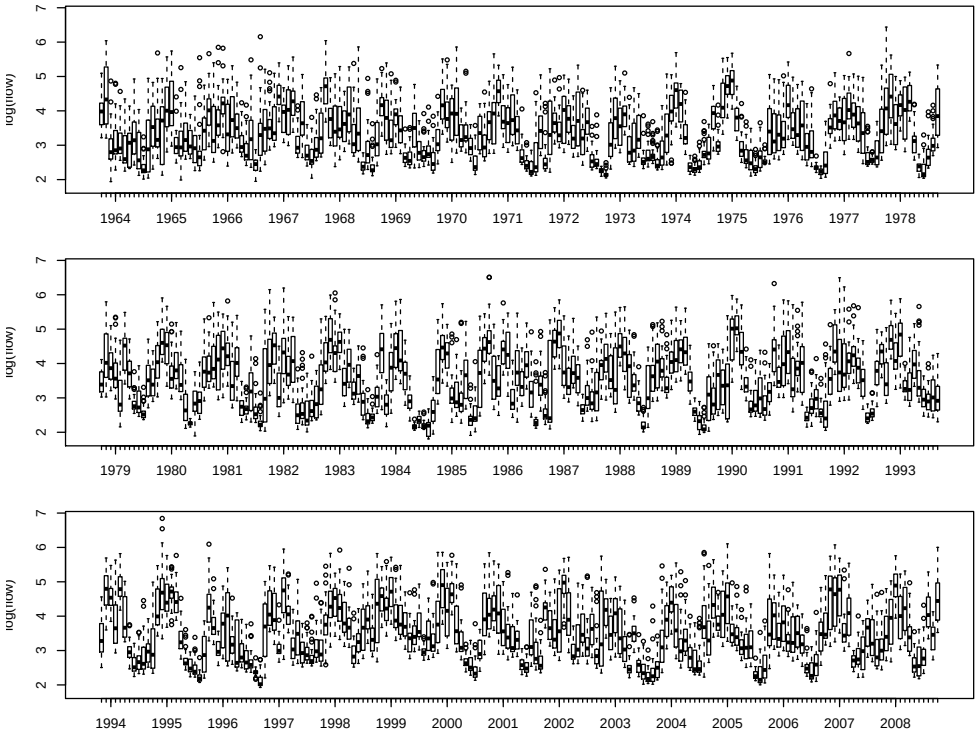

FIGURE 2.12: Monthly boxplot of river flow ($\log(m^3/s)$). River Clyde (Station 84013)

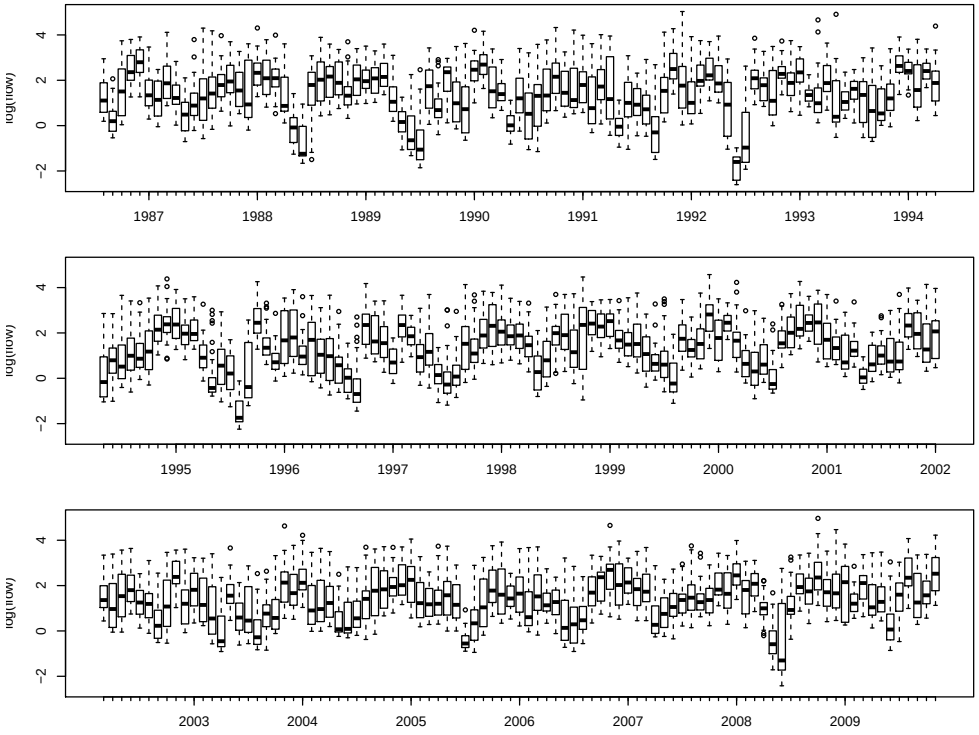

FIGURE 2.13: Monthly boxplot of river flow ($\log(m^3/s)$). Water of Minnoch (Station 81006)

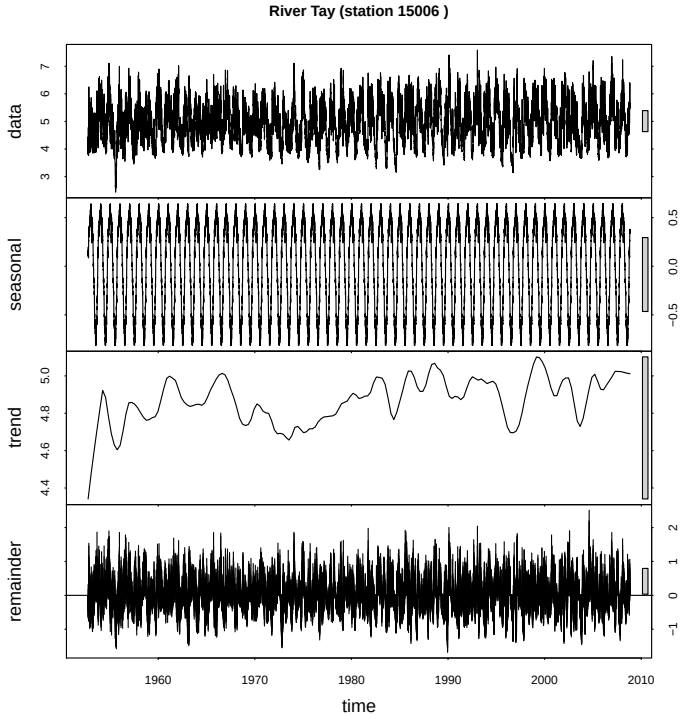

FIGURE 2.14: STL decomposition of daily river flow ($\log(m^3/s)$). River Tay (Station 15006)

Figure 2.15 shows the trends obtained from the stl decomposition for all eight rivers. The trend of River Tweed and Water of Leith shows a marked decrease around July 1973 ((a) in Figure 2.15); this decrease can also be seen in the River Lossie, although earlier in time, and in the River Clyde although in this case it is not as pronounced. The River Lossie and Water of Leith show an increasing trend from 1973 until the beginning of 1986; this increasing trend, although less steep, is also found in rivers Tay and Tweed. The Water of Leith reaches its maximum in March 1986 ((b) in Figure 2.15), when the trend peaks. Rivers Tay and Tweed also show high values at this time point, and the River Lossie peaks slightly earlier. On the other hand, the trend of rivers Ewe and Ness reaches its peak in August 1992 ((c) in Figure 2.15). This peak is followed by a marked decrease in March 1996 ((d) in Figure 2.15) and a peak in January 1999 ((e) in Figure 2.15) common to all western rivers (Ewe, Ness, Clyde, Water of Minnoch) and also present but less pronounced in the eastern rivers (Lossie, Tay, Water of Leith and Tweed). The trend of rivers Ewe and Ness reaches its minimum in January 2003 ((f) in Figure 2.15). A decrease, although smaller than for the former, can also be found, slightly later in time, for the remaining rivers, being more pronounced for rivers on the

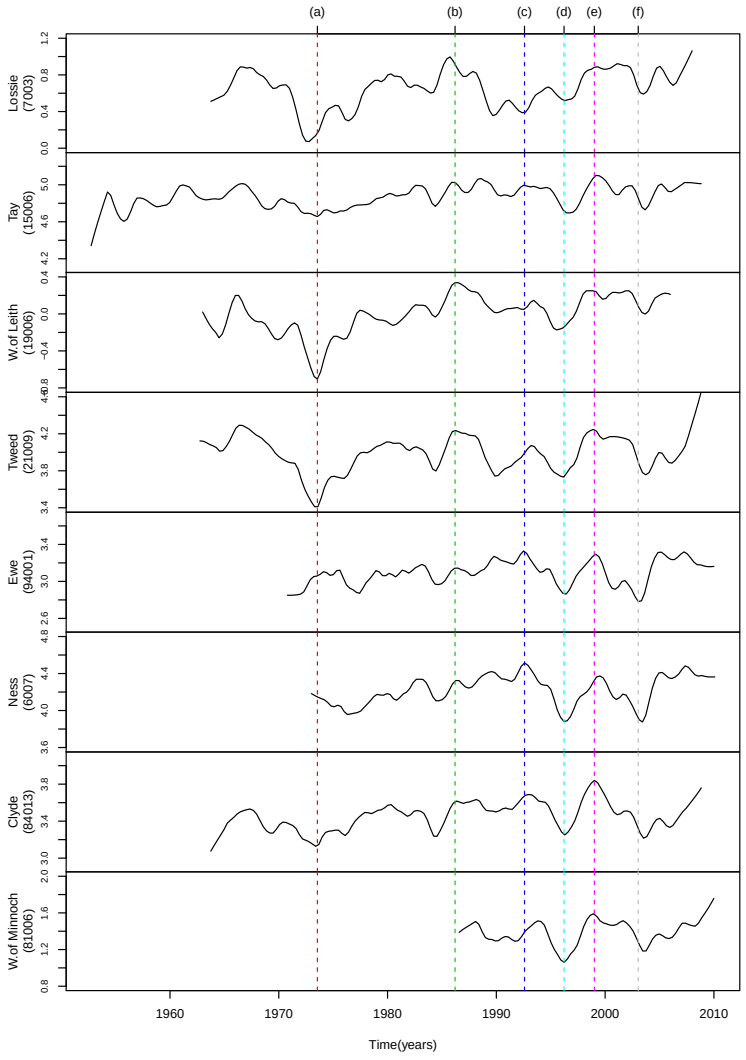

FIGURE 2.15: Trends from stl decomposition for daily river flow series ($\log(m^3/s)$) of all eight rivers. Note the scale on the y axis varies across graphs. Reference lines (a), (b), (c), (d), (e) and (f) has been added to highlight particular features of the data. These are referred to in the text

West than on the East. Overall, trends on the East show greater variability than trends on the West. The Water of Leith and the River Tweed show fairly similar trends, and so do rivers Ewe and Ness. River Lossie shows a similar pattern to that of the Water of Leith and Tweed, although features (a) and (b) in Figure 2.15 tend to happen slightly earlier in time. These three rivers are in the East of Scotland, although River Lossie is in the North while the remaining two are in the South.

The stl decomposition of all eight rivers shows a seasonal cycle repeated over the years. However, as was already pointed out previously by looking at the monthly boxplots,

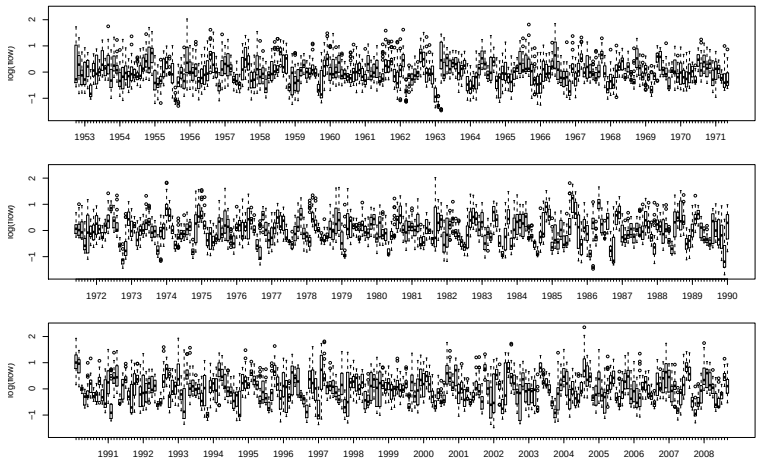

FIGURE 2.16: Monthly boxplot of residuals from stl decomposition ($\log(m^3/s)$). River Tay (Station 15006)

a constant seasonal cycle is very unlikely to be the case. To make sure no seasonality was left in the residuals, a monthly boxplot of those was produced and it is shown in Figure 2.16. Only one figure is included here, as the remaining rivers showed similar patterns. It can be seen that the residual series retains a seasonal pattern. The detrended series was also investigated, as sometimes, by removing a seasonal component which might not be there, one might artificially introduce seasonal patterns in the data, but the boxplot looked very similar to the residual based one from Figure 2.16 and is not included here. The sample autocorrelation function of the residual series is plotted in Figure 2.17. If the stl decomposition of the model was adequate, such a strong pattern should not be present in the residuals. Alternatively, the stl decomposition might be appropriate, but there is long range dependence in the residuals (see Section 2.4).

As an alternative way of identifying the seasonal component of the series, a sinusoidal model was fitted. In order to investigate the presence of any cycles, the periodogram was plotted (Figure 2.18). Prior to doing so, a trend was removed by fitting a 'loess' line, as trends introduce extremely low frequency components in the periodogram. An initial look at the periodogram reveals that all the peaks are concentrated at very low frequencies. Hence, the plot was restricted to a small range of frequency values (0-0.02) so that peaks can be clearly identified. As can be seen in Figure 2.18, there is a periodic component at frequency 1/365, i.e. one cycle per year. The plot suggests no other consistent cyclic components at any other frequencies. Hence, a sinusoidal model was

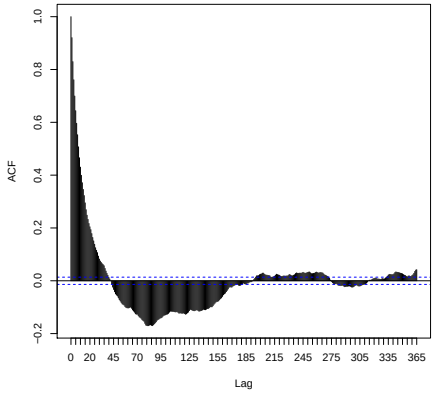

FIGURE 2.17: Sample autocorrelation function of residuals from stl decomposition. River Tay (Station 15006)

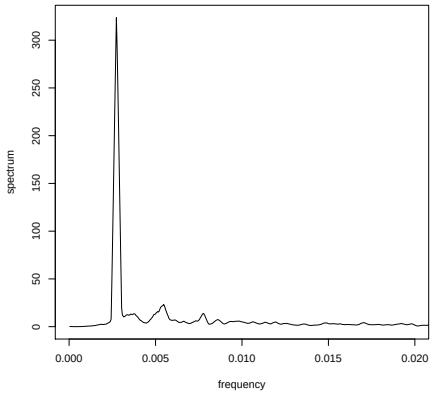

FIGURE 2.18: Periodogram of detrended series for frequencies 0.00-0.02. River Tay (Station 15006)

constructed by fitting the linear model:

$$\log(\text{flow}) = \beta_1 \cos(2\pi \text{time}/365) + \beta_2 \sin(2\pi \text{time}/365) + \varepsilon$$

No intercept was included as it was found to be non significant. Once again, the sample autocorrelation function of the residuals (similar to Figure 2.17) shows not only significant correlation at high lags but a remaining pattern in the data. These results suggest that traditional methods of time series analysis, which assume a constant seasonal component over the years, might not be appropriate for these data.

Time series plots of the monthly variance (Figures 2.19, 2.20) show that the variability of the data is not constant, and hence data are non-stationary an assumption on which many time series modelling techniques are based. Figure 2.17 suggests a complicated correlation structure that cannot be explained with an ARMA(p, q) model. A long-range dependence structure (see Section 2.4) might be appropriate instead.

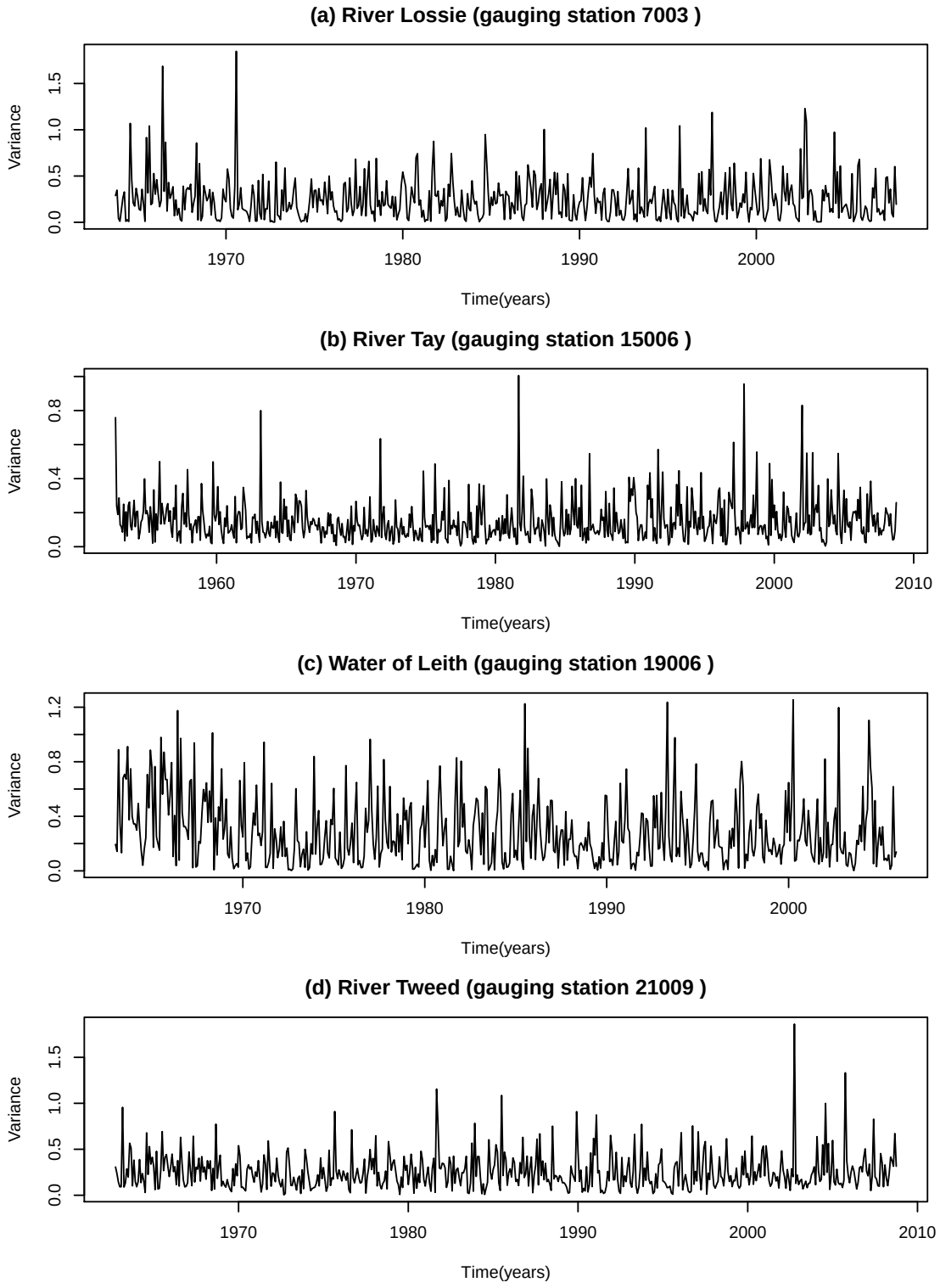

FIGURE 2.19: Time series plot of monthly variance ($\log(m^3/s)$) (eastern rivers). Note the scale on the y axis varies across graphs

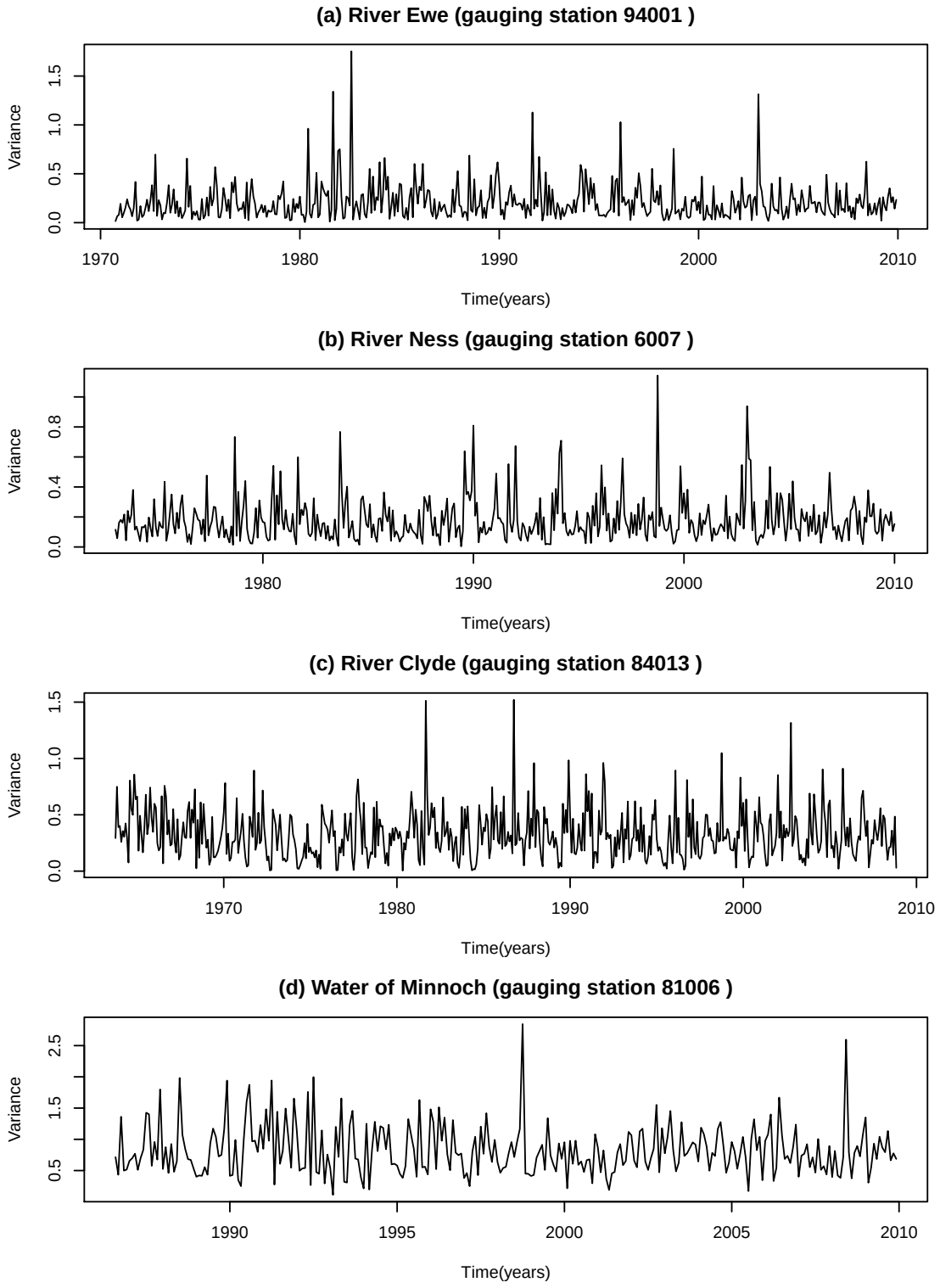

FIGURE 2.20: Time series plot of monthly variance ($\log(m^3/s)$) (western rivers). Note the scale on the y axis varies across graphs

2.4 Long memory

Hydrologic time series are expected to be correlated (Hurst (1951)). Observations close in time tend to be strongly correlated, the strength of the correlation decaying with increasing lag and decreasing frequency. Long range or long memory behaviour appears when observations far apart in time are still correlated (Granger and Joyeux (1980); Hosking (1981); Koutsoyiannis (2002); Doukhan et al. (2003); Beran (1994); Serinaldi (2010)). Even if those correlations are small, they still might have an effect on the time series; it is the simultaneous effect of the correlations that becomes important in a long memory series, rather than the individual correlations on their own (Beran (1994)).

A traditional time series analysis decomposes the response variable as the sum of a trend, seasonal and random noise components. Once the trend and seasonal components of a time series have been identified, the residual correlation can be modelled as an ARMA(p, q) process (Box and Jenkins (1970); Shumway and Stoffer (2006)). The sample autocorrelation and partial autocorrelation functions can be used to get an idea of the orders p and q of the ARMA model, assumed to be stationary. A common technique for dealing with non-stationarity is differencing the series d times, in which case an ARIMA(p, d, q) model is fitted (Box and Jenkins (1970); Shumway and Stoffer (2006)). ARMA processes are short memory processes whose autocorrelation function $\rho(k) \approx c_\rho \theta^k$ decays exponentially with increasing lag k (Doukhan et al. (2003); Beran (1994)), where c_ρ is a positive constant and $|\theta| < 1$. This means that the autocorrelation function has an upper bound $|\rho(k)| \leq a|\theta|^k$ for large k (Beran (1994)) and hence it is summable:

$$\sum_{k=-\infty}^{k=\infty} \rho(k) < \infty$$

Long memory dependence was first identified by Hurst (Hurst (1951)) in daily level data from the river Nile. The strength of long range dependence is measured by the Hurst exponent $H \epsilon$ [0,1], where the case H=0.5 corresponds to independent white noise (Serinaldi (2010); Beran (1994)). Hurst found this value to vary between 0.46 and 0.93 when analyzing different records (Doukhan et al. (2003)). Long memory processes are characterized by an autocorrelation function that slowly decays to zero, as a result of observations far apart in time being correlated, in which case traditional time series

ARIMA models are no longer appropriate. Even though there is general agreement amongst hydrologist that the Hurst phenomenon is "inherent to hydrological time series" (Koutsoyiannis (2002)), it has been largely avoided and ARIMA models have been used instead. In particular, the AR(1), despite being "inconsistent with hydroclimatic reality" (Koutsoyiannis (2002)) is widely used. As Koutsoyiannis (2002) claims, this is probably because understanding and appropriately explaining long memory is more complex that other statistical concepts such as an autoregressive process. Long range dependence has been defined as a "difficult statistical property to work with" (Clegg (2005)), the difficulty lying in going from the theory to the practical application; mathematically the Hurst parameter is well-defined, but it is not easy to measure it in a real data set (Clegg (2005)). An added complication is that of simulating data (Koutsoyiannis (2002)), more difficult than in the ARIMA case. Beran (1994) provides some practical advice that might help identifying a (stationary) long memory process; overall, the series might look stationary, but long periods with high and low levels can be found; when looking at the series as a whole, there does not seem to be any apparent trend but there seem to be cycles or local trends in short periods and the variance of the sample mean, $\text{var}(\overline{x}) \longrightarrow 0$ at a rate $\propto n^{-\alpha}, 0 < \alpha < 1$ (i.e. slower than the normal rate 1/n).

Examples of long memory applications in hydrology include Hurst (1951) and Montanari et al. (1997). From a physical point of view, the reasons underpinning long memory behaviour are not clear and there is great controversy around the subject (Koutsoyiannis (2002); Doukhan et al. (2003)). A number of theories are available; these include nonstationarity (Doukhan et al. (2003)), some sort of Markovian dependence inducing dependence on large scales (Doukhan et al. (2003)), different items of a large system working together is some way (Koutsoyiannis (2002)), and deterministic trends or a natural mechanism inducing long memory behaviour (Koutsoyiannis (2002)). Koutsoyiannis (2002) proposes his own theory, based on absence of memory of the series itself, in which the apparent long memory may be due to irregular changes in the climate. Another possibility would be aggregation of short-memory models that induces an (artificial) long-memory process (Beran (1994); Koutsoyiannis (2002)); this could be the case when the time series analyzed is the result of putting together a collection of short memory series (Beran (1994)).

As Beran (1994) points out, identifying long range dependence and correctly estimating the H parameter is important for statistical inference, as not taking it into account will have a large effect on the bias of the estimated parameters. In particular, long range dependence has a direct effect on the autocorrelation function in the time domain, and equivalently in the spectral density function f(ω), $\omega \epsilon$ [-1/2,1/2], in the frequency domain. The autocorrelation function of a long memory process can be expressed as (Granger and Joyeux (1980); Beran (1994)):

$$\rho(k) \approx c_\rho |k|^{-\alpha} \quad \text{for large } k \tag{2.1}$$

where c_ρ is a positive constant and $0 < \alpha < 1$. This means that even when the lag k is very large, $\rho(k)$ tends to zero very slowly, which translates in the correlations not being summable (Beran (1994)):

$$\sum_{k=-\infty}^{k=\infty} \rho(k) = \infty$$

The corresponding spectral density function (SDF) can be written as (Granger and Joyeux (1980); Beran (1994); Serinaldi (2010)):

$$f(\omega) \approx c_f |\omega|^{-\beta} \tag{2.2}$$

for $-1 < \beta < 3$, $c_f > 0$ and small ω ($\omega \to 0$).

2.4.1 Fractional Gaussian noise, fractional Brownian motion and FARIMA models

Long range dependence can be represented via two stochastic processes, fractional Gaussian noise (fGn) and fractional Brownian motion (fBm) (Serinaldi (2010)). A fractional Gaussian noise would be the equivalent of a white noise process with long range dependence (and hence, despite the dependence structure, it is a stationary process). A fractional Brownian motion can be regarded as the cumulative aggregation of the terms of a fGn series (Serinaldi (2010)). A fractional Brownian motion Y_t is a self-similar process with Hurst parameter $0 < H < 1$. This means that for any positive stretching factor c, the rescaled process $c^{-H} Y_{ct}$ is equal in distribution to Y_t. The stochastic process Y_t is not stationary, but its increments, $X_t = Y_t - Y_{t-1}$ are. The series of increments X_t is a

fractional Gaussian noise with the same Hurst parameter H. Throughout it is assumed that E(X_t)=0.

The autocorrelation function at lag $k \geq 0$ (Koutsoyiannis (2002); Palma (2007); Granger and Joyeux (1980); Taqqu et al. (1995)) of a fGn process with Hurst parameter $H > \frac{1}{2}$:

$$\rho(k) = \frac{1}{2}\left\{|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}\right\} \text{ for } k \in \mathbb{Z}$$

Asymptotically, for large k:

$$\rho(k) \sim H(2H-1)k^{2H-2}$$

Consequently, the spectral density function (Koutsoyiannis (2002); Serinaldi (2010); Taqqu et al. (1995)):

$$f(\omega) = c_f|\omega|^{1-2H}$$

for a positive constant c_f. The fBm related to the fGn is characterized by the same Hurst parameter H.

2.4.1.1 Fractional ARIMA models

Fractionally differenced ARIMA (ARFIMA) models provide a framework that allows incorporation of both short and long memory effects in the model (Doukhan et al. (2003); Beran (1994); Montanari et al. (1997)). The idea is to build models flexible enough so that low-lag autocorrelation can be incorporated as well as long memory effects. An ARFIMA model can be expressed as:

$$\phi(B)(1-B)^d X_t = \psi(B)\varepsilon_t$$

where X_t is a stationary process, B is the backshift operator, defined by $BX_t = X_{t-1}$ $\phi(B)$ and $\psi(B)$ polynomials in B corresponding to the AR and MA parts of the model and ε_t is white noise. Here d (-1/2$\leq d \leq$ 1/2) determines the long-term behaviour, and is related to the Hurst parameter H via the equation $d = H - 1/2$ (Beran (1994); Taqqu et al. (1995)). The terms p and q are the usual ARMA model parameters and relate to the short-range properties of the process. For long memory processes $0 \leq d < 1/2$ results

in an stationary process while for $d \geq 1/2$, the process is not stationary (Doukhan et al. (2003)). The general idea is that once the series has been fractionally differenced using d, an ARMA model can be fitted to the resulting differentiated series. However, this approach requires the series X_t to be stationary, there are problems with parametrization for non-Gaussian data and the asymptotic properties have not been properly investigated for non-Gaussian or non-stationary processes (Doukhan et al. (2003)). FARIMA models are not the objective of this thesis and will not be further discussed. The reader is referred to Doukhan et al. (2003); Beran (1994); Montanari et al. (1997) for additional information.

2.4.2 Hurst parameter estimation

A number of heuristic approaches to estimate the Hurst parameter are available, both in time and frequency domains. Time methods include the R/S, aggregated variance, residuals from regression and Higuchi's methods (Doukhan et al. (2003); Beran (1994); Montanari et al. (1997); Taqqu et al. (1995)), while frequency methods include the periodogram (Taqqu et al. (1995)). More recently, wavelet based estimation (Serinaldi (2010); Clegg (2005); Simonsen and Hansen (1998)) has been proposed. Probability based methods using approximate likelihoods (Whittle's approximation) are available as well (Beran (1994); Taqqu et al. (1995)). R packages devoted to Hurst parameter estimation include `fractal`, `fArma` and `longmemo`. Clegg (2005) compares the R/S, aggregated variance, periodogram, wavelet and Whittle's methods and highlights the weaknesses of some of them, in particular the R/S method, which seems to underestimate H with respect to the remaining methods, for both real and simulated data. Overall, there is lack of agreement amongst the methods. Clegg (2005) also investigates the performance of the methods when the series is perturbed with either AR(1) noise, a seasonal pattern or a trend. All methods perform badly when AR(1) is added, but the wavelet and Whittle's methods do not seem to be affected when a seasonal pattern or a trend are added to the fGn simulated series (Clegg (2005); Simonsen and Hansen (1998)). When performance was investigated with real data, pre-processing the data to remove possible trends and stabilizing the variance did not have much of a influence in the latter two estimates, but it slightly affected the ones in the time domain.

β value	Characteristics	Process
$\beta < 0$	Stationary and anti-persistent	
$\beta = 0$	Stationary and uncorrelated	White Noise
$0 < \beta < 1$	Stationary and weak persistent	fGn
$\beta > 1$	Non-stationary and strong persistent	fBm

TABLE 2.4: Long memory processes classification

Even though both fGn and fBm processes are characterized by the same H parameter, the estimation method used depends on the type of process that the data are assumed to follow. Taqqu et al. (1995) provides a good summary for fGn and FARIMA processes, while Serinaldi (2010) explains how the different methods relate to each other for both fGn and fBm processes. The methods for estimating H under the assumption that the process is fGn rely on the series being stationary, making it necessary to first eliminate the trend and seasonal components and estimate H based on the residuals. In his paper, Serinaldi (2010) points out the importance of identifying the kind of signal (fGn or fBm) prior to parameter estimation. For example, the relationship between the slope β (Equation (2.2)) and H changes depending on whether the process is fGn or fBm. For fGn processes, $\beta\epsilon$(-1,1), while for fBm, $\beta\epsilon$(1,3). The case β=0 corresponds to white noise. Note that $\beta_{fGn} = \beta_{fBm} - 2$ (Serinaldi (2010)). Table 2.4 summarizes the different classes of processes depending on the β value (Serinaldi (2010)).

To assess whether a given time series can be represented as a fGn or fBm process, Serinaldi (2010) suggests first plotting the periodogram in a log-log scale and estimating β according to Equation (2.2). Near the origin, the spectral density function (or its estimate the periodogram I(ω)) should be randomly scattered around a straight line with negative slope (Beran (1994)). This estimate of β can provide a first idea of whether the process is fGn ($\beta\epsilon$ (-1,1)) or fBm ($\beta\epsilon$ (3,5)). Based on this, Serinaldi (2010) provides a series of rules that help deciding whether the process is fGn or fBm using a range of methods, including the R/S, aggregated variance, residuals from regression, Higuchi and wavelets.

There is general agreement that the wavelet estimation method is the most appropriate and least biased amongst the whole range of estimators available, for both fGn and fBm processes, outperforming the periodogram method (Serinaldi (2010); Clegg (2005);

Maxim et al. (2005); Simonsen and Hansen (1998)). The wavelet method will be used in this thesis and is detailed below. Details of alternative methods are not included here. The reader is referred to Doukhan et al. (2003); Beran (1994); Taqqu et al. (1995); Serinaldi (2010); Clegg (2005) for further information.

2.4.3 Wavelet method

Let $W_n(\tau_j)$ be the wavelet transform (see Chapter 3) of the time series x_t at scale τ_j, $j = 1, \ldots, J$ and time n. As stated in Chapter 3 (Section 3.1.4), the variance at scale τ_j can be estimated by (Percival and Walden (2006)):

$$\hat{v}_X^2(\tau_j) = \frac{1}{M_j} \sum_{t=L_j-1}^{N-1} \overline{W}_{j,t}^2 \tag{2.3}$$

where $\overline{W}_{j,t}^2$ is the time series x_t filtered by the chosen wavelet filter (of width L), $L_j = (2^j - 1)(L-1) + 1$ is the width of the j^{th} level wavelet filter and $M_j = N - L_j = 1$ is the number of coefficients not affected by boundary conditions.

The wavelet variance $\hat{v}_X^2(\tau_j) \propto \tau_j^{\beta}$ for both fGn and fBm, with $\beta \epsilon$ [-3,5] (Serinaldi (2010)). For fGn processes, the Hurst parameter $H = (\beta_{fGn}+1)/2$, while for fBm processes, $H = (\beta_{fBm}\text{-}1)/2$. The parameter β can be estimated by plotting the logged wavelet variance versus log(scale) and estimating the slope of the linear relationship.

2.4.3.1 Hurst parameter estimation of daily river flow data

The residuals from the stl decomposition were used to estimate the Hurst parameter using the wavelet based method. An example for the River Tay (gauging station 15006) is shown in Figure 2.21. The Hurst parameter was estimated to be $\hat{H}$=0.79, assuming that the underlying process is fGn and using 5 levels of decomposition in the wavelet transform. This suggests a strong long memory effect, as $\hat{H}$=0.79 is considerably greater than 0.5 (value for which observations are independent). The estimation is based on the residuals from the stl decomposition. The Hurst parameter was also estimated using the original series, in which case $\hat{H}$=0.79 (with 8 levels of decomposition) as well. In

either case, the two first scales, corresponding to 1 and 2 days, were not included in the estimation, as these can be related to short memory effects.

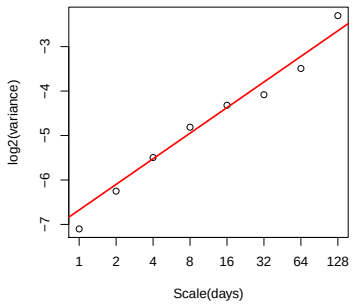

FIGURE 2.21: Plot of log wavelet variance vs log scale for the River Tay (gauging station 15006). The red line shows the fitted linear regression model (including scales of 4 days and above) to estimate the slope β

Table 2.5 summarizes the estimated Hurst parameter for all eight rivers. Both original and residual based estimates suggest a long memory effect in all rivers, although the strength of the dependence varies slightly from river to river. In general, it looks as if large catchments exhibit stronger long range dependence than smaller ones. Original and residual based estimates are fairly similar for all rivers but the Ewe. The presence of long memory will be further investigated in Chapter 4.

Station	River	$\hat{H}$ (original series)	$\hat{H}$ (residual series)
7003	Lossie	0.57	0.56
15006	Tay	0.79	0.79
19006	W. of Leith	0.62	0.58
21009	Tweed	0.72	0.72
94001	Ewe	0.79	1.02
6007	Ness	0.72	0.72
84013	Clyde	0.65	0.62
81006	W. of Minnoch	0.53	0.62

TABLE 2.5: Hurst parameter estimates based on the wavelet method for all eight rivers using the original and residual series

In the previous sections, the main features of Scottish rivers have been explored and the correlation structure, characterized by long range dependence, introduced. Section 2.5 concentrates on the analysis of extreme river flow values. First extreme value theory

is briefly introduced, followed by an application on the eight rivers selected in the exploratory analysis.

2.5 Extreme Value Analysis

Extreme value analysis is based on two approaches, block maxima and peak over threshold series. **Annual maxima** (**AM**) series, a special case of block maxima, are calculated by taking the maximum peak discharge record for every year. Peaks are assumed to be independent. Use of the water year (which depends on the seasonal climatic and flow regimes) instead of the calendar year is recommended (Shaw (1994)). **Points-over-threshold** (**POT**) series are calculated by taking all the observations above a threshold u which has to be defined *a priori*. The resulting series is longer than the AM one, but independence is likely to be an issue. A special type of this kind of series is the **annual exceedance series**, for which the threshold u is chosen such that there are N peaks in the N years of record, but not necessarily one in each year (Shaw (1994)). A range of well known distributions for modelling extreme values are available. The main results and notation used in this section were taken from Coles (2004).

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables with common distribution function F and let $M_n = max\{X_1, X_2, \ldots, X_n\}$ be the block maxima. The "Extremal types theorem" (Coles (2004)) states that if there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$P\left\{\frac{(M_n - b_n)}{a_n} \leq z\right\} \rightarrow G(z) \text{ as } n \rightarrow \infty$$

where G is a non-degenerate distribution function, then G belongs to the generalized extreme value (GEV) family, with distribution function:

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \tag{2.4}$$

defined on $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$. The scale, location and shape parameters (μ, σ, ξ) satisfy that $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$. In particular, the shape parameter $\xi = 0$ corresponds to the Gumbel distribution, while $\xi > 0$ and $\xi < 0$ correspond to the

Fréchet and Weibull distributions respectively.

Extreme quantiles can then be obtained by inverting the cdf:

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \{-\log(1-p)\}^{-\xi}\right] & \xi \neq 0 \\ \mu - \sigma \log\{-\log(1-p)\} & \xi = 0 \end{cases} \tag{2.5}$$

The quantile z_p is defined as a **return level** with return period $1/p$, where $G(z_p) = 1-p$. The **return period** or recurrence interval is defined as the amount of time (on average) until a certain peak z_p is likely to be equalled or exceeded. It can also be defined as the long term average of the intervals between successive exceedances of a flow of magnitude z_p (Shaw (1994); Coles (2004)). Return period estimation is highly dependent on the length and characteristics of the data set. Black and Burns (2002) suggests that for an estimate of the T year return, a record length of 2T years should be appropriate. A return level plot is obtained by plotting z_p against $\log(y_p)$, where $y_p = -\log(1-p)$.

Let X be an arbitrary term in the sequence of X_i, i=1,...,n. The probability that the extreme values (ie, those above the threshold u) exceed the threshold u by y units ($y > 0$) can be calculated as:

$$P(X > u + y | X > u) = \frac{1 - F(u+y)}{1 - F(u)}$$

The distribution F is unknown. However, if F satisfies the "Extremal types theorem" (equation (2.4)), then for sufficiently large u, the distribution of the exceedances $(X-u)$, conditioned on $X > u$, can be approximated by a generalized Pareto (GP) distribution with shape and scale parameters ξ and $\tilde{\sigma} = \sigma + \xi(u - \mu)$:

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi} \tag{2.6}$$

$H(y)$ is defined on $\{y : y > 0, (1 + \xi y/\tilde{\sigma}) > 0\}$.

The choice of the threshold u is arbitrary, but needs to be large enough so that the theory holds. A rule that is used (Shaw (1994)) is to choose u such that there are between 3 and 5 exceedances per year. The mean residual life plot (Coles (2004)) can be used

to graphically assess a suitable range of thresholds, determined by the linearity of the graph. Sensitivity analysis can be used to assess the influence of the threshold choice on parameter estimates.

Return levels can also be obtained. Assuming the GP(ξ, σ) distribution is adequate, then for any $x > u$:

$$P(X > x|X > u) = \left[1 + \frac{\xi(x-u)}{\sigma}\right]^{-1/\xi}$$

which means:

$$P(X > x) = \zeta_u \left[1 + \frac{\xi(x-u)}{\sigma}\right]^{-1/\xi}$$

where $\zeta_u = P(X > u)$. The probability of an individual observation exceeding the threshold u, can be estimated as $\hat{\zeta}_u = \frac{k}{n}$, where k is the number of excesses and n the total number of observations. The m-observation return level x_m in this case is the value that is exceeded on average once every m observations, and hence is the solution to the equation $P(X > x_m) = \frac{1}{m}$. Return levels can the estimated as:

$$x_m = \begin{cases} u + \frac{\sigma}{\xi}\left[(m\zeta_u)^{\xi} - 1\right] & \xi \neq 0 \\ u + \sigma \log(m\zeta_u) & \xi = 0 \end{cases}$$

The N-year return level, defined as the value z_N that is expected to be equalled or exceeded on average once every N years, is defined as:

$$z_N = u + \frac{\sigma}{\xi}\left[(Nn_y\zeta_u)^{\xi} - 1\right]$$

where n_y is the number of observations per year (365 for river flow data). Confidence intervals can be produced using the delta method or profile likelihood (Coles (2004)). The latter is preferred for being more conservative (Coles (2004)).

2.5.1 POT modelling of daily river flow data

Extreme river flow values for the eight rivers described in Section 2.3.1 were analyzed by means of POT series. POT series are a more efficient way of using the data available, as they contain more information than the annual maxima series. Given the differences

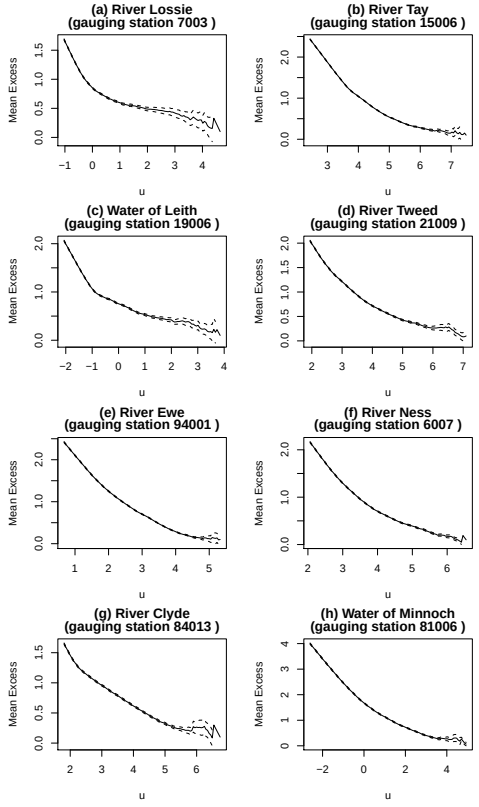

FIGURE 2.22: Mean residual life plots

between catchments and hence flow values, a common threshold cannot be used for all rivers. Mean residual life plots were produced for each river (Figure 2.22). Stability of model parameters was assessed over a range of thresholds. The resulting estimates can be seen in Figures 2.23 and 2.24. After inspection of Figures 2.22-2.24, the thresholds reported in Table 2.6 were selected. Model parameters were estimated using the R package `ismev`. 95% confidence intervals for $\hat{\xi}$ were calculated using the profile likelihood to assess the significance of the shape parameter, which was found to be significant in all cases except for gauging station 81006, where the 95% confidence interval contained zero. The resulting estimates (and their standard errors) are summarized in Table 2.6. The percentage of the data that the excesses represent varies from river to river. For gauging station 81006 (Water of Minnoch), in particular, it is considerably lower (2.94%) than for the rest of the rivers.

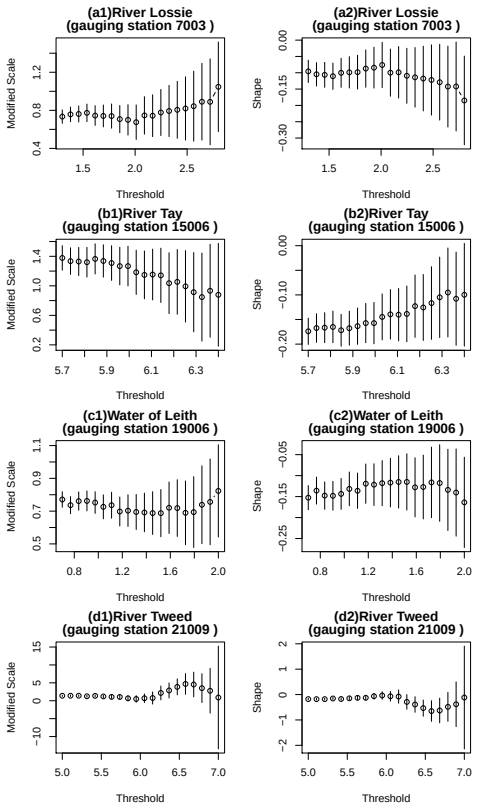

FIGURE 2.23: Generalized Pareto parameter estimates against threshold (eastern rivers). Note the scale on the y axis varies across figures

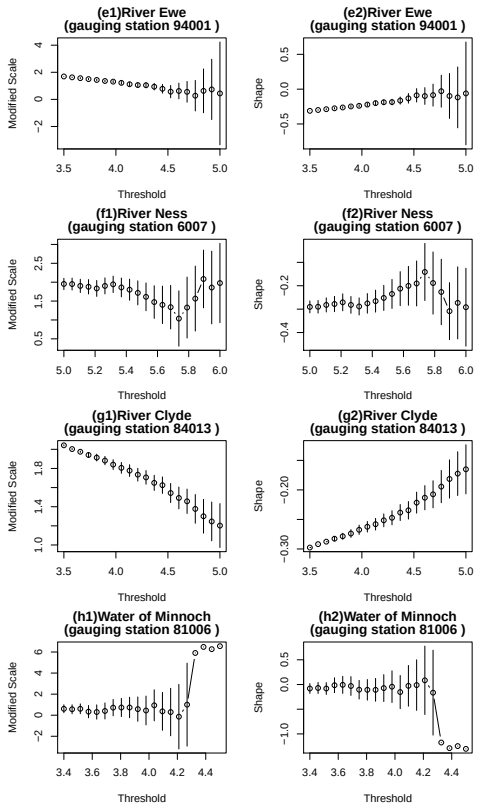

FIGURE 2.24: Generalized Pareto parameter estimates against threshold (western rivers). Note the scale on the y axis varies across figures

Station	River	Threshold u ($\log(m^3/s)$)	Number of excesses	% of data	$\hat{\sigma}$	$\hat{\xi}$
7003	Lossie	2	851	5.27	0.533(0.026)	-0.086(0.034)
15006	Tay	6	1210	5.91	0.324(0.012)	-0.157(0.022)
19006	W. of Leith	1.2	1412	9.00	0.554(0.020)	-0.118(0.025)
21009	Tweed	5.2	1491	8.86	0.464(0.015)	-0.177(0.020)
94001	Ewe	4.2	1208	8.44	0.266(0.009)	-0.186(0.018)
6007	Ness	5.3	1089	8.05	0.410(0.014)	-0.285(0.019)
84013	Clyde	4.9	1149	6.98	0.400(0.014)	-0.171(0.019)
81006	W. of Minnoch	3.5	251	2.94	0.230(0.027)	-0.037(0.065)

TABLE 2.6: Summary of GP models for river flow data

2.5.1.1 Goodness of fit

Adequacy of the fitted models can be assessed graphically by comparing the empirical and fitted distributions and the empirical and fitted quantiles. If the model is adequate, a plot of empirical vs fitted values should be close to a straight line. A histogram of the data with the fitted density function superimposed can also be used to assess goodness of fit. These, along with return level plots, are shown in Figures 2.25-2.26. Overall, the models are a reasonable fit, although most of them show some divergence towards the upper tail.

2.5.1.2 Return levels

The parameters of the GP distribution are not the main interest when performing an extreme value analysis, but rather the return level estimates that can be obtained using the fitted model. 100-year return levels for the eight rivers are shown in Figure 2.27, along with 95% confidence intervals. The 100-year return level, or the value that is expected to be equalled or exceeded with a probability of 0.01 in any given year, is considered as a "medium probability" flood in the Flood Risk Management Act (2009). The resulting estimates were transformed back to the original scale and are summarized in Table 2.7. The confidence intervals for most of the rivers are quite wide, suggesting a high degree of uncertainty. Bankfull levels, defined as the flow value at which the river level reaches the top of its banks, are included as well for comparison. These were obtained from Marsh and Hannaford (2008). For gauging stations 7003, 15006, 94001, 6007 and 84013 the bankfull level is below the lower limit of the 95% confidence interval, while for gauging stations 19006, 21009 and 81006 the bankfull level is included in the interval. This suggest that the 100-year event might lead to flooding in the former set of rivers, but may not in the latter. Stations 7003, 15006, 94001 and 6007 are situated in the North of Scotland. Gauging station 84013 is in the South, but is a much larger catchment than the rest of the catchments in the South studied here.

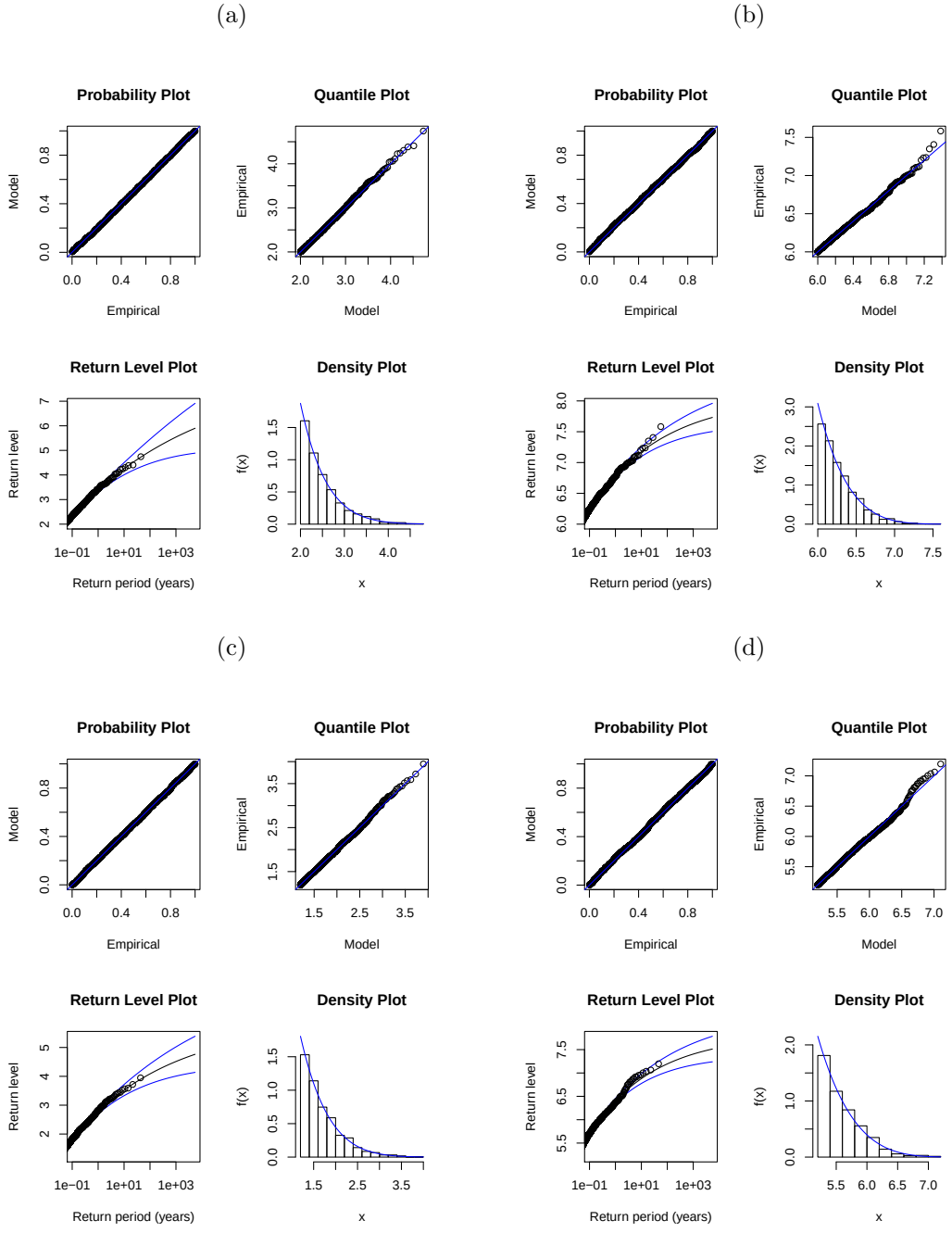

FIGURE 2.25: Diagnosis plots - Rivers (a) Lossie (station 7003), (b) Tay (station 15006), (c) Water of Leith (station 19006) and (d) Tweed (station 21009). For the model to be a good fit, the probability and quantile plots should show a straight line. Points in the return level plot are expected to lie within the confidence bands and the histogram is expected to agree with the fitted density function.

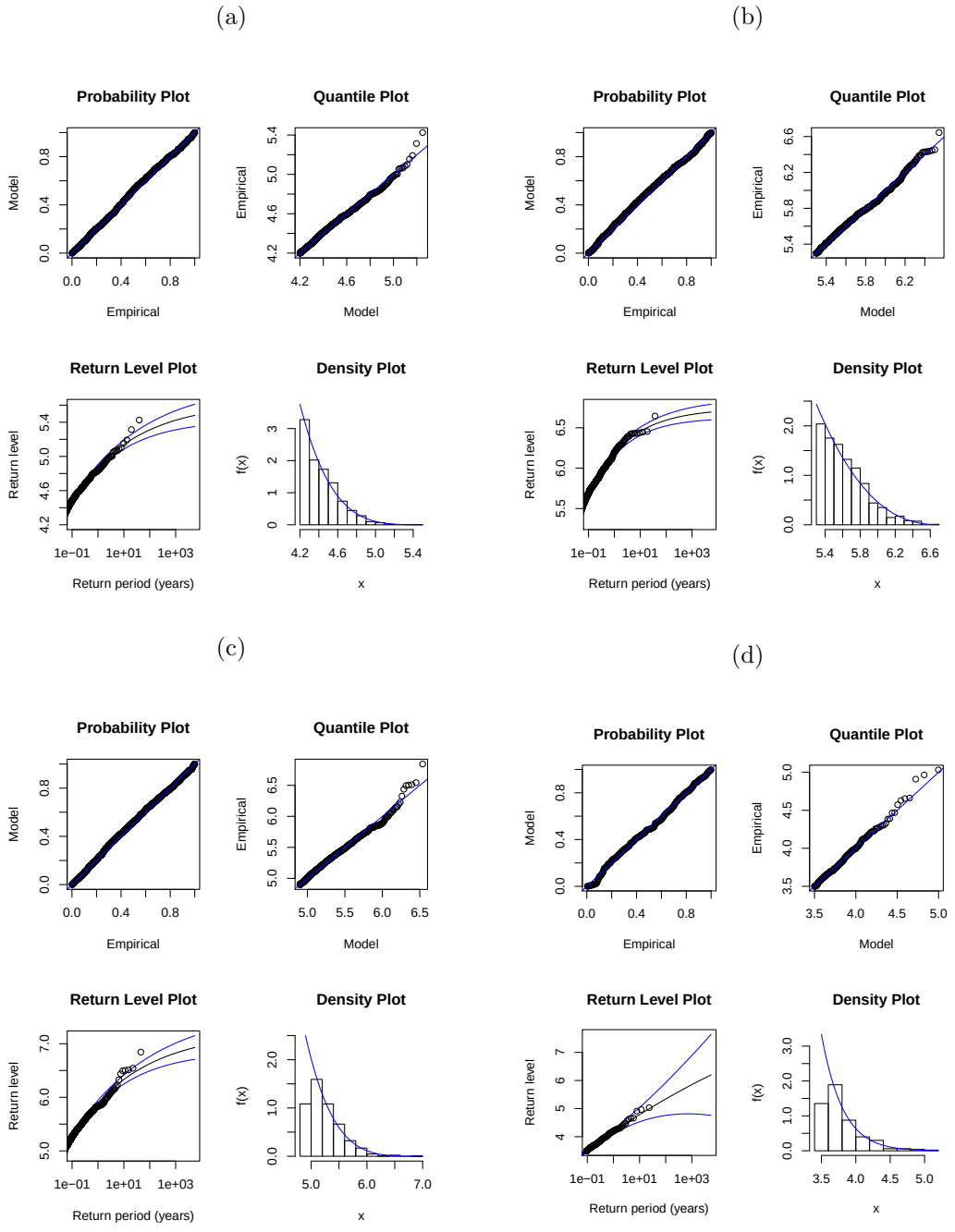

FIGURE 2.26: Diagnosis plots - Rivers (a) Ewe (station 94001), (b) Ness (station 6007), (c) Clyde (station 84013) and (d) Water of Minnoch (station 81006). For the model to be a good fit, the probability and quantile plots should show a straight line. Points in the return level plot are expected to lie within the confidence bands and the histogram is expected to agree with the fitted density function.

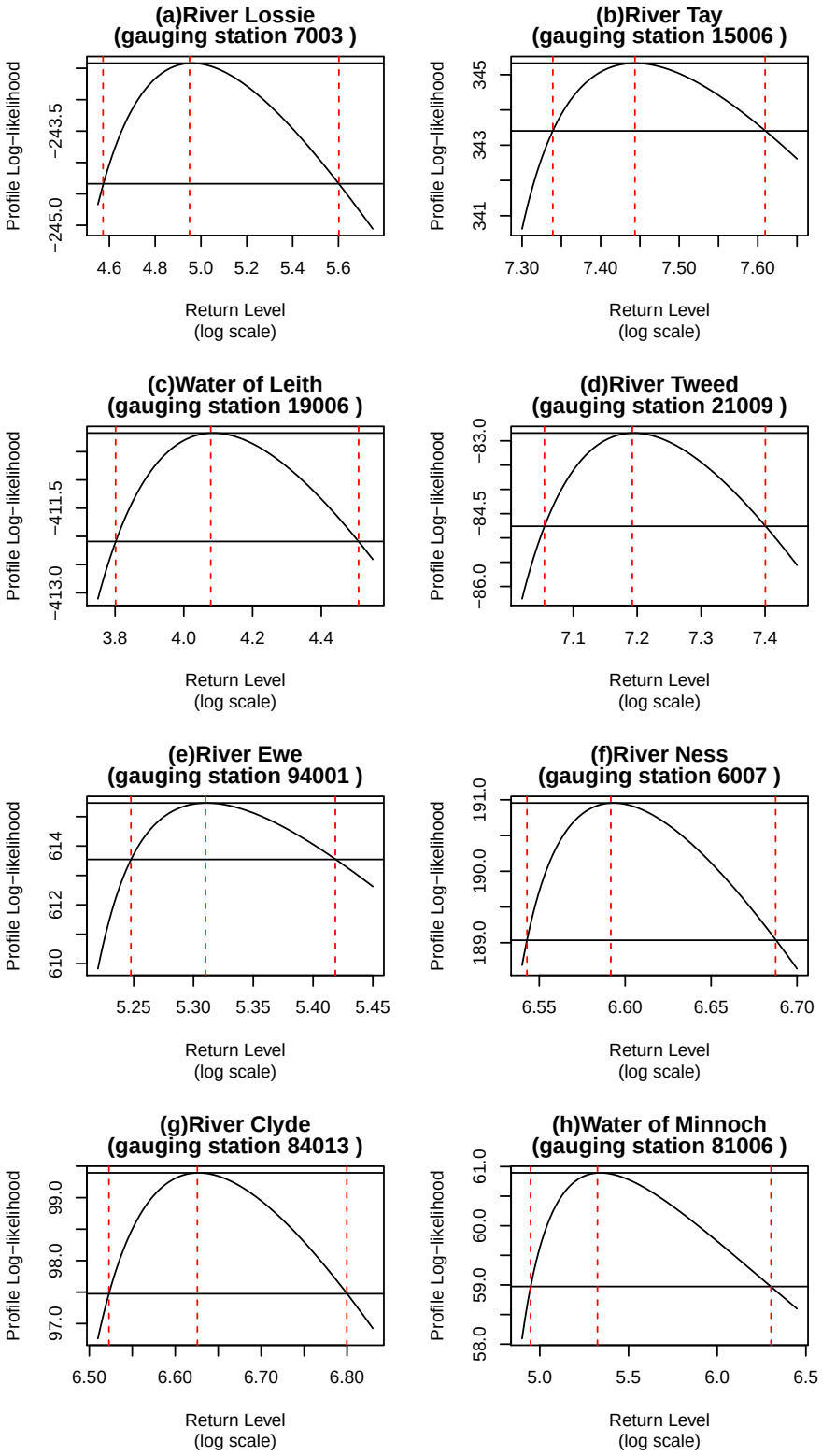

FIGURE 2.27: 100-year return levels. Units are in $\log(m^3/s)$. The black horizontal lines correspond the maximum (top line) and 95% confidence (lower line) values of the profile log-likelihood. The red vertical dashed lines highlight the point estimate of the return level (central line) and 95% confidence interval

Station	River	100-year return value (m^3/s)	95% CI	Bankfull level (m^3/s)
7003	Lossie	141.2	(96.8, 270.9)	35.7
15006	Tay	1709	(1539, 2017)	620
19006	W. of Leith	59.1	(44.8, 90.8)	86
21009	Tweed	1329	(1159, 1637)	1300
94001	Ewe	202.3	(190.1, 225.6)	62.3
6007	Ness	729	(694.3, 802.4)	647
84013	Clyde	754.3	(680.5, 897.6)	370
81006	W. of Minnoch	205.4	(140.9, 546.4)	150

TABLE 2.7: Summary of 100-year return values. Bankfull level source: Marsh and Hannaford (2008)

2.6 Large scale climatic indices

The attention drawn to climate over the last few years coupled with the greater availability of data as a result of the recent development of climate models has led researchers to try to relate patterns in rainfall and river flow to various climate signals. In Europe, the main influence comes from the North Atlantic Ocean, for which two main associated signals have been identified, the North Atlantic Oscillation (NAO) and the Atlantic Multidecadal Oscillation (AMO).

The NAO is a large-scale signal of natural climate variability, calculated as the atmospheric pressure difference (at sea level) between Iceland and the Azores (Hurrell (1995)). The resulting index is positive when the pressure is high in the Azores and low in Iceland, and negative when the situation is reversed. A high index is associated with strong westerly winds, cool summers, mild winters and frequent rain in the north of Europe, while a low index is linked to scarce winds, extreme temperatures and dry conditions (with localized storms) (Macklin and Rumsby (2007); Bouwer et al. (2008)).

Less well known and far less studied being a relatively 'new' concept is the Atlantic Multidecadal Oscillation (AMO) (Kerr (2000); Delworth and Mann (2000)), a further signal of climatic variability, related to anomalies in the sea surface temperature (SST) and sea level pressure (Delworth and Mann (2000)) and calculated based on the former. Similarly to the NAO, the resulting index can be either positive (warm phase) or negative (cold phase). It has been attributed with an oscillation period of about sixty

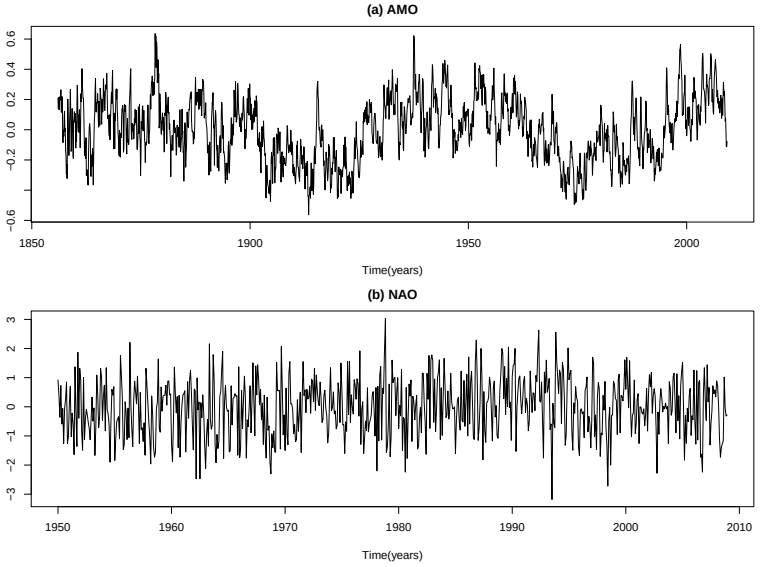

FIGURE 2.28: (a) AMO and (b) NAO indices

years and it is thought to be correlated to air temperatures and rainfall over much of the Northern Hemisphere, in particular, to European summer climate (Knight et al. (2006); Sutton and Hodson (2005)).

The NAO data were downloaded from *http://www.cdc.noaa.gov/data/climateindices/List/* and consists of a monthly series covering fifty-one years (January 1950-December 2008). Data for the AMO were downloaded from *http://www.esrl.noaa.gov/psd/data/timeseries/AMO/* and consists of a monthly series running from January 1856 to April 2009. Both series can be seen in Figure 2.28. The influence of these climatic indices in the set of Scottish rivers selected in Section 2.3.1 are explored in Chapter 3.

2.7 Summary

This chapter describes the main features of river flow data in Scotland. Weather patterns and catchment characteristics have a direct effect on river flow, hence a summary of these two factors was included to help understand the differences across rivers in Scotland. In particular, eight selected rivers covering a range of geographical locations and catchment sizes are discussed in detail. These rivers will be further analyzed in the following chapters. Differences in the long-term term of river flows were found between

the East and the West of Scotland. In particular, trends in the East seem to be slightly more variable than trends in the West.

The data suggest changes in the seasonal pattern over time, with periods where the seasonality is very pronounced and periods where the seasonal cycle is weak. These changes might be related to changes in the weather patterns that affect Scotland and also to the geographical location and catchment characteristics of the rivers. Non-stationarity issues, mainly due to non-constant variability, were identified. The variability of the river flow series will be investigated in detail in Chapter 3 by means of wavelet analysis.

Traditional time series ARMA(p, q) models could not satisfactorily express the complex correlation structure of the data; instead, a long memory correlation structure was preferred. The Hurst parameter, estimated using the wavelet variance, was greater than 0.5 for all eight rivers, suggesting long range dependence.

Extreme value analysis was performed in the form of peak-over-threshold series. Even though the model diagnostic plots suggested a reasonably good fit for all eight rivers, in some cases the estimated return levels were lower than the corresponding empirical ones. Also, the generalized Pareto parameter estimates showed considerable variability across thresholds for some of the rivers. An alternative way of studying extreme values is discussed in Chapter 4. Namely, a quantile regression approach is presented.

Chapter 3

Wavelet Analysis

The exploratory analysis presented in Chapter 2 suggests that the river flow series analyzed are non-stationary, mainly related to non-constant variability. Wavelet analysis allows to identify the main sources of variability in a time series as well as how the variability changes over time and scale. The chapter is organized as follows. First, an introduction to wavelets theory, both in the discrete and continuous case, is presented. The results from a wavelet analysis applied to monthly maxima for the eight rivers are summarized next and the influence of two large scale climatic indices (NAO, AMO) in river flow is assessed using wavelet coherency.

Time series analysis can be approached from two different perspectives: the time domain (classical time series analysis) or the frequency domain. In hydrology, Fourier analysis (also called spectral analysis) has traditionally been used for the latter. However, most commonly used methods assume that the time series is stationary and that it can be expressed as a "linear superposition of linear, independent and non-evolving cycles" (Labat (2005)), conditions that are rarely met by hydrologic series.

Natural processes tend to be composed of one or more cycles that operate on different time scales or frequencies, e.g. annually, every two years, multi-decadally. Wavelet analysis is a useful tool for non-stationary time series which allows detection of the local behaviour at different frequencies (Percival and Walden (2006)). The result is a time-scale decomposition of the original time series. Wavelet analysis provides an alternative

way of looking at the time series, showing features that would not be visible using other methods (Percival and Walden (2006)). The time series and its wavelet transform are two representations of the same thing. Wavelets have been defined by Labat (2005) as a "sort of microscope with magnification $1/s$ (where s refers to scale) and location given by the parameter n (time)". A wavelet ('small wave') is a real valued function $\psi(\cdot)$ defined over the real axis which integrates to zero and is square integrable (Percival and Walden (2006)). Critically this function is localized in both time and frequency (Percival and Walden (2006); Torrence and Compo (1998)), meaning that we do not need to restrict the analysis to either the time or frequency domain but work with both simultaneously. A further condition (necessary for reconstructing a time series from its wavelet transform) is that of *admissibility*, which translates into its Fourier transform $\Psi(f) = \int_{-\infty}^{\infty} \psi(u) e^{-2\pi i f u} du$ satisfying the condition:

$$0 < C_\psi = \int_0^\infty \frac{|\Psi(f)|^2}{f} df < \infty$$

where C_ψ is a reconstruction constant (see Section 3.2). Both discrete and continuous versions of the wavelet transform are available. The choice between the two depends on the aim of the study. The discrete wavelet transform (DWT) provides a detailed description of the time series at different time scales, while the continuous version (CWT) provides a more general picture. The main theory and notation here have been drawn from Percival and Walden (2006) and Torrence and Compo (1998).

3.1 The discrete wavelet transform (DWT)

Let $\{X_t, t = 1, \ldots, N\}$ be a time series of length N. Assume it is equally spaced, with time between observations δt. Assume also that the length of the time series is a power of 2 ($N = 2^J$). Define the DWT coefficients as $\{W_n : n = 1, \ldots, N\}$, then:

$$\mathbf{W} = \mathcal{F}\mathbf{X}$$

where $\mathbf{W}$ is a $N = 2^J$x1 column vector whose n^{th} element is the n^{th} DWT coefficient W_n, $\mathbf{X}$ is a $N = 2^J$x1 column vector (the time series) and $\mathcal{F}$ is a $N \times N$ real valued matrix such that $\mathcal{F}^T\mathcal{F} = I_N$, constructed based on the chosen wavelet filter (see Section

3.1.1). The first N-1 coefficients $\{W_1, \ldots, W_{N-1}\}$ are called **wavelet coefficients** and they relate to differences across time at various scales. Note that each W_j is not a single element but a series that contains $N/(2\tau_j)$ coefficients which are associated with changes on a scale $\tau_j \equiv 2^{j-1}$, $j = 1, \ldots, J$, and localized in time. The last one, W_N, is called the **scaling coefficient**, and relates to variations in the time series at scales 2^J and higher. These wavelet coefficients will be high or close to zero depending on whether there is considerable variation or not in the series at the corresponding scale and time. The wavelet and scaling coefficients vary with time, so we refer to them as $W_{j,t}$, where j refers to scale (τ_j=2^{j-1}) and t refers to time. The coefficients within each scale are approximately uncorrelated.

The time series $\mathbf{X}$ can be recovered as $\mathbf{X}$=$\mathcal{F}^T\mathbf{W}$, and $||\mathbf{W}||^2 = ||\mathbf{X}||^2$. The **multiresolution analysis**(MRA) of $\mathbf{X}$ is defined as:

$$\mathbf{X} = \mathcal{F}^T\mathbf{W} = \sum_{n=1}^{N} W_n \mathcal{F}_{n\bullet} = \sum_{j=1}^{J} \mathcal{F}_j^T W_j + \mathcal{V}_J^T V_J = \sum_{j=1}^{J} D_j + S_J \tag{3.1}$$

where $_{n\bullet}$ represents the n^{th} row of the matrix $\mathcal{F}$, $\mathcal{F}_j$, $\mathcal{V}_J$ are submatrices of $\mathcal{F}$ and D_j is the **wavelet detail**, a N-dim column vector which contains a time series related to variations in $\mathbf{X}$ at scale τ_j. S_J is the J^{th} level **wavelet smooth** of $\mathbf{X}$, a smooth version of $\mathbf{X}$ which becomes smoother as J increases, and is associated with scales τ_{J+1} and higher. S_J can be considered as the long-term trend of the time series.

Hence, the multiresolution analysis of a time series can be used to identify the long term trend and cyclic components of the series, but also how these vary with time. In order to carry out a wavelet analysis, a wavelet filter has to be chosen, which will then be used to build up the matrix $\mathcal{F}$.

3.1.1 Filtering

The spectral density $f(\omega)$ describes the distribution of the variance of a time series as a function of frequency (ω). To extract a particular signal from it, $f(\omega)$ can be transformed

in a predictable way by using a linear filter. A filter is just a sequence $\{a_t\}$ that can be represented using its discrete Fourier transform (DFT), also called the transfer function:

$$A(\omega) \equiv \sum_{t=-\infty}^{\infty} a_t e^{-2\pi i t\omega}, \qquad -\infty < \omega < \infty$$

The polar representation of this function is:

$$A(\omega) = |A(\omega)|e^{i\theta(\omega)}$$

and its associated squared function is defined as $\mathcal{A}(\omega) = |A(\omega)|^2$. Two special types of filters are of interest for wavelet analysis: zero phase filters and linear phase filters. A **zero phase** filter is a filter such that its phase function $\theta(\omega)$=0 $\forall\omega$. This kind of filter is of interest because events in the filtered and original series can be aligned in time when using them to do the wavelet transform.

Denote the periodized filter $\{a_l^o$, l=0,. . . , N-1$\}$. Its discrete Fourier transform is:

$$A_k^o \equiv \sum_{l=0}^{N-1} a_l^o e^{-2\pi i l k/N} = A(\omega_k) \quad where \quad \omega_k = \frac{k}{N}$$

We can circularly filter a time series $\{X_t\}$ with the periodized filter $\{a_l^o\}$, obtaining:

$$Y_t \equiv \sum_{l=0}^{N-1} a_l^o X_{t-l mod N}, \qquad t = 0, \ldots, N-1$$

The filtered series $\{Y_t\}$ can be rewritten in terms of the inverse DFT:

$$Y_t = \frac{1}{N} \sum_{k=0}^{N-1} A_k^o \mathcal{X}_k e^{2\pi i t k/N}$$

where $\mathcal{X}_k$ is the DFT of $\{X_t\}$ and $A_k^o\mathcal{X}_k$ is the DFT of $\{Y_t\}$. The transfer function of the filter can be expressed in its polar form as $A(\omega)$=$|A(\omega)|e^{i\theta(\omega)}$. If $\{a_l\}$ is a zero phase filter, ie, $\theta(\omega_k)$=0 $\forall k$, then $A(\omega_k)$=$|A(\omega_k)|$, A_k^o=$|A_k^o|$ and therefore:

$$Y_t = \frac{1}{N} \sum_{k=0}^{N-1} |A_k^o| \mathcal{X}_k e^{2\pi i t k/N}$$

This means that $\{X_t\}$ and $\{Y_t\}$ differ in the amplitudes of the sinusoids but not in their phases, which implies that events in $\{Y_t\}$ can be aligned with events in $\{X_t\}$.

The second special case are linear phase filters. A **linear phase** filter is a filter such that its phase function is of the form $\theta(\omega) = 2\pi\omega\nu$ for some real-valued constant ν. Define *circularly* advancing the filter output $\{Y_t\}$ by ν units as:

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

where $\nu \in$ Z, $1 \leq |\nu| \leq N$-1. The effect that circularly advancing $\{Y_t\}$ has on the filter $\{a_l\}$:

$$Y_t^{\nu} = Y_{t+\nu mod N} = \sum_{l=0}^{N-1} a_l^o X_{t+\nu-l mod N} = \sum_{l=-\nu}^{N-1-\nu} a_{l+\nu}^o X_{t-l mod N} = \sum_{l=0}^{N-1} a_{l+\nu mod N}^o X_{t-l mod N}$$

i.e.: advancing $\{Y_t\}$ by ν units corresponds to using a filter $\{a_l\}$ whose coefficients have been advanced circularly by ν units. Let $\{a_l^{\nu} = a_{l+\nu}, l = \ldots,$-1,0,1,$\ldots\}$ be the filter $\{a_l\}$ advanced by ν units. By periodizing this filter to length N, we get the circular filter $\{\mathrm{a}_{l+\nu mod N}^o, l = 0, \ldots, N-1\}$. The transfer function for the advanced filter $\{a_l^{\nu}\}$ is given by $A^{(\nu)}(\omega) = \mathrm{e}^{2\pi i f\nu} A(\omega)$. If $\{a_l\}$ has zero phase, i.e. $A(\omega) = |A(\omega)|$, then $A^{(\nu)}(\omega) = |A(\omega)|e^{2\pi i\omega\nu}$, which means that the phase function of $\{a_l^{\nu}\}$ is given by $\theta(\omega) = 2\pi\omega\nu$ and therefore $\{a_l^{\nu}\}$ is a linear filter. If $\nu \in$Z, then a linear phase filter can be easily shifted to get a zero phase filter (by just shifting it -ν units).

The choice of the filter depends on the characteristics of the time series being studied and on the aim of the analysis. The matrix $\mathcal{F}$ is built up in J steps following a pyramid algorithm. There are two types of filters: wavelet filters and scaling filters. A **wavelet filter** of width L (where L is constrained to be an integer even number) is a high-pass filter formed by an infinite sequence $\{h_l\}$ where there are at most L elements different from zero, the first and last elements are non-zero, and h_l=0 for $0 > l \geq L$. For it to be a wavelet filter, it must satisfy all of the following:

1) $\sum_{l=0}^{L-1} h_l = 0$

2) $\sum_{l=0}^{L-1} h_l^2 = 1$ (unit energy)

3) $\sum_{l=0}^{L-1} h_l h_{l+2n} = \sum_{-\infty}^{\infty} h_l h_{l+2n} = 0 \; \forall n \neq 0 \; \epsilon \; \mathrm{Z}$ (orthogonal to its even shifts)

The **scaling filter** of width L is a low-pass filter formed by a sequence $\{g_l\}$ that can be calculated from the corresponding wavelet filter as:

$$g_l \equiv (-1)^{l+1} h_{L-1-l}$$

and it satisfies:

1) $\sum_{l=0}^{L-1} g_l = \sqrt{2}$ (or -$\sqrt{2}$)

2) $\sum_{l=0}^{L-1} g_l^2 = 1$ (unit energy)

3) $\sum_{l=0}^{L-1} g_l g_{l+2n} = 0$

To calculate the first level wavelet coefficients W_1, the time series X_t is circularly filtered with the wavelet filter $\{h_l\}$ (multiplying by $2^{1/2}$ to preserve energy). Only the values corresponding to odd indices are retained (downsampling):

$$W_{1,t} \equiv 2^{1/2} \tilde{W}_{1,2t+1} = \sum_{l=0}^{L-1} h_l X_{2t+1-l \bmod N} \quad t = 0, \ldots, \tfrac{N}{2} - 1$$

Similarly, the time series is circularly filtered with the scaling filter $\{g_l\}$ to calculate the first level scaling coefficients V_1:

$$V_{1,t} \equiv 2^{1/2} \tilde{V}_{1,2t+1} = \sum_{l=0}^{L-1} g_l X_{2t+1-l \bmod N} = \sum_{l=0}^{N-1} g_l^o X_{2t+1-l \bmod N} \quad t = 0, \ldots, \tfrac{N}{2} - 1$$

W_1 and V_1 are orthonormal sets of length $N/2$ that together constitute a set of N orthonormal vectors (Percival and Walden (2006)). The remaining wavelet/scaling coefficients are calculated in a recursive way, using what is known as the pyramid algorithm. At every subsequent step j, $j = 2, \ldots, J$, V_{j-1} is treated as X_t in the first stage, ie, the elements of V_{j-1} are filtered separately with $\{h_l\}$ and $\{g_l\}$ and the outputs are subsampled by two to obtain W_j and V_j. At every step, the wavelet and scaling coefficients capture the high and low frequency behaviour respectively. For further details on the pyramid algorithm, the reader is referred to Percival and Walden (2006).

3.1.2 Daubechies Filters

Daubechies (1992) defined a particular type of filter that, when applied to a time series, provide a discrete wavelet transform that is easy to interpret. The Daubechies wavelet filters are high-pass filters. Their associated scaling filters are then low-pass. The width of the filter (L) has an effect on the filtered series; as L increases, the squared gain function of the Daubechies scaling filter $\mathcal{G}^{(D)}(\omega)$ converges to the squared gain function of an 'ideal' low-pass filter. Also, the number of sequences $\{g_l, l = 0, \ldots, L-1\}$ whose squared gain function is $\mathcal{G}^{(D)}(\omega)$ increases. This results in a wide variety of possible sequences $\{g_l\}$ to choose from. To decide which one is best to use for a particular time series, Daubechies (1992) proposes various options:

1. Daublets D(L): the *extremal phase* scaling filter $\{g_l^{(ep)}\}$ (also known as minimum delay filter) satisfies:

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

   for any other filter $\{g_l\}$ with the same squared gain function.

2. Symmlets LA(L) ($L \geq 8$): the *least asymmetric* filter $\{g_l^{(la)}\}$ is the sequence $\{g_l\}$ (amongst all those with the same squared gain function) whose phase function is as close as possible to a linear phase filter.

An alternative to Daubechies filters are Coiflet filters, but their use is not as extensive as Daubechies filters, as they are likely to introduce artifacts into the wavelet transform due to their shape; hence, they will not be explored here. For a detailed description of Coiflet filters see Percival and Walden (2006).

3.1.2.1 LA(L) filters

The LA(L) scaling and wavelet phase functions can be considered to be approximately linear:

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

Note that ν is always odd and negative. Hence, to get a zero phase filter, it is enough to circularly advance the filtered series by $|\nu|$ units for the scaling filter and by $|L-1+\nu|$

for the wavelet filter. As a result of these shifts, the wavelet and scaling coefficients will be aligned in time with events in the original time series.

In practice, the filter choice will depend on the objective of the analysis. When applying circular filtering, the time series is assumed to be a portion of a larger periodic sequence with period N. This might induce some problems at the beginning and end of the series; the coefficients which might be affected by this are known as boundary coefficients (and can be easily identified). Note that if the time series has a strong annual component and the length is close to an integer multiple of a year, the assumption of periodicity would be reasonable. Once the filter is chosen, the width L of the filter has to be chosen too. Percival and Walden (2006) recommend use of the smallest L that gives reasonable results, bearing in mind that small values of L (2 to 6) might introduce artifacts in the DWT, while large values of L could decrease the degree of localization of the coefficients as well as increase the number of coefficients influenced by boundary conditions, with a larger computational cost. An LA(L) filter should be used when it is of interest to align the DWT in time with the original series. With respect to the sample size restriction, if it is not proportional to a power of 2, the series can be either padded with zeroes or truncated. Finally, note that the wavelet transform does not necessarily have to be done for all J, but it can be stopped at $J_0 < J$, and the same results apply. An alternative would be to use the maximal overlap discrete wavelet transform (MODWT) instead of the DWT (see Section 3.1.3).

3.1.3 Maximal Overlap Discrete Wavelet Transform (MODWT)

Instead of modifying the time series by either padding or truncating it (and hence losing some information), the maximal overlap discrete wavelet transform (MODWT) can be used on a time series whose length is not a power of two. It has the advantages of not being restrictive in terms of sample size and that neither the starting point of the series has an influence on the results, nor is the choice of the filter crucial, as any circularly shifted version of the original series results in the same MODWT decomposition. The cost for these advantages is loss of orthogonality (it is a highly redundant transformation) and a higher computational cost. The wavelet and scaling filters are now defined by $\{\tilde{h}_l = h_l/\sqrt{2}\}$ and $\{\tilde{g}_l = g_l/\sqrt{2}\}$. The process is the same; the time series is circularly

filtered at each step of the pyramid algorithm, only that now there is no downscaling involved. The original time series can still be recovered as the sum of a number of detail and smooth components:

$$X = \sum_{j=1}^{J_o} \tilde{D}_j + \tilde{S}_{J_o} \tag{3.2}$$

where J_0 is the maximum level of decomposition and $\tilde{D}_j$, $\tilde{S}_{J_o}$ are associated with zero phase filters (and therefore, aligned with the events in the original series).

The DWT theory can be easily adapted to the MODWT. Throughout this thesis, the MODWT will be used instead of the DWT.

3.1.4 The Wavelet Variance

Consider a stochastic process $\{X_t \,/\, t = \cdots, -1, 0, 1, \cdots\}$ and filter it using the MODWT wavelet filter $\{\tilde{h}_{j,l}\}$ (associated with scale τ_j). Denote the filtered series by $\overline{W}_{j,t}$:

$$\overline{W}_{j,t} = \sum_{l=0}^{L_j-1} \tilde{h}_{j,l} X_{t-l}, \quad t = \cdots, -1, 0, 1, \cdots$$

where $L_j = (2^j - 1)(L - 1) + 1$ is the width of the j^{th} level wavelet (or scaling) filter. The (time-independent) wavelet variance for scale τ_j is defined as the variance of the wavelet coefficients at that scale:

$$v_X^2(\tau_j) = var\{\overline{W}_{j,t}\}$$

There are two main reasons for interest in the wavelet variance. The first one is that it provides a decomposition of the time series variance (σ_X^2) on a scale by scale basis:

$$\sum_{j=1}^{\infty} v_X^2(\tau_j) = \sigma_X^2$$

Note this is similar to what the spectral density f(ω) does, only wavelet coefficients are easier to calculate. The variance at scale τ_j can be estimated by:

$$\hat{v}_X^2(\tau_j) = \frac{1}{M_j} \sum_{t=L_j-1}^{N-1} \overline{W}_{j,t}^2 \tag{3.3}$$

where $M_j = N - L_j + 1$ is the number of coefficients not affected by boundary conditions. This guarantees the estimator to be unbiased. As $N \to \infty$, this estimator is asymptotically normally distributed with mean $v_X^2(\tau_j)$ and large sample variance $2A_j/M_j$, were $A_j = \int_{-1/2}^{1/2} f_j^2(\omega)d\omega$ and f is the spectral density function. A confidence interval based on this might include negative values. To avoid this problem, renormalization is done to work with a χ^2 distribution instead. The following approximation is used:

$$\frac{\eta \tilde{v}_X^2(\tau_j)}{v_X^2(\tau_j)} \sim \chi_\eta^2$$

where $\eta = \frac{2(E\{\tilde{v}_X^2(\tau_j)\})^2}{var\{\tilde{v}_X^2(\tau_j)\}}$ are the "equivalent degrees of freedom" (EDF) (Percival and Walden (2006)).

In practice, the degrees of freedom are estimated as either $\hat{\eta}_1 = \frac{M_j \hat{v}_X^4(\tau_j)}{\hat{A}_j}$ or $\hat{\eta}_2 = max\{M_j/2^j, 1\}$, where $\hat{A}_j$ can be estimated using the periodogram. If the number of coefficients is large enough (M_j >128) then $\hat{\eta}_1$ should be a good approximation. For smaller sample sizes, $\hat{\eta}_2$ is recommended. There is a third option, $\hat{\eta}_3$, but it involves some knowledge of the shape of the spectral density function. The corresponding CI is calculated as:

$$\left[\frac{\eta \hat{v}_X^2(\tau_j)}{Q_n(1-p)}, \frac{\eta \hat{v}_X^2(\tau_j)}{Q_n(p)}\right]$$

where $Q_n(p)$ is the p^{th} quantile of a χ_η^2.

It follows that the sample variance σ_X^2 can be estimated as:

$$\hat{\sigma}_X^2 = \sum_{j=1}^{\infty} \hat{v}_X^2(\tau_j) \tag{3.4}$$

This is of interest as the usual estimator of the sample variance is biased when the process mean is unknown, while now the mean of the filtered series $\overline{W}_{j,t}$ is known a priori (zero). Note that if the width of the filter is too small the estimate of the sample variance based on the wavelet variance might provide misleading estimates.

The results stated so far assume the variability to be constant over time, which might not be case, specially if the time series is not stationary. The time-dependent wavelet

variability for scale τ_j can be estimated as a running average of N_S points:

$$\hat{v}_{X,t}(\tau_j) = \frac{1}{N_S} \sum_{u=-(N_S-1)/2}^{(N_S-1)/2} \widetilde{W}_{j,t+|\nu_j^{(H)}|+umodN} \tag{3.5}$$

where $\widetilde{W}_{j,t+|\nu_j^{(H)}|+umodN}$ is just the vector $\widetilde{W}_j$ circularly advanced by $|\nu_j^{(H)}|$ units so that events are aligned with the original time series. The choice of N_S depends on the nature of the data; e.g, if we are working with daily data, a natural choice may be be N_S=30. Confidence intervals for the time dependent wavelet variability can be calculated similarly to the time independent case.

The second reason is linked to long-memory processes (see Chapter 2, Section 2.4). The relationship $v_X^2(\tau_j) \propto \tau_j^{-\alpha-1}$ can be used to investigate the existence of long-memory behaviour by plotting $\log(v_X^2(\tau_j))$ vs $\log(\tau_j)$. Linear variation suggests long-memory, and the slope of the relationship can be used to estimate α.

3.2 The continuous wavelet transform (CWT)

The theory results stated in this section have been drawn from Percival and Walden (2006), Torrence and Compo (1998), Torrence and Webster (1999) and Labat (2005).

Let $\psi(\cdot)$ be a wavelet function, the continuous wavelet transform (CWT) of a time series $\{X_t\}$ is defined as the convolution of the time series with a scaled and translated version of the chosen (normalized) wavelet function $\psi_o(\eta)$:

$$W_n(s) = \sum_{t=0}^{N-1} x_t \psi * \left[\frac{(t-n)\delta t}{s}\right]$$

where * indicates the complex conjugate, δt is the time spacing of the series $\{x_t\}$, n denotes time and s denotes scale. Normalization at each scale is done in order to preserve unit energy. One of the most used wavelet functions in time series analysis is the Morlet wavelet, defined as:

$$\psi_0(\eta) = \pi^{-1/4} e^{i\omega_0\eta} e^{-\eta^2/2}$$

where η is a nondimensional time parameter and ω_0 is a nondimensional frequency (scale). In the case of the Morlet wavelet, ω_0=6 (Torrence and Compo (1998)).

The continuous wavelet transform is a two dimensional signal (scale, time) that depends on a one dimensional signal (time), hence it is highly redundant; it is not an orthogonal transformation, which means that the transformed values at adjacent times are highly correlated (Percival and Walden (2006); Torrence and Compo (1998)).

Most of the wavelet functions $\psi(\cdot)$ are complex functions, and so is the corresponding continuous wavelet transform $W_n(s)$. As such, it can be expressed in its polar form as $W_n(s) = |W_n(s)|e^{i\theta}$. The wavelet power spectrum (WPS) is defined as:

$$\text{WPS} = |W_n(s)|^2 \tag{3.6}$$

The wavelet power spectrum provides a measure of the variability of the time series at each time point n and scale s. The expected value of the WPS is $\text{E}[|W_n(s)|^2]=N\text{E}[|\tilde{x}_k|^2]$, where $\tilde{x}_k$ is the discrete Fourier transform of x_t. If x_t is white noise, then $\text{E}[|\tilde{x}_k|^2]=\sigma^2/N$, and therefore $\text{E}[|W_n(s)|^2]=N\sigma^2/N = \sigma^2 \; \forall \; n, s$. This means the WPS can be normalized dividing by σ^2, so that it represents a measure of variability with respect to white noise. This standardization is useful to decide whether the variability in the series is significant or can be considered as natural random variation or to compare different wavelet power spectra (Torrence and Compo (1998)).

As was the case with the discrete wavelet transform, the wavelet function needs to be chosen a priori. Things to be considered when making this decision are the width and shape of the function, as well as the range of scales to investigate. The width of the wavelet function will have an influence on the resulting CWT, in the form that a narrow function (time wise) will have good time resolution but poor scale resolution, and vice versa for a broad function. The ideal choice is to find a balance between time and scale resolution, although this mainly depends on the interest of the study. The shape choice of the function depends on the features of the time series, although it does not

really affect the qualitative results obtained from the WPS. A further choice to make is which range of scales to look at. Torrence and Compo (1998) suggest looking at scales $s_j{=}s_0 2^{j\delta j}$, $j = 0, 1, \ldots, J$, where s_0 is the smallest resolvable scale, δj is the difference between adjacent scales and $J = \delta j^{-1} log_2(N\delta t/s_0)$ is the largest scale. The choice of δj depends on the width of the wavelet function. Similarly to the discrete case, the Fourier transform assumes the data to be cyclic, which means that one should be careful when interpreting the results at the beginning and end of the series; to highlight the regions that might be affected by edge effects, the cone of influence is usually plotted along with the WPS. A final observation to be made is that the wavelet scale (s) does not exactly correspond to the Fourier frequency (λ). However, for the Morlet wavelet (which is the one that the analysis presented in this thesis is based on) the relationship is $\lambda = 1.03s$ and therefore they can be considered to be roughly the same.

Similarly to the discrete transform, the time series can be reconstructed from its continuous wavelet transform. While for the discrete case it was fairly straightforward, now it is a bit more complex given the redundancy in time and scale. For details on how to calculate the reconstructed series, see Torrence and Compo (1998). The discrete transform preserves energy and so does the continuous transform. That means the sample variance σ^2 can be estimated from the CWT as:

$$\hat{\sigma}^2 = \frac{\delta j \delta t}{C_\delta N} \sum_{n=0}^{N-1} \sum_{j=0}^{J} \frac{|W_n(s_j)|^2}{s_j} \tag{3.7}$$

where C_δ is a (wavelet function specific) reconstruction constant (C_δ=0.776 for the Morlet wavelet).

3.2.1 Significance Testing

By choosing a background spectrum and assuming that different realizations of the process will be distributed about it, a significance test can be developed (Torrence and Compo (1998)). For geophysical phenomena, an appropriate choice appears to be either white noise, which has a flat Fourier spectrum or an AR(1), whose Fourier spectrum is characterized by increasing power with decreasing frequency. Once a theoretical wavelet power spectrum has been chosen, a hypothesis test for the significance of a peak in the

spectrum can be established. While for the Fourier transform a theoretical spectrum can be derived, it is not possible for the wavelet transform; hence significance testing is usually based on MC simulations. However, Torrence and Compo (1998) suggest a theoretical approximation.

Assume that the time series $\{X_t\}$ is an AR(1), $X_t = \rho X_{t-1} + Z_t$, where ρ is the lag 1 autocorrelation and $Z_t \sim$ white noise. For this model, the discrete Fourier power spectrum (after normalizing):

$$P_\omega = \frac{1-\rho^2}{1+\rho^2-2\rho\cos(2\pi\omega/N)} \qquad \omega = 0, \cdots, N/2 \tag{3.8}$$

The continuous wavelet transform can be seen as a series of bandpass filters applied to the original time series. Assuming the time series to be an AR(1) process, the local wavelet power spectrum (where 'local' refers to a set period of time, shorter than the total length of the record) is given by P_ω; this is supported by results from a simulation study in Torrence and Compo (1998), where the local wavelet power spectra of 100000 simulated AR(1) series were compared with the Fourier spectrum, being identical on average.

The null hypothesis H_0 is defined by Torrence and Compo (1998) (pag. 69) as follows: "assume that the time series has a mean power spectrum given by P_ω. If a peak in the wavelet power spectrum is significantly above the background spectrum P_ω, then it can be assumed to be a true feature with a certain %confidence."

Assuming the wavelet coefficients to be normally distributed, the wavelet power spectrum $|W_n(s)|^2$ follows a χ^2_2. If the mean background spectrum is P_ω (Equation (3.8)), then the distribution for the local wavelet spectrum at each time n and scale s is:

$$\frac{|W_n(s)|^2}{\sigma^2} \sim \frac{1}{2} P_\omega \chi^2_2$$

Here ω is the Fourier frequency corresponding to scale s. Note that the distributional result is independent of the wavelet function chosen. Inside the cone of influence, the distribution is still χ^2_2, but the results need to be interpreted carefully if the series has

been padded with zeroes. Something to remember when using this significance test is that, if the confidence is say 95%, it means that still 5% of the wavelet power will be above the background spectrum by chance. A non-random distribution of the significant areas indicates that they are likely to be true features of the data and not just random noise.

Even though some authors claim that the significance test should be non-stationary, Torrence and Compo (1998) provide several reasons for which a stationary test like the one they developed is preferable. First of all, the assumption of stationarity allows detection of non-stationarity. The test is robust, as it does not depend on the wavelet function chosen, and based on a known distribution (χ^2). Also, many wavelet transforms of real data are similar to transforms of AR(1) processes. Torrence and Compo (1998) recommend use of a stationary test like this one at first, and further development if necessary for the time series under study.

3.2.2 Smoothing in time and scale

The wavelet power spectrum can be smoothed to increase the degrees of freedom and improve confidence. The smoothing can be done in time or scale domain (or both).

3.2.2.1 Time

Each vertical slice of the wavelet spectrum plot is a measure of the local spectrum for a particular period of time. The ***time averaged wavelet spectrum*** over a certain period (n_1, n_2) is defined as:

$$\overline{W}_n^2(s) = \frac{1}{n_a} \sum_{n=n_1}^{n_2} |W_n(s)|^2$$

where n is the midpoint between n_1 and n_2 and $n_a = n_2 - n_1 + 1$. By doing this at each time step, we get a wavelet plot smoothed by the chosen window.

The ***global wavelet spectrum*** (GWS):

$$\overline{W}^2(s) = \frac{1}{N}\sum_{n=0}^{N-1}|W_n(s)|^2 \tag{3.9}$$

is just the average over all time points N. This is an unbiased and consistent estimator of the true power spectrum (Percival (1995)). By smoothing the wavelet spectrum in time, we can increase the degrees of freedom of each point and the significance of peaks. To determine the degrees of freedom, we need the number of independent points. For the Fourier spectrum, the power at each frequency is independent of the others. The average of the power at M frequencies, each frequency with 2df, is χ^2_{2M} (Spiegel (1975), as reported in Torrence and Compo (1998)). However, for the time averaged wavelet spectrum, the points we are averaging, which are χ^2 distributed, are not independent, but correlated in both time and scale, with correlation in time increasing as scale increases and the wavelet function broadens. If we have v degrees of freedom, and n_a is the number of points we average over, we would expect v $\propto n_a$, and v $\propto 1/s$. An approximation proposed by Torrence and Compo (1998) is v=$2\sqrt{1+(\frac{n_a\delta t}{\gamma s})}$, where γ is a wavelet specific decorrelation factor (γ=2.32 for the Morlet wavelet (Torrence and Compo (1998))).

3.2.2.2 Scale

The ***scale averaged wavelet power*** over scales (s_{j_1},s_{j_2}) is defined as:

$$\overline{W}^2_n = \frac{\delta j \delta t}{C_\delta}\sum_{j=j_1}^{j_2}\frac{|W_n(s)|^2}{s_j} \tag{3.10}$$

which provides a measure of how the variability changes over a particular range of scales. Comparing this expression to the expression of σ^2 (Equation 3.7), it can be seen that the scale-averaged power is simply a time series of the average variance within a certain band. Hence, it can be useful for examining variations in the frequencies of two time series, or even changes in frequency within the same time series (Torrence and Compo (1998)).

As with time-averaging, the degrees of freedom increase by smoothing the series in scale direction. To derive the degrees of freedom, first we normalize the wavelet power by the expectation value for white noise, which is $\frac{\delta j \delta t \sigma^2}{C_\delta S_{avg}}$, where σ^2=var(x_t), and $S_{avg} = (\sum_{j=j_1}^{j_2} \frac{1}{s_j})^{-1}$.

Let n_a = j_2 - j_1 + 1 be the number of scales, then:

$$\frac{C_\delta S_{avg}}{\delta j \delta t \sigma^2} \overline{W}_n^2 \sim \overline{P} \frac{\chi_v^2}{v}$$

where $\overline{P} = S_{avg} \sum_{j=j_1}^{j_2} \frac{P_j}{s_j}$ is the scale-averaged theoretical background spectrum and the degrees of freedom:

$$v = \frac{2 n_a S_{avg}}{S_{mid}} \sqrt{1 + \left(\frac{n_a \delta_j}{\delta j_0} \right)^2}$$

where $S_{mid} = s_0 2^{0.5(j_1+j_2)\delta j}$ and δj_0 is a wavelet specific decorrelation distance (δj_0=0.60 for the Morlet wavelet (Torrence and Compo (1998))).

3.3 Wavelet Cross-Correlation

The wavelet cross-correlation (Torrence and Compo (1998)) is analogous to the well known time series cross-correlation, with the difference that now it is possible to investigate the relationship between two time series not only across time but also for different time scales. There is a discrete version, but this is not as widely used as the continuous one and is not included here. The reader is referred to Whitcher et al. (2000) for further details. The cross-wavelet spectrum of two time series X and Y is defined as:

$$W_n^{XY}(s) = W_n^X(s) W_n^{Y*}(s) \tag{3.11}$$

where $W_n(s)$ is the continuous wavelet transform (Torrence and Compo (1998)) at time n and scale s and * indicates the complex conjugate. However, defined this way the cross-wavelet spectrum is not a reliable tool, for peaks of high correlation can appear even when the two series are independent, reflecting just peaks of high variability of the individual series (Maraun and Kurths (2004)). The alternative is to use a normalized version of the cross-wavelet spectrum, the wavelet coherency. The cross-wavelet spectrum is a complex number, so it can be re-written as $W_n^{XY}(s) = |W_n^{XY}(s)| e^{n\phi_n(s)}$. The wavelet coherency

(Grinsted et al. (2004); Torrence and Webster (1999)) is then defined as:

$$WCO_n^{XY}(s) = \frac{|\langle s^{-1}W_n^{XY}(s)\rangle|^2}{\langle s^{-1}|W_n^X(s)|^2\rangle\langle s^{-1}|W_n^Y(s)|^2\rangle} \tag{3.12}$$

Its value ranges from 0 to 1. The symbol $\langle\cdot\rangle$ indicates smoothing. The smoothing, which can be done in time or/and scale direction, is necessary because otherwise the wavelet coherency would always be equal to 1, for every time point and scale (Torrence and Compo (1998)). The common practice is to smooth both in time and scale.

The wavelet coherency provides information about how strong the association between the two time series is, but it does not carry information about the time lag at which the two series are correlated. The phase function $\phi_n(s)$ provides a measure of the lag difference between the two time series at time n, scale s and it can be calculated as:

$$\phi_n(s) = \tan^{-1}\left(\frac{Im\{\langle s^{-1}W_n^{XY}(s)\rangle\}}{Re\{\langle s^{-1}W_n^{XY}(s)\rangle\}}\right) \tag{3.13}$$

where $s^{-1}W_n^{XY}(s)$ has been smoothed in time and scale.

A significance test for the wavelet coherency was proposed by Maraun and Kurths (2004) and Grinsted et al. (2004), the null hypothesis being that the processes are not significantly correlated. However, the fact that neighboring times and scales are correlated is problematic for deriving the distribution of the test-statistic under the null hypothesis. To overcome this problem, Maraun and Kurths (2004) and Grinsted et al. (2004) make use of Monte Carlo simulation to generate 10000 realizations of two independent Gaussian white noise processes or AR(1) processes. The simulated sample can be used to derive the empirical distribution of the test-statistic under H_0, from which a critical value for the chosen significance level can be obtained. Critical values depend on the amount of smoothing and appear to be scale dependent (Grinsted et al. (2004)) even though Maraun and Kurths (2004) suggests that the ideal is to find the right amount of smoothing so that they are scale independent.

3.4 Some wavelet applications

Percival and Mofjeld (1997) applied MODWT to subtidal coastal water levels in California to characterize the non-stationary behaviour of the data. Rossi et al. (2009) investigated the Mississippi river using wavelet analysis, finding annual variability and multi-year cycles (in both river flow and rainfall) that are not constant through time, with a clear shift around 1970. Kisi (2010) makes use of wavelets in a regression context to forecast river flow in the short term (1,2 and 3 days ahead). The idea is to use some of the detailed components and/or the smooth component from the discrete wavelet decomposition as explanatory variables in a regression model, where the response variable is the river flow at the time of interest. Kisi (2010) claims that such a model performs better in forecasting than artificial neural network (ANN) and ARMA models, even though extreme values are under/overestimated. However, the filter used and how the sample size limitation is dealt with are not mentioned. Cannas et al. (2006) also uses wavelet transform (discrete and continuous) in a regression context. Their forecasting model (based on artificial neural networks) performance is improved when wavelets are used to preprocess river flow data, characterized by non-stationarity and irregular seasonal patterns.

Wavelet analysis has been widely used to investigate large scale climate signals and their influence in river flow and rainfall variability. These climatic signals include the North Atlantic Oscillation (NAO), the Atlantic Multidecadal Oscillation (AMO) and El Niño/La Niña Southern Oscillation (ENSO), whose oceanic and atmospheric components are represented by the sea surface temperature (SST) (a measure of ENSO amplitude) and Southern oscillation index (SOI) respectively. These climatic signals are not independent of each other; wavelet coherence results between the ENSO and NAO during 1956-2000 show significant peaks around 1920 (scale 4-8 years) and 1940 at scales 4-14years (Maraun and Kurths (2004)). Torrence and Webster (1999) report a close relationship between SST and SOI. Wavelet results for the NAO series during 1900-1999 do not show a clear dominant frequency, but enhanced non constant (along time) variability at periods 0.5,1,3,8-14 years, with a trend towards high periods (Markovic and Koch (2005)).

Smith et al. (1998) carried out a wavelet based classification of a large set of US rivers. Their classification successfully groups the rivers according to the different climatic regimes. Markovic and Koch (2005) used continuous wavelet analysis to analyze extreme monthly precipitation in Germany and relate patterns of rainfall variability to the NAO. The results from 27 locations reveal a clear annual cycle common to all locations, plus additional low frequency components in some of them. There seems to be a spatial pattern, with the influence of the NAO being different for different geographical locations. However, their results are based on the cross wavelet spectrum rather than on the wavelet coherency, hence it might be that they are not actually identifying autocorrelation but significant variability of the individual series.

Torrence and Compo (1998) and Torrence and Webster (1999) analysed monthly sea surface temperature over the period 1871-1997. Their results from a continuous wavelet analysis (using the Morlet wavelet) show most of the SST variability concentrated on scales 2 to 8 years ("ENSO band") but also some power at higher scales. They identified temporal changes in the so called 'ENSO band' for SST and SOI, with high (significant) variability during 1875-1920 and 1960-1990 (which would translate into many warm and cold events of large amplitude) and a shift in the latter towards higher scales, and low during 1920-1960, suggesting a 12-20 year amplitude modulation of ENSO events (Torrence and Webster (1999)).

Torrence and Webster (1999) extended the analysis in Torrence and Compo (1998) to investigate the influence of SST in Indian rainfall, as well as its relationship with the SOI. In their analysis, periods of high ENSO variability correspond to periods of high monsoon variability. El Niño seems to be associated with deficient monsoon rainfall while La Niña is associated with abundant monsoon rainfall, even though, as the authors point out, but there might be other processes affecting this relationship. Significant coherency (over 0.8) is found between SST and rainfall in the 1 year and 2-8 years band; for the latter, the strength is not constant over time, being lower from 1920 to 1960 and particularly strong during 1871-85,1895-1925 and 1960-1990 (Torrence and Webster (1999)). Towards the end of the record (1960-1990) there is a shift in the coherency from about 4-8 year band to the 2-4 year band. The results for phase difference show the two series being out of phase, not in a constant way though. SST appears to be leading rainfall,

with a delay of about 4months on the annual cycle, and about 6months on the ENSO band. This seems to be explained by a "lag between the indian monsoon season and the boreal winter peak of ENSO events" (Torrence and Webster (1999)).

Labat et al. (2005) studied the influence of two climatic indices, the SOI and the NAO, during the period 1900-2000, on 4 large rivers (Amazon, Parana, Orinoco and Congo) using monthly data. Annual cycles with clear temporal variation were identified in all 4 rivers by means of the global wavelet spectrum, as well as peaks at higher scales (ie, lower frequencies) that vary from river to river, with the Amazon river having the most complex structure amongst the four. Results from wavelet coherence indicate that the relationship between climatic indices and river flow is not constant over time, being stronger during 1940-1970, and varies from river to river. For the NAO and Congo, Orinoco and Parana rivers, the correlation is concentrated on a scale of about 8 years, while for the SOI and Amazon, Congo and Parana rivers, the correlation is concentrated on a scale of about 20 years.

Sen (2009) identified a strong (significant at 5%level) intermittent annual cycle in a set of 15 rivers in England and Wales during the period 1865-2002 by means of the wavelet power spectrum. Rivers on the South East also show (intermittent) power at the 2-8 year band and 11-12 year band. Sen (2009) argues reasons for this intermittency are due to ground water and runoff natural intermittency. Sen (2009) also makes an attempt to connect river flow and NAO, but based on the fairly weak argument that if both have intermittent patterns of variability, they should be related.

Wavelet analysis has successfully been used in the literature to investigate changes in river flow and rainfall variability and their relationship with large scale climate indices. Independently of the index studied, the relationship appears to be complex and variable over time and scale. The remainder of the chapter illustrates an application of the wavelet transform, both discrete and continuous, on the eight rivers presented in Chapter 2. First the River Tweed (gauging station 21009) is discussed in detail. Then a comparison amongst the eight rivers is presented in terms of their wavelet analysis,

and the influence of the NAO and the AMO is investigated using the wavelet coherency. The results presented are based on the logged river flow series of monthly maxima.

3.5 Case Study: the River Tweed at Norham (gauging station 21009)

The preliminary river flow data analysis reported in Chapter 2 suggests that the time series analyzed are non-stationary. In particular issues with non-constant variability and changes in the seasonal pattern were detected. Hence, wavelet analysis (both discrete and continuous) was carried out on the set of eight Scottish rivers described in Table 2.3. Further, the wavelet power spectrum (Equation 3.2) allows detailed investigation of the time series variability. All the analysis was done with the statistical software R and the packages `sowas` and `wmtsa`. Due to space limitation, only the results for the River Tweed are presented here in detail. Results for the remaining 7 rivers presented in Chapter 2 can be found in Appendix B. The series of monthly maxima was preferred to the daily series because we are mainly interested in high flow values, as well as due to smaller computational cost.

3.5.1 MODWT

The MODWT was preferred to the DWT so that there is no restriction about the length of the series. Maximal overlap discrete wavelet analysis was applied to the monthly maxima series. Note that since the analysis is based on the MODWT rather than on the DWT, the choice of filter is not vital, but given it is of interest to align events in time, an LA(8) filter was chosen. The level of decomposition J_0 was set to four (the maximum level of decomposition J_0 possible for this series is 6). Circular extension should not be a problem here as the data set covers approximately 46 complete years (plus one further month) and we know from the periodogram that there is an annual component.

Figure 3.1 shows $\hat{v}_X^2(\tau_j)$ vs τ_j along with corresponding confidence intervals. The degrees of freedom were estimated as $\hat{\eta}_2 = max\{M_j/2^j, 1\}$. It can be seen that the main contributor to the sample variance of the time series is the vector related to changes

at a scale $\tau_3 = 2^{3-1}$=4 months. This seems reasonable, since changes over a scale of 4 months correspond closely to a seasonal component.

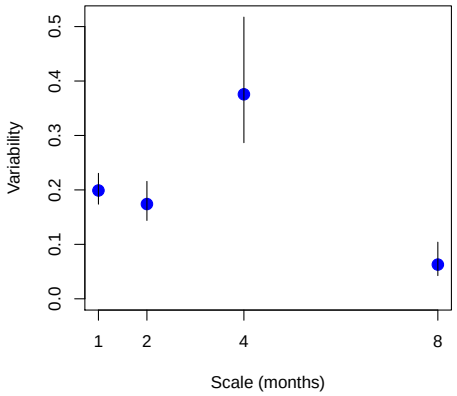

FIGURE 3.1: Wavelet based variance for scales 1, 2, 4 and 8 months. Confidence intervals are based on a χ^2 distribution with $\hat{\eta}_2 = max\{M_j/2^j, 1\}$ degrees of freedom. The time series variance is estimated to be $\hat{\sigma}^2$=0.87. River Tweed (gauging station 21009)

The multiresolution analysis of the monthly maxima series can be seen on Figure 3.2. Each of the detail series D_j reflect changes in the monthly maxima time series on a scale of τ_j months (here,1, 2, 4 and 8 respectively). S_4 can be seen as the trend (it is a smoothed version of data) and it relates to variations of 16 months and higher. D_1 is highly variable and likely to be influenced by weather events (lasting a few days up to 2 weeks). D_2 is also very variable and can be related to persistent weather events, e.g. blocking of high pressure or sequences of fronts in disturbed westerly airflows. D_3 can be seen as an approximation of the seasonal component, while D_4 is less variable than D_3 and can be considered a multi-seasonal component. The near-seasonal component D_3 again appears as the main contributor to the sample variance. We would expect values in D_3 to be large when an average over roughly 4 months differs from values surrounding it. Is is very clear from the plot that the seasonal component is not constant over the years. As Figure 3.3 shows, the peak usually occurs in winter although that is not always the case.

S_4 suggests a complex trend, with a decrease until about 1973, when the trend reaches its minimum, to then increase in a non-monotonic way. Given that component D_3 is the most significant, the time dependent D_3 wavelet variance (Equation (3.5)) was produced (Figure 3.4).

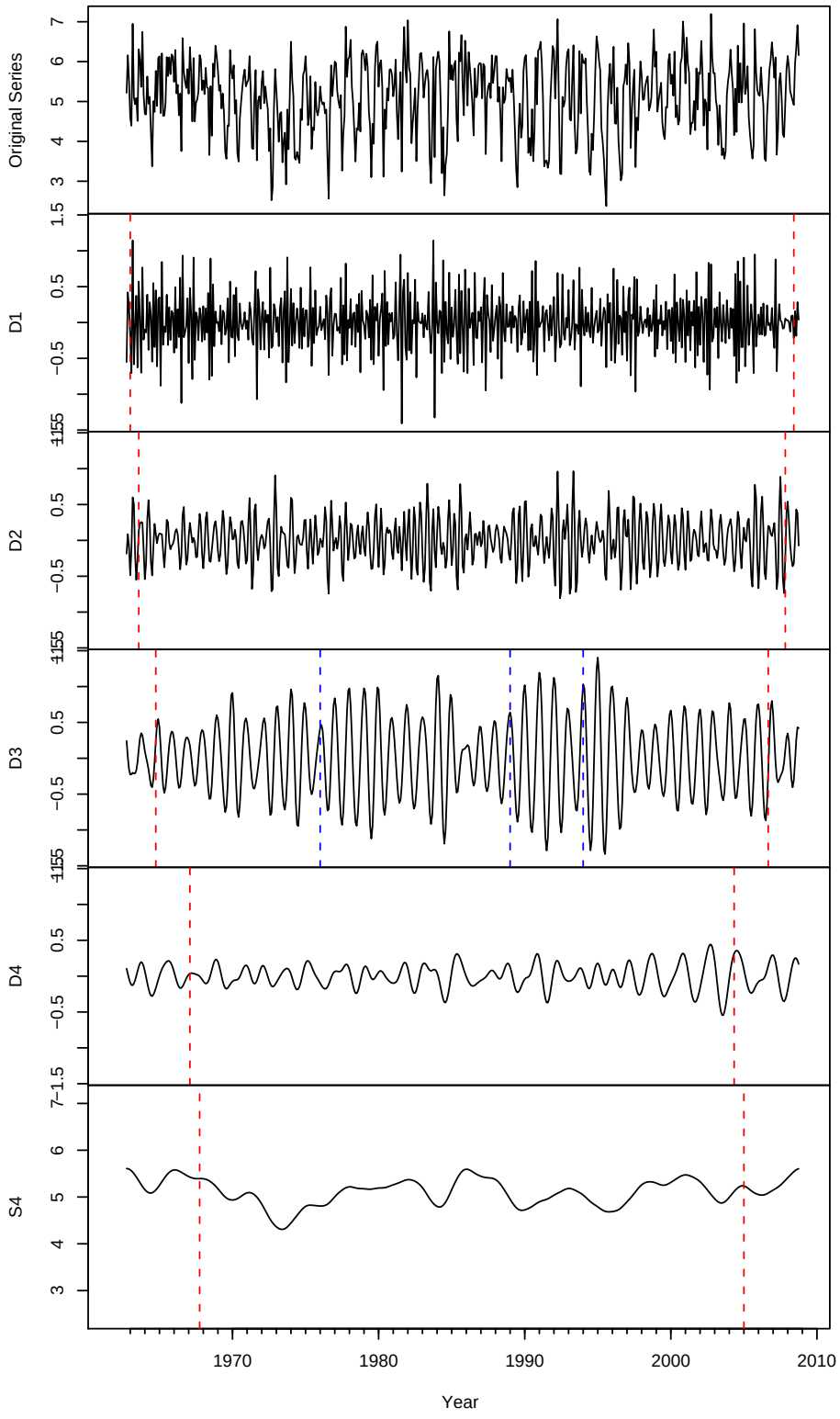

FIGURE 3.2: Multiresolution analysis of monthly series - River Tweed. All four detail components D_1-D_4 are on the same scale, different from the original series (top) and S_4 (bottom). Red dashed lines indicate the areas that might be affected by boundary coefficients. The blue dashed lines on component D_3 correspond to the time points identified in Figure 3.4 at which variability increased. Tick marks on the x axis correspond to 1^{st} of January

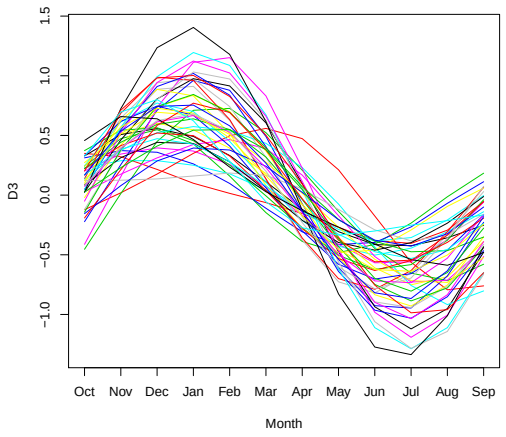

FIGURE 3.3: Seasonal cycle based on the wavelet decomposition - River Tweed. The first and last cycles have been omitted to avoid boundary effects

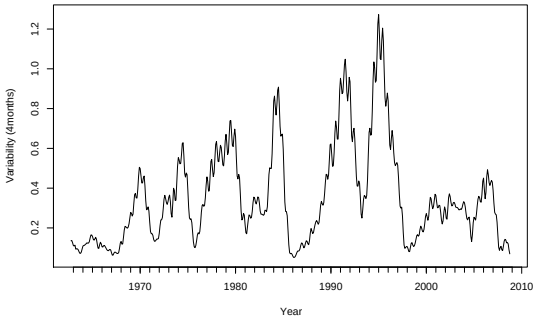

FIGURE 3.4: Time dependent wavelet variability for the near-seasonal component D_3 - River Tweed

Figure 3.4 shows that the variability is not constant over time. For most of the years, the variability seems to increase around April and then again around October. This is easily visible in Figure 3.3. Two periods can be identified as having fairly stable variability, from the beginning of the record until 1979 and 1985-88. This last period also shows a decrease in variability. We can identify three times when there was a clear increase in variability, 1976, 1989 and 1994. If we go back to Figure 3.2 we can see that the conclusions drawn about variability from the time dependent wavelet variance plot are in agreement with the behaviour of the monthly flow series D_3.

3.5.2 CWT

A continuous wavelet analysis was also carried out. While the DWT provides a very detailed description of the data which is useful for identifying the trend and seasonality of a non-stationarity time series, the CWT provides a more global picture so that understanding can be gained on how the series behaves globally, as well as identifying whether there is significant variability at higher scales. The wavelet power spectrum (Equation (3.2)), plotted on Figure 3.5, provides a measure of the variability of the time series at each scale s (y axis) and time t (x axis). The thick contour lines represent 90% confidence regions (assuming the background spectrum to be an AR(1)). Variability ranges from zero (blue) to one (red) as it has been standardized (dividing by σ^2). The observed variability is concentrated on the 1year band. However, it is not constant across the whole time span, and periods of high (significant) variability (1972- mid 1985, 1988-1997 and 1999-2008) alternate with periods of non-significant variability. This is a clear indication of non-stationarity.

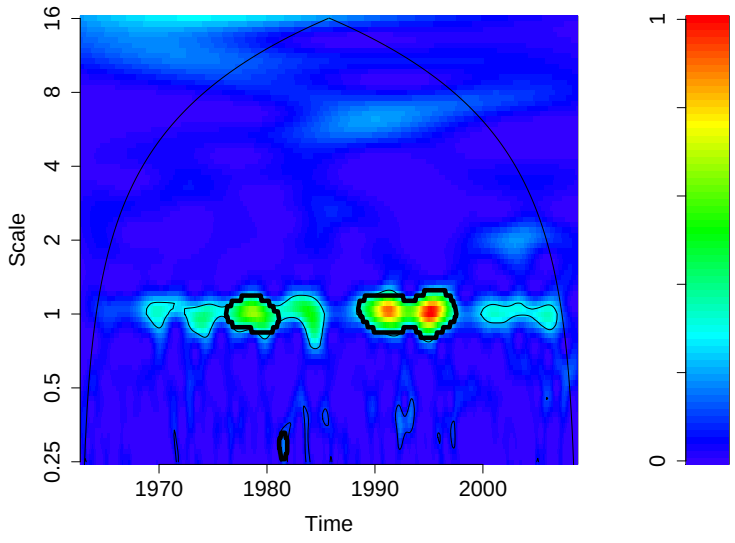

FIGURE 3.5: Wavelet power spectrum of monthly maxima series - River Tweed (gauging station 21009). Both time and scale axis are in years. The thick black line highlight areas of significant variability at the 90% level

The global wavelet spectrum (Figure 3.6) was then calculated (Equation (3.9)) by averaging along time. This plot is equivalent to the periodogram and it shows that the only significant periodic component is the annual cycle, although it illustrates how the non-stationarity could be missed from a purely spectral analysis.

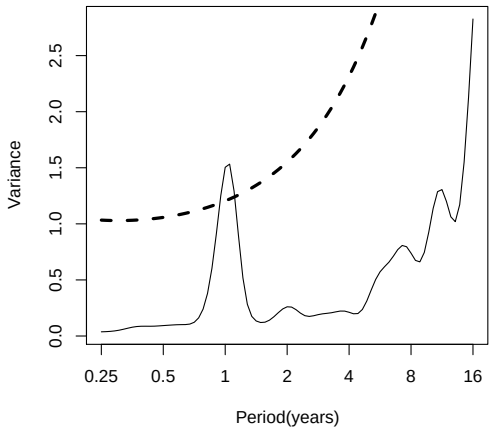

FIGURE 3.6: Global wavelet spectrum of monthly maxima series - River Tweed (gauging station 21009). The dashed line shows the 95% significance level assuming an AR(1) background spectrum

Figures 3.5 and 3.6 show that the annual component is the main contributor to overall variability, and annual variability changes over time. To further investigate these changes, the scaled averaged wavelet power spectrum for the 1 year band (scales 0.86-1.16 years) was calculated (Equation (3.10)) and is shown on Figure 3.7:

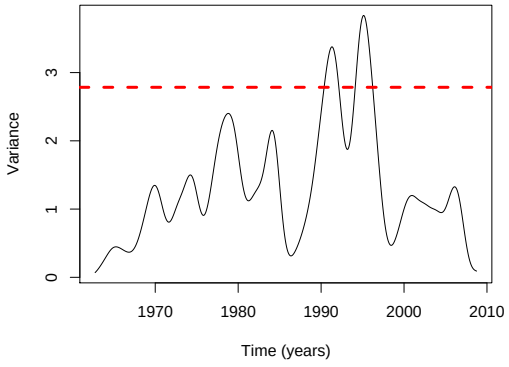

FIGURE 3.7: Scale averaged wavelet power spectrum (annual cycle) - River Tweed (gauging station 21009). The dashed line shows the 95% significance level assuming an AR(1) background spectrum

Looking at the overall picture, there is a clear change point around mid 1986, when the variability is minimum (followed by another period of very low variability around the beginning of 1998). This means that monthly river flow maxima was quite stable in that year. Grew and Werritty (1995) defined flood poor (1964-1973) and rich (late 1980s -

early 1990s) periods, which coincide with low and high variability respectively on the graph. The period of time from 1977 to 1986 has been characterized as the wettest period on record for the UK as a whole (Marsh (1995)), which corresponds with the 'cluster' of high variability and flood events just before the low variability change point, followed by a second 'cluster' of peaks which corresponds to the flood-rich period. The decrease in variability since summer 1995 until it reaches a minimum point in 1998 could be explained by a remarkable drought in April-August 1995 followed by a dry winter 1995/1996 which resulted in the second driest summer for Scotland on record (Marsh (1995)).

3.6 Wavelet based river comparison

The eight rivers introduced in Chapter 2 (Figure 2.5) are analyzed here using wavelet analysis. A similar analysis to Section 3.5 was completed for the remaining seven rivers with the aim of comparing the timing and scales of variability. Results presented in this section are also based on the (logged) series of monthly maxima. MODWT (based on an LA(8) filter (Percival and Walden (2006)) with 4 levels of decomposition was applied to the monthly maxima series (normalized to the overall mean). Each of the river series was decomposed as $\sum_{j=1}^{4} D_j + S_4$.

The near-seasonal component D_3 is the main contributor to the sample variance for all eight rivers. Figure 3.8 confirms that the seasonal component varies over the years and suggests some consistency between the different sites. The variability of the seasonal cycle is explored further in Figure 3.10.

The trend series (component S_4 from wavelet decomposition), plotted on Figure 3.9, suggest some differences between rivers in the East (top four plots in Figure 3.9) and the West (lower four plots in Figure 3.9) of Scotland. There is a sharp decrease around April 1973 ((a) on Figure 3.9) that, even though it is present in the Clyde trend too, is not as marked as for the rivers in the East. This decrease is very small in the River Tay, despite being in the East, and it appears a bit earlier in time in rivers Lossie and Ewe, both relatively small catchments. The trend of the River Clyde reaches its minimum

in April 1984 ((b) on Figure 3.9). This decrease is also present in the trends of the Eastern rivers, but it is not as marked. River Tweed and Water of Leith's trends peak in February 1986 ((c) on Figure 3.9). This peak also appears in the River Lossie, but a bit earlier, and in the River Clyde. There is another decrease common to all rivers in April 1996 ((d) on Figure 3.9) although in this case it is more prominent in the West than in the East, where, apart from the Tay, the trough appears a bit earlier in time. River Clyde's trend peaks in November 1998 ((e) on Figure 3.9), a feature common to all rivers although this is most pronounced for the Clyde. It is difficult to say whether there is an overall increasing or decreasing trend as the trend series are not monotonic; however, it looks as if there has been a slight increase in the East from 1973 but not in the West.

The scaled averaged wavelet power spectrum for the 1 year band (scales 0.86-1.16 years) was calculated (Equation 3.10) to get a measure of fluctuations in the yearly cycle variability. This is equivalent to calculating the time dependent wavelet variability (Equation 3.5). The resulting series are plotted in Figure 3.10. For all rivers there is a clear indication of non-stationarity with periods of high variability alternating with periods of very low variability. In particular, there is a clear change point in 1986 for both eastern and western rivers (Figure 3.10(c)), when the variability is minimum. Note that this decrease in the variability series corresponds to an increase in the trend series (Figure 3.9(c)). River Tweed appears to show the greatest variability amongst the eight rivers and River Lossie shows the smallest.

3.7 Relationship with climatic indices

In Europe, the main climatic influence comes from the North Atlantic Ocean, for which two main associated signals have been identified, the North Atlantic Oscillation (NAO) and the Atlantic Multidecadal Oscillation (AMO). The reader is referred to Chapter 2 for further details about the NAO and the AMO.

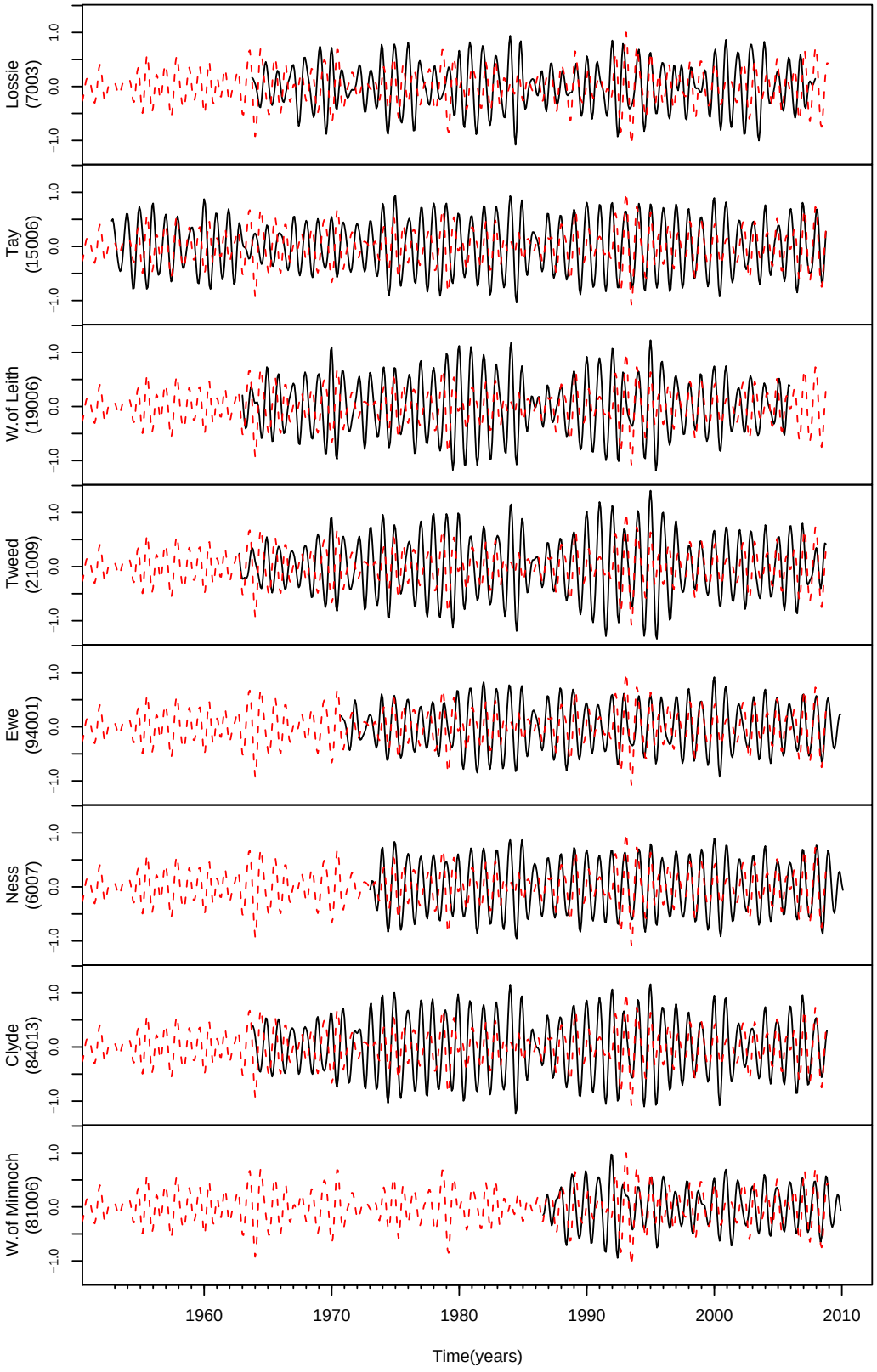

FIGURE 3.8: Seasonal component (D_3) for all rivers. The dashed red line represents the seasonal component for the NAO.

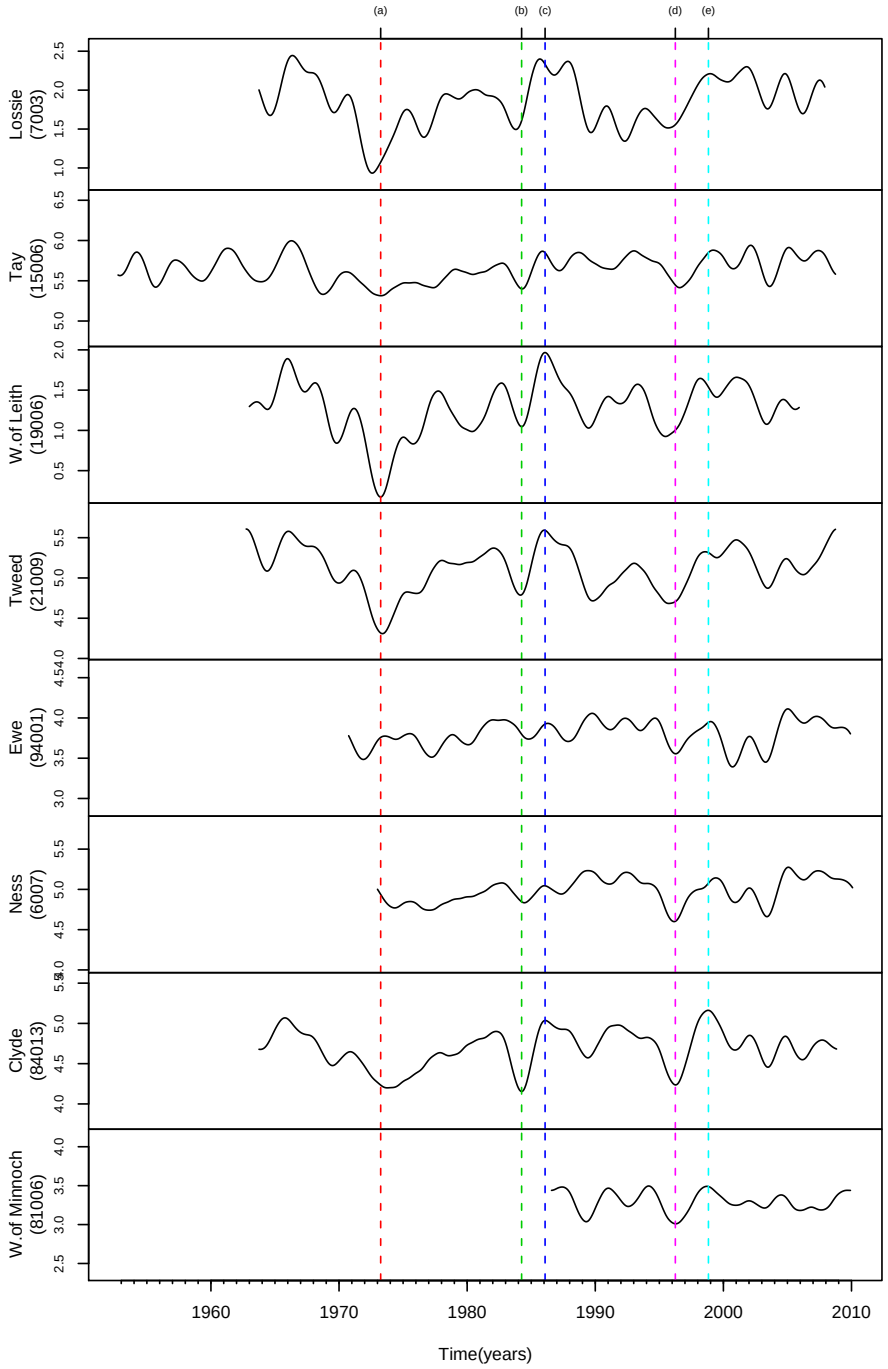

FIGURE 3.9: Trend (S_4) from wavelet decomposition for all rivers. Units are $\log(m^3/s)$. Scale on the y axis changes across rivers. The vertical lines (a), (b), (c), (d) and (e) highlight particular features of the data and are referred to in the text

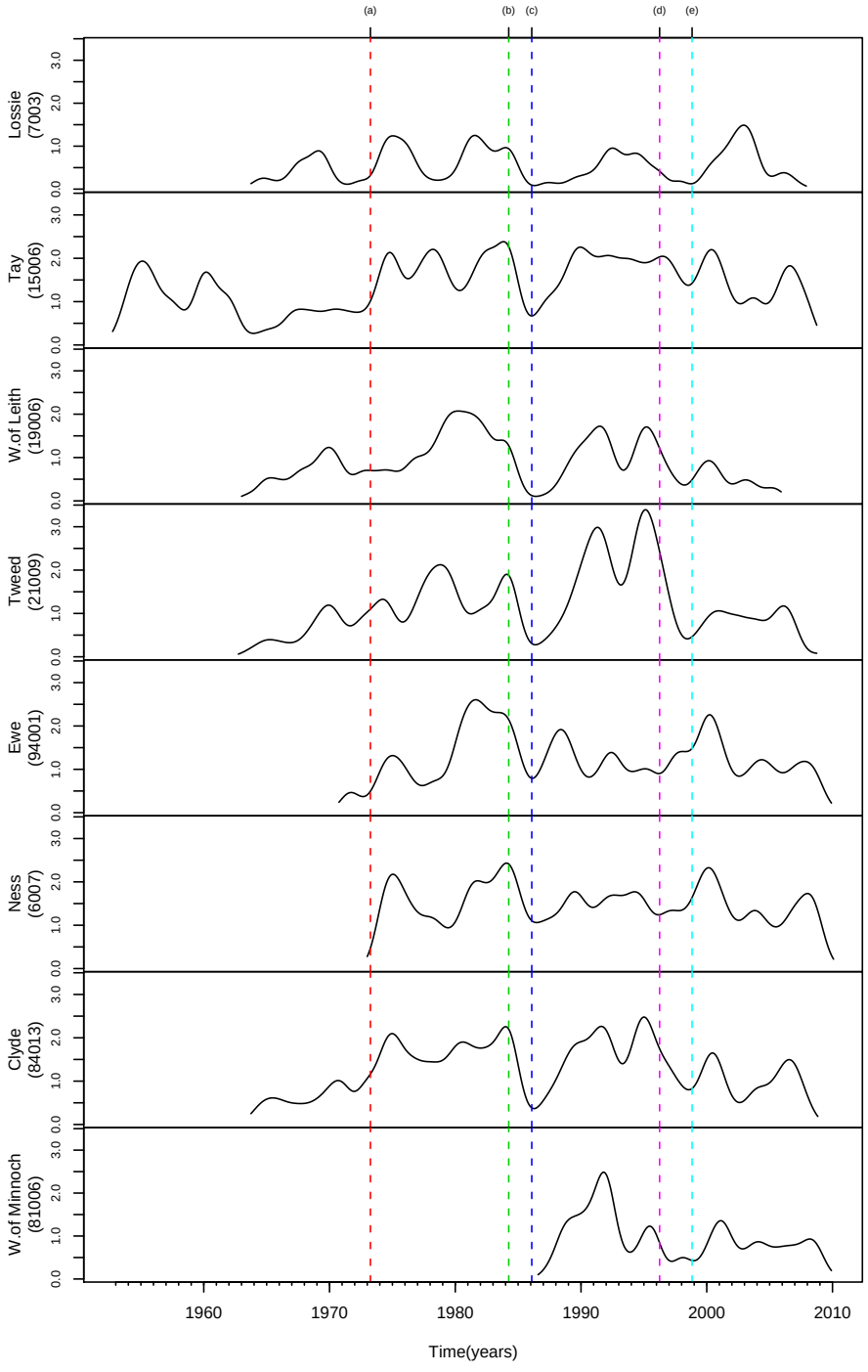

FIGURE 3.10: Seasonal time dependent variability based on component D_3 for all rivers. The reference lines (a),(b),(c),(d) and (e) are the same as in Figure 3.9 for ease of comparison.

Even though some authors restrict the NAO influence to winter months, NAO has been shown to be significantly correlated with precipitation over Northern Europe outside the winter months (Markovic and Koch (2005)). The results presented here are based on the whole NAO series. The length of the data record varies from river to river; for comparison purposes, only values from October 1973 onwards have been used. The Water of Minnoch (gauging station 81006) was analyzed on its own as the time series was considerably shorter than for the rest of the rivers (starting in August 1986), to avoid losing a large amount of data from the other rivers. The wavelet coherence (top) and phase difference (bottom) (smoothed in time and scale direction) between NAO and each of the rivers is shown in Figures 3.11 (Eastern rivers) and 3.12 (Western rivers). In general, the correlation takes values between 0.4 and 0.6 for most of the time and scales. The thick black contour lines denote regions of statistically significant correlation. Critical values were calculated as the 95^{th} percentile of the empirical distribution of the simulated wavelet coherencies following Grinsted et al. (2004) and Maraun and Kurths (2004).

Amongst the Western rivers (Figure 3.12) the coherency for the Ewe and Ness has a similar structure and appears to be the strongest one out of all the rivers considered here, with statistically significant correlation around 1987-1990 and then again during 1999-2004 on the 1year band and around 1987-1994 for the Ewe and 1980-1990 for the Ness on the 2 to 4 years band. The River Clyde also shows a period of statistically significant correlation around 1987-1990 on the 1year band. It is difficult to make a definite statement about the Water of Minnoch given the limited number of observations, but it seems plausible that the peak of significant correlation around 1987-1990 that appears in the rest of the rivers appears here too. In terms of the phase difference, river flow and NAO seem to be out of phase, but this relationship is not constant over time. For the 1 year band, the phase difference changes from about the mid 1980s onwards, while for the 2 year band, it looks as if the phase difference can be divided into 3 different stages.

Amongst the rivers in the East (Figure 3.11), the Tay, Water of Leith and Tweed also show significant correlation on the 1 year band around 1987-1990 and 1998-2001, plus a period of significant correlation on the 4 year band that can also be seen for the River Tay. The River Lossie seems to have a different pattern, with small patches of high

correlation along the 1 year band around 1991-1994 and 2002, ie, with a delay of about 1 year with respect to the rest of the rivers. All four Eastern rivers show intermittent periods of high correlation at higher scales. As was the case for the rivers in the West, NAO and river flow maxima are out of phase but the phase difference is not constant for every scale and time point.

To gain a better understanding of the phase difference, the 1 year band filtered NAO series was added to Figure 3.8. It can clearly be seen, for example looking at the plot for the River Ewe, how the NAO series is being slowly shifted from being completely out of phase around 1976-1977 to being completely in phase at the beginning of the 1990s. A similar plot was produced for the AMO series, but the variability for the 1 year band filtered series was too small compared to the variability of the river maxima series and as a result the plot was not informative and is not shown here.

The correlation structure between river flow and AMO (Figures 3.13 and 3.14) resembles that of the NAO, although the relationship seems to be stronger than for the NAO. As before, there is moderate correlation for most of the time and scales, but it is not constant. The highest correlation is concentrated on the 1 year band, suggesting that AMO has a strong influence on the seasonal cycle of river maxima. A peak of high, significant correlation can be found for all rivers but for the Lossie around 1987-1990, independently of whether they are in the East or West of the country. It is interesting to see that the AMO influence appears to be stronger in the second half of the record (from 1987 onwards) than it is in the first. Two rivers show slightly different patterns; these are the River Lossie, for which no significant correlation is found, and the Tweed, which shows high (significant) correlation during the period 1977-1982 on the 2 year band.

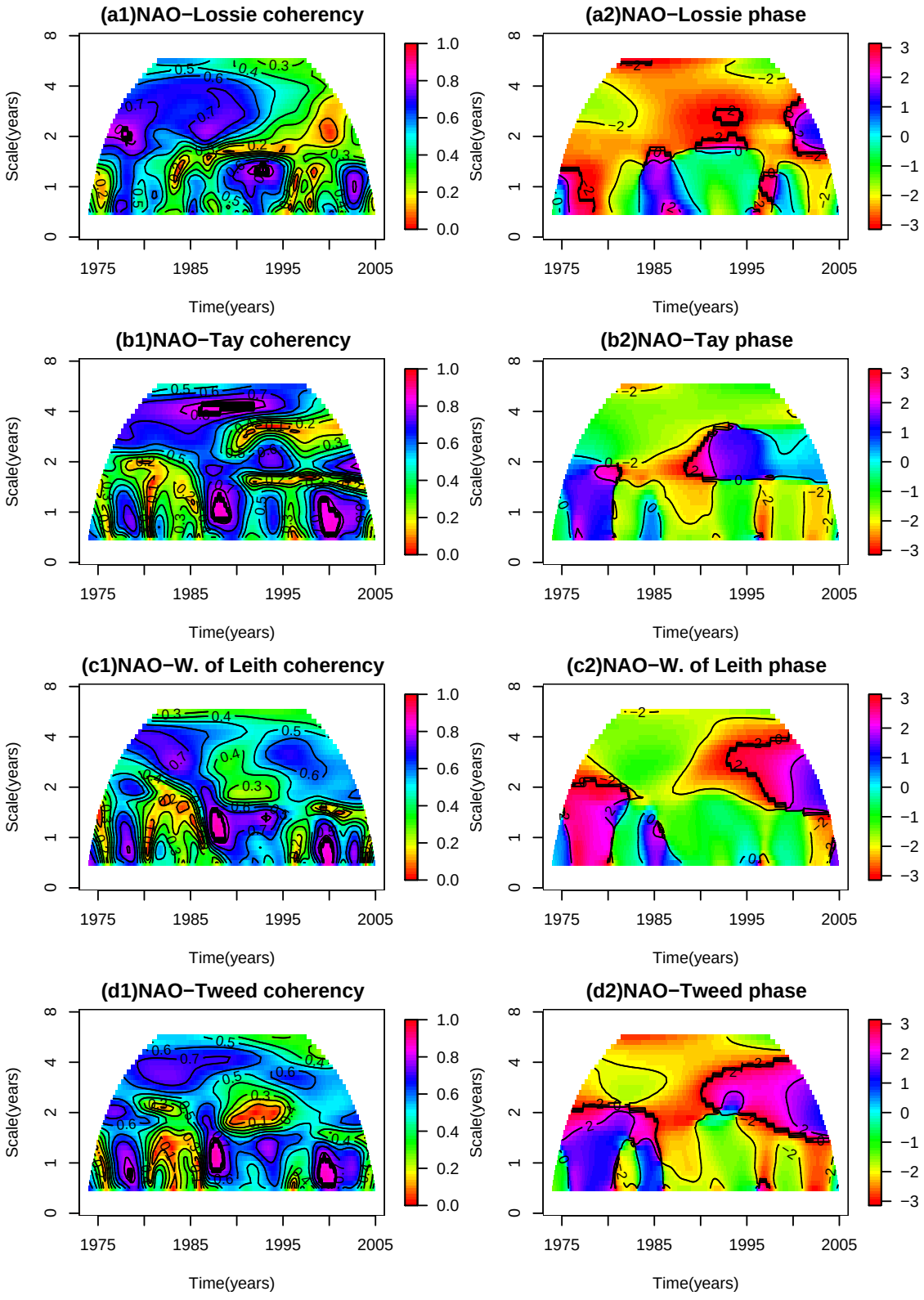

FIGURE 3.11: Wavelet coherency (left) and phase (right) between NAO and rivers (a)Lossie, (b)Tay, (c)Water of Leith and (d)Tweed. The thick black contour lines on the wavelet coherency plots denote regions of statistically significant correlation based on MC simulation. The black lines on the phase plots represent contour lines.

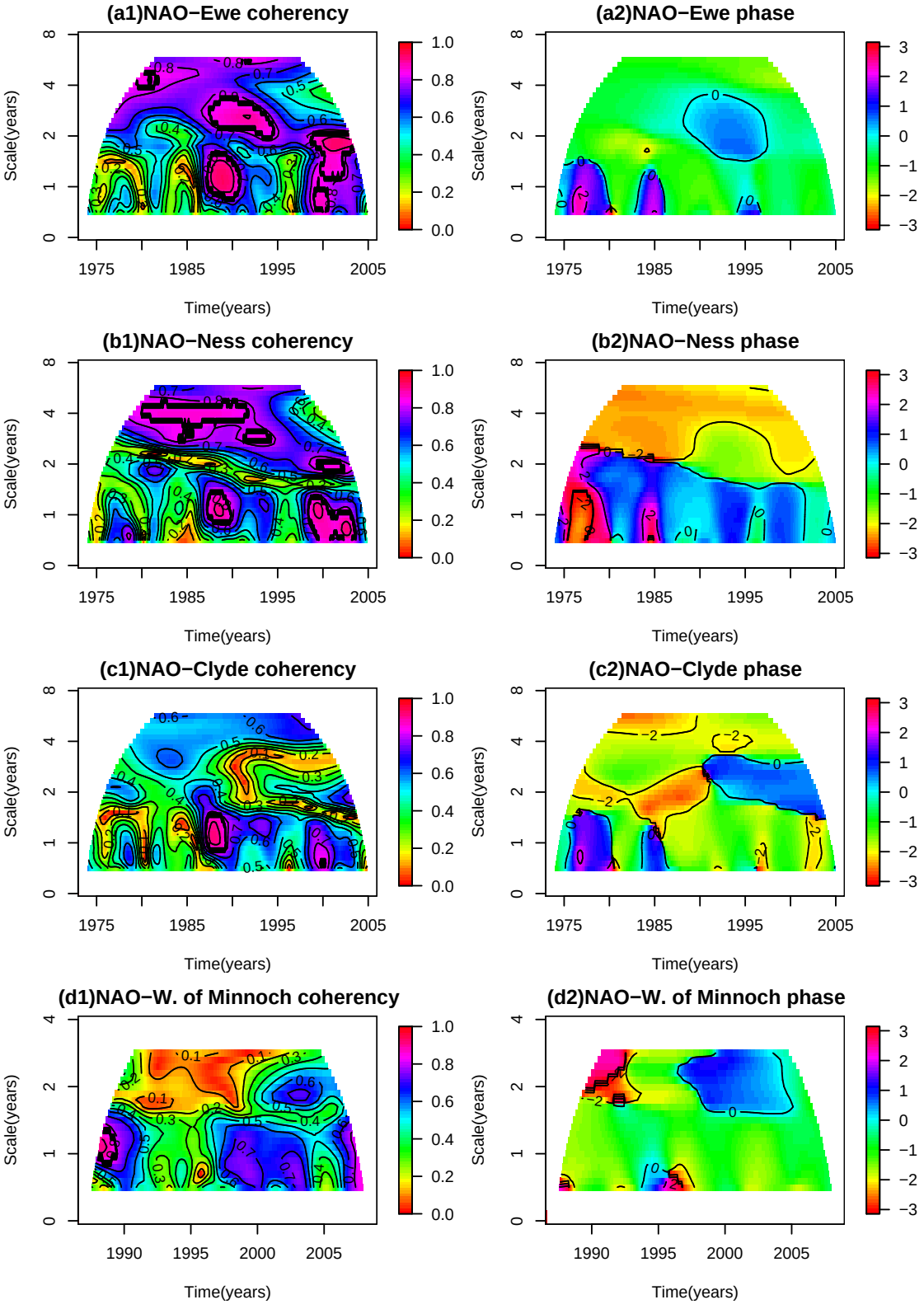

FIGURE 3.12: Wavelet coherency (left) and phase (right) between NAO and rivers (a)Ewe, (b)Ness, (c)Clyde and (d)Water of Minnoch. The thick black contour lines on the wavelet coherency plots denote regions of statistically significant correlation based on MC simulation. The black lines on the phase plots represent contour lines. Note the timescale on the x axis is different for the Water of Minnoch (Figures (d1) and (d2))

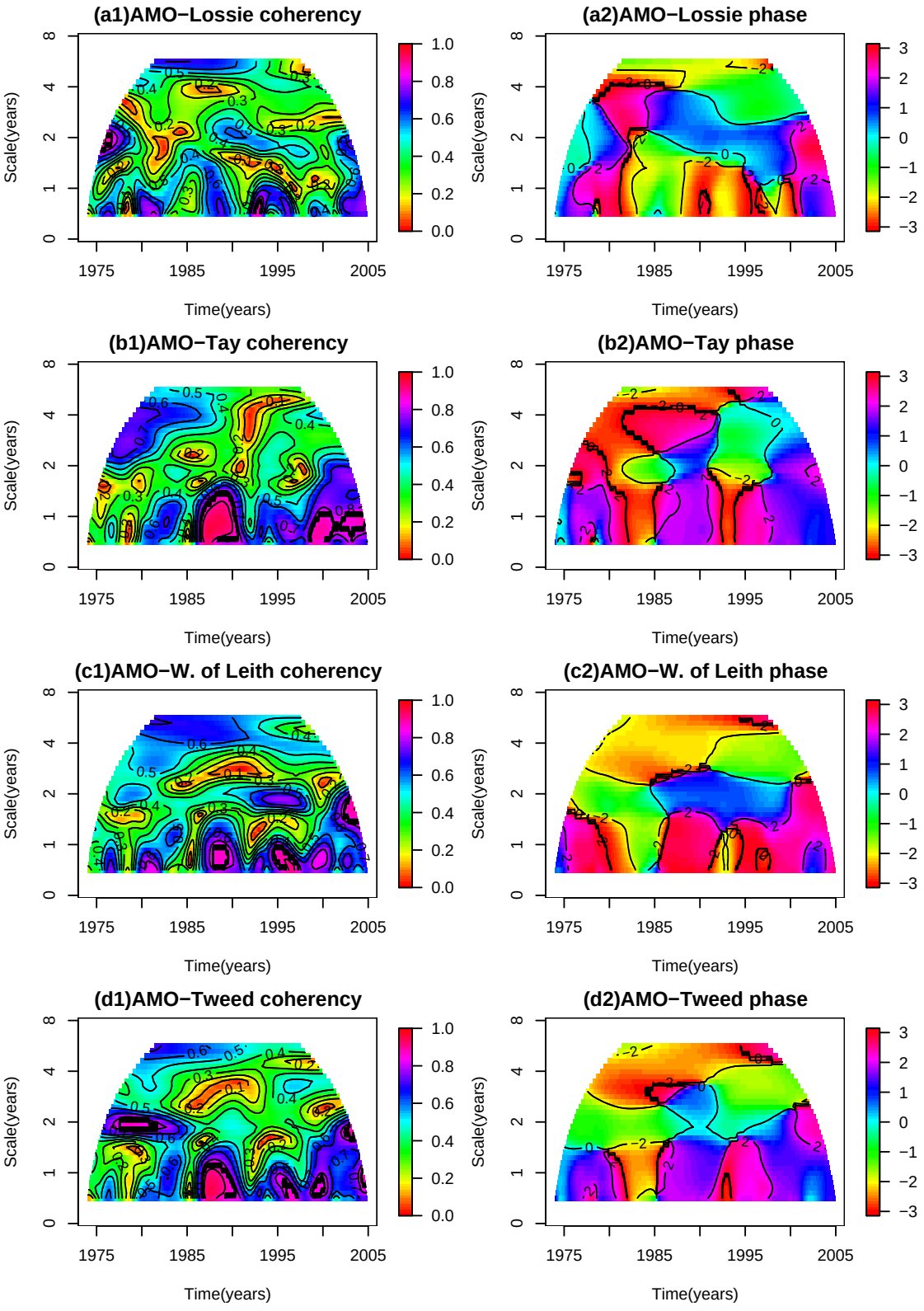

FIGURE 3.13: Wavelet coherency (top) and phase (bottom) between NAO and rivers (a)Lossie, (b)Tay, (c)Water of Leith and (d)Tweed. The thick black contour lines on the wavelet coherency plots denote regions of statistically significant correlation based on MC simulation. The black lines on the phase plots represent contour lines.

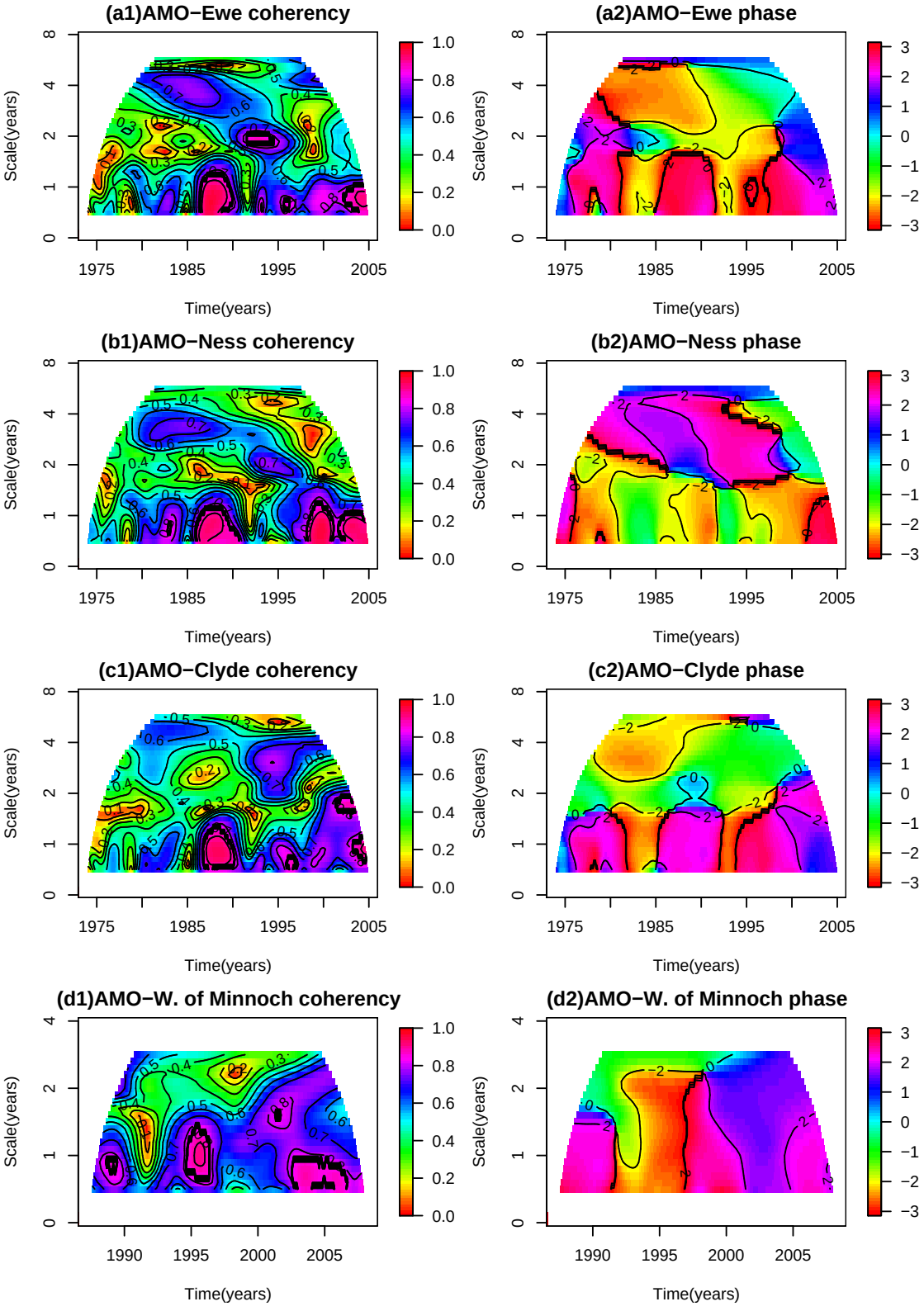

FIGURE 3.14: Wavelet coherency (top) and phase (bottom) between AMO and rivers (a)Ewe, (b)Ness, (c)Clyde and (d)Water of Minnoch. The thick black contour lines on the wavelet coherency plots denote regions of statistically significant correlation based on MC simulation. The black lines on the phase plots represent contour lines. Note the timescale on the x axis is different for the Water of Minnoch (Figures (d1) and (d2))

3.8 Summary and discussion

Wavelet analysis is presented here as a useful method for detecting and comparing trends, investigating spatial heterogeneity and periods of significant variability in non-stationary environmental time series. By subsequently filtering the original series, we obtain sequences of results which relate to variations at different scales (frequencies). The result is a time-scale decomposition of the original series that provides a more informative way of looking at the time series. Results from a set of Scottish rivers revealed significant changes in the variability of the seasonal pattern over the last 40 years, with periods of high and low variability associated with rich and poor flood periods respectively. The results suggest a difference in the long term trend of river flow maxima between the East and the West that had been pointed out by previous studies (Black and Burns (2002); Black (1996); Werritty (2002)). The differences in trends, however, are very subtle in some cases.

The trends in monthly maxima identified by means of wavelet analysis (Figure 3.9) are qualitatively very similar to those previously identified in Chapter 2 for the daily series using the stl decomposition (Figure 2.15). However, the wavelet analysis provides a complex seasonal pattern that the stl decomposition was not able to capture, as well as how the variability in the seasonal cycle evolves over time.

3.8.1 Hydrological findings

Differences in long term trends and seasonal variability between the East and the West were found, with variability being higher for rivers in the East of Scotland. This is in agreement with rainfall variability being greater in the East than in the West, as discussed in Chapter 2.

A clear indication of non-stationarity was found for both eastern and western rivers. 1986 was detected as a 'change point', when the seasonal variability is minimum. This is in agreement with Black (1996), who suggest a shift towards a flood rich period in the late 1980s. The cluster of high variability just before that (from about 1977 to 1986), especially in the West, would correspond to the wettest period on record for the UK

(Marsh (1995)).

The decrease around 1995-1996 in both the long term trend and annual variability and common for most rivers corresponds to a period of drought that commenced in the summer of 1995, recognized as the second driest Scottish summer, and that affected the winter of the following year, particularly dry in the UK (Marsh (1995)). Interestingly, the Strathclyde floods (Black and Bennett (1994)) that, amongst others, affected the River Clyde in December 1994, do not correspond to the highest peak in the trend of monthly maxima for this river, highlighting the importance of considering other factors rather than just high river flow when assessing flood risk. The 1994 Clyde event also had a lot of snowmelt as a result of a warm front that led to very fast melting and a lot of rain. As Katz and Brown (1992) point out, the probability of extreme events is highly influenced by changes in variability, more than changes in the mean. In the annual variability series, in fact, December 1994 corresponds to the peak of the series for the Clyde. High variability may reflect less stable weather in the previous and/or following months to the peak. This peak of high variability, common mostly on rivers in the South of Scotland, might reflect the strong contrast between the constant 48 hour rainfall period of December 1994, originated by a slow moving frontal system, and the following unusually dry winter. The rivers Tay and Ness, in particular, seem to be more stable than the rest in terms of variability. This might be due to the fact that they are affected by hydropower regulation (see Chapter 2). The River Tweed shows the greatest variability amongst all eight rivers investigated. The catchment of the Tweed has some reservoirs but their effects are not significant and hence are not expected to control the variability of flows. The Water of Leith, despite being a very small catchment, shows great variability too.

3.8.2 Climate influence

The influence of large scale climatic indices such as the North Atlantic Oscillation (NAO) and the Atlantic Multidecadal Oscillation (AMO) was investigated using the wavelet coherency, an equivalent of the 'traditional' cross-correlation, with the difference that the wavelet coherency allows to investigate correlation across different scales. The results suggested a complex relationship between the climatic indices and monthly maxima river

flow, with the AMO being more influential than the NAO on the annual river flow cycle. Peaks of significant correlation following Maraun and Kurths (2004) and Grinsted et al. (2004) have been identified at localized time periods and scales. While the correlation structures (as in coherency plots) between river maxima and AMO and NAO have common features (meaning that they are mainly concentrated on the 1 year band and that they are not constant over time) the phase difference plots are distinct. Even though river flow and AMO are also out of phase and the phase difference is not constant, the difference along the 1 year band appears to be nearly the opposite of what was observed for the NAO (Kerr (2000)).

The NAO is known to exert an important influence on European climate and its variability (Markovic and Koch (2005)) particulary in the winter months (Markovic and Koch (2005); Shorthouse and Arnell (1997)). Since the NAO has such a strong influence on rainfall, the expectation is that this influence will extend to river flow too. However, the relationship between NAO and river flow is not simple. Previous studies have linked it to winter river discharge globally (Dettinger and Diaz (2000); Labat (2010)) and in Europe (Shorthouse and Arnell (1997); Kingston et al. (2009); Bouwer et al. (2008)). Macklin and Rumsby (2007) investigated the relationship between NAO and floods in upland catchments in the UK over the last 250 years, arguing that the non-stationarity of the flood series was due to large scale climatic fluctuations. The results presented in this chapter for the eight Scottish rivers show a clear influence of the NAO and the AMO on the annual cycle of river flow, so it may be possible that part of the non-stationarity of the series is linked to large scale climatic variations as Macklin and Rumsby (2007) suggest.

Macklin and Rumsby (2007) claim that the relationship between NAO and floods changes both spatially and temporally, with the NAO influence over Scotland being different to that in England and Wales (Macklin and Rumsby (2007)) and within Scotland itself, where 3 upland regions were studied, Glencoe, An Teallach, in the West of Scotland and the Cairngorms, in the East. Their results suggest an increase in the frequency of floods in the second part of the 19th century and again in the 1980s, the latter associated with high autumn rainfall and positive NAO (Macklin and Rumsby (2007)), contrary to data from England, which suggest higher frequency floods when NAO is negative (Macklin

and Rumsby (2007)). In agreement with the findings of Macklin and Rumsby (2007), Shorthouse and Arnell (1997) report strong positive correlation between regional runoff and NAO in Scotland, especially in winter (December,January,February), although the strength of the correlation varies across the country. The results from the wavelet coherency analysis presented in this chapter also suggest that the influence of NAO varies slightly from catchment to catchment, with rivers in the North-West showing periods of correlation at scales higher than the 1 year band, the latter being common to most of the rivers in both the East and West. In particular, rivers Ewe and Ness show the strongest correlation pattern. This can be explained by their geographical location, in the North-West of Scotland. Weather patterns in this area (and consequently river flow) tend to be more affected by low pressure systems than the rest of Scotland, and the NAO measures pressure gradients.

In Europe, results from a cross-wavelet analysis suggested strong correlation with river flow during 1900-1950 on a scale band of 8-15 years (Labat (2010)). However, data during that period of time was not available for the Scottish rivers investigated here and a direct comparison is not possible. Shorthouse and Arnell (1997) also investigated the relationship between NAO and river flow during the period 1961-1990 looking at 744 river basins across Europe, finding spatial and temporal patterns in the NAO influence. However, their results, despite being indicative of what the relationship might be, must be interpreted carefully, for the conclusions are based on the results of a Pearson correlation analysis, which assumes independence of observations (something that is rarely the case when dealing with time series) and a linear relationship between the two series.

Even though the AMO has only been defined recently, its oscillatory nature was observed in the early 1970s (Kerr (2000)), when researchers noticed an increase in the North Atlantic sea surface temperature during 1910-1940 that was accompanied with an increase in global air temperature, after which a phase of cooling down began, both in sea surface and global air temperature, to then warm up again in the 1980s. Phases of warm/cold AMO index have been related to anomalous regional climate, particularly in the North West of Europe. This dependence seems to vary seasonally, with the highest impact during the summer months (June, July and August) (Knight et al. (2006); Sutton and

Hodson (2005)).

The relationship between the AMO and river maxima also changes slightly from catchment to catchment, but the influence along the 1 year band appears to be stronger than that of the NAO, especially at the end of the 1980s. During this period of high correlation, the AMO has been said to be in negative or cold phase, although unusually high values (to be in a cold phase) were recorded, regarded as a period of 'transition' between the cold and warm phase. From that point onwards, the correlation with river maxima is higher, suggesting that rivers might be more affected when AMO is in its warm phase. At this point, this results should be interpreted as merely indicative, and further research is needed to draw definite conclusions. The AMO has an oscillation period of about 60 years and here we are looking at possible relationships over 30 years. It would be useful to extend the length of the records to see how this oscillation has influenced river flow in the long term, to investigate whether the effect has always been similar or has changed through time. Further, AMO is claimed to be linked to summer rather than winter climate (Knight et al. (2006); Sutton and Hodson (2005)). A seasonal analysis to gain a better understanding of the relationship would be informative.

3.8.3 Statistical issues

In the discrete case, both the wavelet filter and the the width of the filter need to be decided a priori. An LA(8) filter was preferred. A least asymmetric filter was chosen so that events in the individual components of the wavelet decomposition could be aligned with events in the original time series. This allows identifying features of interest like peaks in trends or how the seasonal pattern varies with time. The width of the filter was set to eight after carrying out a sensitivity analysis using different width values. As Percival and Walden (2006) point out, values smaller than eight introduced artifacts, mainly in the form of sharp peaks, in the wavelet decomposition. On the other hand, values greater than eight just increased the number of boundary coefficients without really affecting the decomposition itself. The Morlet wavelet was chosen for the continuous wavelet analysis as it is the one mostly used and recommended in the literature. Alternative wavelet functions were not explored.

The level of decomposition in the discrete case was chosen to be four. This choice can be justified based on two arguments. First, the smooth component S_4 reflects variations on a scale of 16 months and longer. Variations in river flow on that time scale can be considered as a trend. Second, further decomposition was investigated, but graphical inspection of the resulting components suggested that no further information was obtained from increasing the level of decomposition.

The length of the time series considered was never a power of two. For that reason, the maximal overlap discrete wavelet transform was preferred to the discrete wavelet transform. While this avoids the sample size restriction of the discrete wavelet transform, it has some limitations, as the wavelet coefficients are no longer independent. This is an important assumption for wavelet based trend estimation (Percival and Walden (2006); Craigmile and Percival (2002)). The trend and seasonal component have been chosen here as the S_4 and D_3 components of the multiresolution decomposition of the series. This is a fairly simple approach and more sophisticated techniques are available in the literature. The main focus of this chapter however was to investigate the variability of the time series, as well as to look at the relationship with large scale climatic indices. For the first purpose, the MODWT is appropriate.

Wavelet based cross-correlation appears to be a good alternative to traditional correlation, even though some authors (Torrence and Compo (1998)) argue that smoothing is somehow contrary to the purpose of using wavelets, which is that of improving the localization of events, as localization is decreased by smoothing. Nonetheless, cross-correlation is highly informative and allows exploration of relationships in a broader time frame, with the assumption of independence no longer being made. With respect to the linearity of the relationship, the interpretation is less clear. While some authors define coherency as the strength of the linear relationship between two series, it has been suggested that the phase difference might be interpreted not only as the time lag but also the degree of linearity (Velasco and Mendoza (2008)); a value close to the extremes $-\pi$ and π would suggest a (negative/positive)linear relationship, while values in between would mean that the relationship is not linear. It that was the case, the phase difference results for the eight Scottish rivers would suggest that the relationship is close to linear

for the AMO, while for the NAO it is not.

Seasonality is a feature common in environmental data. However, probably it has not been given the attention that it deserves. While many studies of environmental series attempt to removing any seasonal effects by just subtracting seasonal means, that means that the seasonal pattern itself cannot be identified or explored. This approach has several issues; one is how seasons are defined, not clearly specified in the literature. Also, the timing of changes between seasons can vary even if they remain at three months long, and the lengths of hydrological seasons may be changing. Non-constant variability is a strong feature that has to be taken into account when analyzing time series data, in the context of stationarity. The wavelet results presented in this chapter show a powerful way of investigating seasonality, as it provides a detailed description of how the variability has changed over time. First, the MODWT provides a visual way of informally assessing whether a constant seasonal cycle is plausible, as well as identifying additional cyclical components. Further, the wavelet power spectrum allows to formally test whether the variability of the seasonal cycle has significantly changed over the years. Finally, by using the wavelet coherency, the influence of external climatic indices can be assessed.

Chapter 4

Temporal Quantile Modelling

By means of wavelet analysis, long term trends in monthly maxima river flow were identified in Chapter 3. Given that daily data were available, analysis of these rather that the monthly data is preferred to make best use of the available information. Trends in the mean were already identified and described in Chapter 2, where a first exploratory analysis was carried out. In that same chapter, a first approach to analysis of extreme events by means of extreme value distributions was used. In particular, peak over threshold series were obtained from each river and a GP distribution fitted (under the assumption of stationarity). The results from the wavelet analysis (Chapter 3) show significant changes in the variability of the river flow series investigated. This is expected to have a direct impact on extreme flows; if the variability changes over time, it might be that the underlying distribution of river flow changes with time too, which in turn would have an effect on the quantiles of the distribution. The aim of this chapter is to investigate and assess the presence of trends in extreme river flow values. This will be achieved using quantile regression. The wavelet analysis introduced in Chapter 3 will be used for estimating the residual correlation structure.

Linear regression has the goal of estimation of the expected value of the response variable and its dependence on any set of explanatory variables. However, there might be situations in which the mean of the distribution is not informative, e.g. if one is interested in the high values of a given variable. Quantile regression (Koenker (2005)) allows estimation of the relationship between response and explanatory variables at any

percentile of the distribution of the response (conditioned on the explanatory variables). As a result, rates of change in the response variable can be estimated for the whole distribution and not only in the mean. The work presented in this chapter concentrates on the upper part of the distribution, of relevance for high river flows.

There are two approaches to quantile modelling; one is to model the conditional cumulative distribution function of the response variable and then invert it to obtain the quantiles of interest. This approach is only briefly introduced here and will not be used. The other approach, preferred for modelling river flow data, consists of building a regression model for the quantile of interest. Techniques include parametric and non-parametric methods. Parametric models are described briefly, but they are not used being too restrictive for river flow data. Instead a nonparametric quantile regression model is proposed. The conditional quantile can be expressed as $Q_Y(\tau|X = x) = F_Y^{-1}(\tau|X = x)$, where $\tau \in (0,1)$, Y is the response variable with cumulative distribution function F_Y and $X = (X_1, \ldots, X_p)$ is a vector of explanatory variables (Koenker (2005); Cade and Noon (2003)). For example, if τ=0.9, then $Q_Y(0.9|X = x)$ is the 90^{th} percentile of the distribution of $Y|X = x$. Usually the only assumption made about the distribution of the error terms is that they are independent.

Quantile models have a set of interesting properties that make them an attractive modelling choice. Quantiles are equivariant under monotonic transformations of the response variable (Koenker (2005); Koenker and Bassett (1978)), implying that the data can be log transformed without having to worry about the effect of such transformation on the model; to express the model in the original scale, it is enough to transform back the corresponding fitted model. Quantile regression is more robust to the presence of outliers in the response than mean regression (Koenker and Hallock (2000); Koenker and Bassett (1978)) and better captures the heteroscedasticity of the data (if present) by exploring the whole distribution of the response variable. However, sample size can be an issue. Sampling variation depends on the value of τ, increasing close to the extremes. Hence for a good fit in very high/low quantiles, a reasonable amount of data is necessary. Also the number of parameters needed may increase towards the extremes; in some situations "overfitting is needed to reduce the asymptotic bias" (Horowitz and Lee (2005)).

The chapter is organized as follows. First a brief introduction to quantile modelling via the cumulative distribution function is presented. The methodology is not described in detail as it will not be used to model river flow. Next quantile regression, both in the parametric and non-parametric case, is introduced. In particular, non parametric quantile modelling using P-splines is described and a fitting procedure using the penalized reweighted least squares algorithm (PIRLS) proposed. The chapter ends with an application to river flow data.

4.1 Cumulative distribution function estimation

A conditional quantile $Q_Y(\tau|X = x)$ can be estimated by inverting the cumulative distribution function:

$$Q_Y(\tau|X = x) = F_Y^{-1}(\tau|X = x)$$

This requires estimating the conditional distribution function $F(y|x)$ at every value of the vector of explanatory variables $X = x$. One approach is to estimate the conditional cdf as (Yu et al. (2003)):

$$\hat{F}(y|x) = \sum_{i=1}^{n} w_i(x) I(y_i \leq y)$$

were weights $w_i(x)$ are given only to observations below the corresponding value y of the response variable and subject to the restriction $\sum_{i=1}^{n} w_i(x) = 1$. Defined in this way, the estimator $\hat{F}(y|x)$ is not a distribution function because it is not monotone and it is not restricted to take values in the interval [0,1] (Yu et al. (2003)). To solve this problem, Hall et al. (1999) and Cai (2002) propose a weighted version of the Nadaraya-Watson estimator:

$$\hat{F}(y|x) = \frac{\sum_{i=1}^{n} p_i(x) K_h(x_i - x) I(y_i \leq y)}{\sum_{i=1}^{n} p_i(x) K_h(x_i - x)}$$

where K_h is a Kernel function with bandwidth h and $p_i(x)$ are weights that can be calculated maximizing the function $\sum_{i=1}^{n} \log p_i(x)$ subject to certain constraints (Cai (2002)). Another modification that has been proposed is to replace the indicator function with a continuous function (double-kernel approach) to avoid quantile crossing (Yu et al. (2003)). Further references as well as discussion of bandwidth selection and asymptotic results include Samanta (1989); Sheather and Marron (1990) and Hall et al. (1999). While this is an attractive approach as it does not require specification of the

relationship between response and explanatory variables, it has the downside of having to estimate the conditional cdf at every point of the set of explanatory variables, making it computationally expensive, specially for large data sets.

4.2 Regression context

The general objective function in a quantile regression model is (Koenker (2005); Koenker and Hallock (2001, 2000); Koenker and Bassett (1978)):

$$R(\beta) = \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T \beta) \tag{4.1}$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$, for $\tau \in (0,1)$, is the check function (Koenker (2005)) (see Figure 4.1), I is an indicator function and $x^T\beta$ is a linear function of the parameters $\beta = (\beta_1, \ldots, \beta_p)$. Estimating the parameters β in Equation (4.1) involves minimizing a sum of weighted absolute deviations, where the weights are asymmetric functions of τ (Cade and Noon (2003)). The function $R(\beta)$ is piecewise linear and continuous (Koenker (2005)), being differentiable at every point except at those whose residuals $y_i - x_i^T\beta$ are zero. This problem can be circumvented by taking directional derivatives (Koenker (2005)). For a model with p parameters, any regression quantile estimate will fit through at least p observations (Koenker (2005); Cade and Noon (2003)). The solution, however, is not unique, and a set of inequalities has to be set to define the 'optimal' solution (Koenker (2005)). The minimization problem defined by Equation (4.1) cannot be solved using least squares. Instead, linear programming methods like the simplex method or the interior point method (Koenker (2005); Koenker and Hallock (2000)) are used.

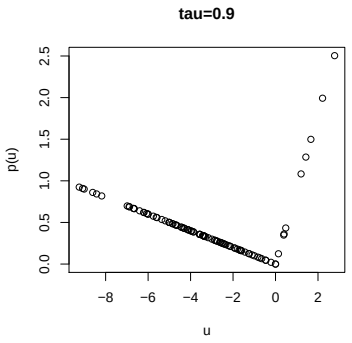

FIGURE 4.1: Illustration of the check function for $\tau = 0.9$

Quantile regression methods include parametric and non-parametric techniques. In the parametric context, the quantiles of the conditional distribution of $Y|X = x$ are expressed as linear (or non linear) combinations of the explanatory variables, and the residuals minimized (Yu et al. (2003)). Non-parametric methods include local polynomial regression and additive models. The R package `quantreg` allows a wide range of models to be fit (Koenker (2005, 2006)).

4.2.1 Parametric regression

These are regression models in which the deterministic part is parametric. Both linear and non-linear models are available. In the linear case the regression model can be expressed as:

$$Q_y(\tau|x) = \beta_0(\tau) + \beta_1(\tau)x_1 + \ldots + \beta_p(\tau)x_p \tag{4.2}$$

where $X = (X_1, \ldots, X_p)$ is a set of explanatory variables. Note that the model parameters $\beta(\tau) = (\beta_1(\tau), \ldots, \beta_p(\tau))$ are estimated for every quantile τ. The interpretation of the parameters is similar to that of the linear model (Koenker (2005); Cade and Noon (2003)). For a regression model of this form, changes across quantiles can occur either in the location, scale or shape of the response distribution (Cade and Noon (2003)). A range of techniques are available for constructing confidence intervals for the model parameters. These include rank score test inversion, direct order statistical approach, dispersion permutation test and asymptotic methods (Cade and Noon (2003)). Koenker and Hallock (2000) suggest calculating standard errors using asymptotic methods or bootstrapping. The asymptotic properties of the parameters, however, are usually based on the assumptions of iid errors and a location-shift model (i.e. changes across quantiles only affect the location parameter of the distribution, but the scale and shape parameters remain the same). In that case, the estimated parameters $\hat{\beta}(\tau)$ are asymptotically normally distributed (Schulze (2004); Koenker (2005); Koenker and Machado (1999)). If the errors are independent but not identically distributed, Koenker and Hallock (2000) recommend using the sandwich formula or an inversion of a rank test. Bootstrapping can be used for both iid or independent but not identically distributed error terms. However, none or little reference is made to the dependent errors case. Ways of assessing goodness of fit are also available (Koenker and Machado (1999)).

Applications of parametric quantile regression in economics and nutrition include Koenker and Hallock (2001) and Koenker and Hallock (2000). Both examples show how inappropriate use of mean regression can occur when the relationship between the response and the explanatory variables changes with quantile. An environmental example is presented in Sousa et al. (2009), where a comparison between multiple linear regression and quantile regression in investigating the effect of a number of explanatory variables in ozone shows that the latter is more appropriate, as the effect of the explanatory variables is not the same across all quantiles. Statistically significant effects are identified by means of quantile regression that were not detected when performing a multiple linear regression.

Parametric models are not discussed further as this thesis is centered on nonparametric methods. For additional details the reader is referred to Koenker (2005); Cade and Noon (2003) and Koenker and Machado (1999).

4.2.2 Nonparametric regression

On occasions, imposing a linear relationship might be too restrictive. Non-parametric regression allows flexible models to be built by not specifying the form of the relationship between the response and explanatory variables parametrically. In the quantile regression context, two approaches are available: local polynomial regression and additive models. The former will be introduced briefly. Then, B-splines and P-splines, used in the construction of additive models, are introduced. Finally, and additive model for river flow data in the context of quantile regression is developed and its application to the eight Scottish rivers selected in Chapter 2 discussed.

4.2.2.1 Local polynomial regression

When a regression model can not be satisfactorily expressed with a straight line, an alternative is to fit the regression model locally by using only data values in the neighbourhood of each value of the explanatory variable x_i, $i = 1, \ldots, n$. This can be done using a kernel function, that defines the weight given to each observation depending on how far/close they are from x_i (Koenker (2005)). Zhou and Wu (2009) propose a local linear quantile regression model using the Epanechnikov kernel. They argue that the

method can be applied to non-stationary series, assuming that it is locally stationary. Draghicescu et al. (2003) propose an alternative approach consisting of a non-parametric two-step method that also assumes the series to be locally stationary. The first step consists of a local linear quantile regression, for which the series is divided into blocks and a linear quantile regression is fitted to each of them. The collection of fitted lines is then smoothed in a second step using a kernel. The optimal block size is chosen based on the mean square error.

When there is more than one explanatory variable, the local polynomial regression method might become computationally expensive. In that case, an additive model approach is preferred. Before additive models are presented, B-splines and P-splines are briefly introduced, as they constitute the 'building blocks' of the additive models used in this thesis.

4.2.2.2 B-Splines and P-splines

A spline is a function defined piece-wise by polynomials. In particular, cubic splines are widely used because of their good theoretical properties, being the "smoothest possible interpolants through any set of data" (Wood (2006)). A cubic spline is a function made up of pieces of cubic polynomials joined together at a set of knots $\xi_1, \ldots, \xi_n$, that need to be specified. The resulting function is continuous and twice differentiable, and can be expressed as a linear combination of a set of basis functions. Of special interest is the B-splines basis (Wood (2006)). The order of the basis is defined as $m+1$, e.g. m=2 for cubic B-splines. The basis functions are local, so that each basis function is only non-zero over the intervals between $m+3$ adjacent knots; i.e. for cubic B-spline bases, each basis function takes values over five adjacent knots (four intervals). For a basis with k elements and order $m+1$, $k+m+1$ knots $\xi_1 < \xi_2, \ldots, \xi_{k+m+2}$ need to be specified. These are usually evenly spaced. The corresponding basis functions are defined using a recursive formula (Wood (2006)). The B-spline basis functions can be used to re-express a function $f(x)$ as (Wood (2006); Eilers and Marx (2009)):

$$f(x) = \sum_{j=1}^{k} B_j^m(x)\beta_j$$

This allows a matrix formulation of the regression model $y = B\beta$, where the columns of the design matrix B correspond to the B-spline basis functions evaluated at the explanatory variable x:

$$B = \begin{bmatrix} B_1(x_1) & B_2(x_1) & \dots & B_k(x_1) \\ B_1(x_2) & B_2(x_2) & \dots & B_k(x_2) \\ \dots & \dots & \dots & \dots \\ B_1(x_n) & B_2(x_n) & \dots & B_k(x_n) \end{bmatrix}$$

and the parameters β can be estimated by minimizing $||y - B\beta||^2$. However, the smoothness of the fitted model is highly dependent on the location of the knots and the dimension of the basis. Alternatively, Eilers and Marx (1996) proposed using P-splines. The idea is to use a large B-spline basis and add a penalty to the parameters β to control the smoothness, so that the minimization problem becomes:

$$||y - B\beta||^2 + \lambda\beta^T P\beta$$

where λ is a smoothing parameter and P is a penalty matrix. Eilers and Marx (2004) suggest creating the penalty matrix P as the product of $D_d^T D_d$, where D_d is a matrix such that when applied to β it calculates the d^{th} differences of β.

4.2.2.3 Additive models

Additive models (Hastie and Tibshirani (1990); Wood (2006)) have been extended to the quantile regression context (Koenker (2005); Hendricks and Koenker (1991); Koenker et al. (1994); Koenker (2011)). A quantile additive model (Koenker (2005)) can be expressed as:

$$Q_y(\tau|x) = \beta_0(\tau) + \sum_{j=1}^{p} g_j(x_j) \tag{4.3}$$

where $g_j(x_j)$, j=1,. . . , p are assumed to be smooth functions of the explanatory variables. As in the case of parametric quantile regression, the model can be written as a linear programming problem and the parameters estimated using a simplex method. Each of the smooth functions $g_j(x_j)$ can be parameterized using a spline basis, so that:

$$g_j(x_j) = B(x_j)\beta_j \text{ for } j = 1, \dots, p$$

Cubic splines are preferred, as they allow restrictions such as monotonicity, convexity and periodicity to be included and they are easy to estimate (Hendricks and Koenker (1991)). In particular, B-splines are used throughout this thesis. In the simplest case, where there is only one smooth function and no intercept, the model can be expressed as (Hendricks and Koenker (1991)):

$$Q_y(\tau|X) = \sum_{i=1}^{k} B_i(x)\beta_i(\tau)$$

where k is the dimension of the B-spline basis $B_i(x)$. Penalties can be added to the model to control the amount of smoothness (Koenker (2005); Koenker et al. (1994); Koenker (2011); Ng and Maechler (2007)). If the response quantile $Q_y(\tau|X) = g(x)$ where $g(x)$ is a smooth function that can be parameterized using a set of basis functions, the problem is now to minimize:

$$\sum_{i=1}^{n} \rho_\tau(y_i - g(x_i)) + \lambda \int \{g''(x)\}^2 dx \tag{4.4}$$

where $\lambda \epsilon R_+$ is the smoothing parameter. Solving the integral in Equation (4.4) poses some computational problems, which are avoided by replacing it with the total variation function of g', $V(g') = \int_0^1 |g''(x)|$ (Koenker et al. (1994)), provided that g' is absolutely continuous. The function to minimize now is:

$$R_{\tau,\lambda}(g) = \sum_{i=1}^{n} \rho_\tau(y_i - g(x_i)) + \lambda \int_0^1 |g''(x)| dx \tag{4.5}$$

over the Sobolev space W_1^2 (Koenker et al. (1994)) (the Sobolev space is a vector space of functions equipped with a norm that is a combination of L^p-norms of the function itself as well as its derivatives (up to second order in this case)). The minimum of $R_{\tau,\lambda}(g)$ cannot be found in W_1^2 (Koenker et al. (1994)) and the problem needs to be reformulated as an optimal interpolation problem (Koenker et al. (1994)). Equation (4.5) can be written as a linear program with a very sparse constraint matrix (Koenker (2011)). The function $g \epsilon U^2$ (continuous and differentiable) that minimizes $R_{\tau,\lambda}(g)$ is a "piecewise linear spline with knots at the observed x_i (in the univariate case)" (Koenker (2011); Koenker et al. (1994)). However, while writing down the penalty matrix is relatively easy in the context of mean regression, in the case of quantile regression is not as intuitive.

4.2.2.4 Smoothing parameter choice

A number of approaches are available in the literature for selecting the smoothing parameter λ. Koenker et al. (1994); Koenker (2011) and Horowitz and Lee (2005) propose using the value λ that minimizes the Schwarz information criterion (SIC):

$$SIC(\lambda) = \log\left[\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(y_i - \hat{g}(x_i))\right] + \frac{1}{2n}df_\lambda \log n$$

where df_λ are the effective degrees of freedom of the model, calculated as the trace of the hat matrix. However, Reiss and Huang (2012) argue that this method "has never been rigorously justified for nonparametric quantile regression". Alternative approaches include approximate cross-validation and multifold cross-validation, but these appear to overfit for extreme quantiles (Reiss and Huang (2012)). Reiss and Huang (2012) propose a likelihood based approach. Selection of smoothing parameters is outside the scope of this thesis and is not further discussed here. The reader is referred to Reiss and Huang (2012) for a summary of available methods and detailed explanation of the likelihood approach. For the purpose of this thesis, smoothing parameters selection is based on a sensitivity analysis, in which model adequacy is assessed graphically (see Section 4.4.3.1).

4.2.3 Time series modelling

The extension of quantile regression models to the time series context is an ongoing subject of research. Over the last few years, auto-regressive quantile regression models that incorporate the temporal dependence of the data have been proposed (Galvao et al. (2009); Xiao and Koenker (2009); Draghicescu et al. (2009); Cai (2002); Koenker and Xiao (2006)). Koenker and Xiao (2006) propose a class of quantile autoregressive models of order p, denoted as QAR(p), that can be expressed as:

$$Q_{y_t}(\tau|y_{t-1},\ldots,y_{t-p}) = \alpha(\tau) + \beta_1(\tau)y_{t-1} + \ldots + \beta_p(\tau)y_{t-p} \tag{4.6}$$

Model (4.6) allows for changes in the location, scale and shape of the distribution of the response variable y_t. The coefficients $\theta_\tau = (\alpha(\tau), \beta_1(\tau), \ldots, \beta_p(\tau))$, which can be interpreted as the coefficients of a random coefficient model, are not assumed to be

independent random variables but to be functionally dependent via a single random uniform(0,1) variable. Galvao et al. (2009) extend the work of Koenker and Xiao (2006) suggesting a quantile autoregressive distributed lag model (QADL). This model is based on the linear time series model, where the response is expressed as a linear combination of past values of the response variable itself and a set of covariates. Cai (2002) proposes including y_{t-1} as an explanatory variable in a non parametric time series model and then estimating the cdf using a weighted version of the Nadaraya-Watson estimator. Quantiles are then obtained by inverting the estimated cdf function. Draghicescu et al. (2009) propose a method for dealing with non-stationary time series by using localized fitting. The data are divided into blocks that are assumed to be stationary and then a quantile regression model is fitted within each block. Once all the fits from the different blocks are pulled together a second round of smoothing is carried out using a kernel. Interpretation of these models is somewhat different from a regression model, as now the explanatory variables describe how the conditional quantile changes through time.

Time series models available in the quantile regression literature include the lagged response variable as an explanatory variable. However, this is not a suitable approach for the river flow data; as shown in Chapter 2, the large lag one correlation coefficient (close to one) and the presence of long range dependence suggest that traditional time series auto regressive models might not be appropriate.

4.2.3.1 Long range dependence in the quantile context

Ghosh et al. (1997) propose a non-parametric approach using kernel smoothing. The long range dependence is accounted for by including the Hurst exponent (see Chapter 2, Section 2.4) in the optimal bandwidth selection expression. The Hurst exponent is estimated using the residuals from a mean regression that removes the trend. The cumulative distribution function is estimated as:

$$\hat{F}_x(y) = \frac{1}{nb_n} \sum_{i=1}^{n} K\left(\frac{x_i - x}{b_n}\right) I\{Y_i \leq y\}$$

where the bandwidth b_n is selected as:

$$b_n = Cn^{m(2H-2)/(4+m(2-2H))}$$

C and m are constant and H is the Hurst exponent.

At the completion of this thesis, the only reference found in the literature for fitting a quantile model in the presence of long range dependence is that of Ghosh et al. (1997).

4.3 Environmental Examples

Most of the theory developed for quantile regression has been motivated by its application in economics (Buchinsky (1998); Horowitz and Lee (2005); Hendricks and Koenker (1991)). Other fields of application, particularly using non parametric models, include biology (Koenker et al. (1994)) and nutrition (Koenker (2011)). Recently, interest in quantile regression has grown in the environmental community, motivated by an increasing interest in extreme values and recognizing the usefulness of studying high quantiles rather than just averages. Draghicescu et al. (2003) and Draghicescu et al. (2009) apply non-parametric quantile regression to ozone concentrations. Zhou and Wu (2009) apply local linear quantile regression to global temperature data and precipitation. Applications in hydrology appear to be limited, the only one found being Sankarasubramanian and Lall (2003), who investigate the relationship between annual flow values and large-scale climatic indices, and no extensive analysis has been carried out on river data. The work presented in this chapter contributes to the current literature by developing a quantile model suitable for river flow data. The model is built on an additive regression framework and the parameters are fitted using a penalized least squares approach.

4.4 A quantile regression model for river flow

The model proposed here builds upon the work of Reiss and Huang (2012). Instead of using linear programming methods to estimate the model parameters, a weighted least squares approach is preferred. Note that the objective function (4.1) is nothing but a weighted sum of absolute residuals. Reiss and Huang (2012) suggest approximating the absolute residuals with the squared residuals, so that least squares methods can be applied. This has the advantage that standard regression methods can easily be extended

to the quantile regression context in a GAM setting. Reiss and Huang (2012) concentrate on smoothing parameter selection and do not discuss at any point the quantile model itself. Here an additive quantile model for river flow data is proposed. Appropriate penalty matrices for each of the terms included in the model are specified and the correlation is incorporated in the inference process.

The proposed model is:

$$Q_{log(flow)_i}(\tau|x) = s_1(x_i) + s_2(x_i) + \varepsilon_i \tag{4.7}$$

where x_i represents time, i=1,...,n and $s_1(x_i)$ and $s_2(x_i)$ are smooth functions of time corresponding to the trend and seasonal components. At this point the errors ε_i are assumed to be independent, but no distributional assumption is imposed. Once the deterministic part of the model has been fitted, the residual correlation structure will be investigated and incorporated into standard error estimation. The trend and seasonal components are parameterized using large B-spline bases with k_1 and k_2 bases functions respectively:

$$s_1(x) = \sum_{j=1}^{k_1} B_{1,j}(x)\beta_{1,j}$$
$$s_2(x) = \sum_{j=1}^{k_2} B_{2,j}(x)\beta_{2,j}$$

A first order difference penalty $P_1 = \sum_{j=2}^{n}(\beta_{1,j} - \beta_{1,j-1})^2$ is used for the long term trend (Eilers and Marx (2009)), while the circular penalty $P_2 = \sum_{j=3}^{n}(\beta_{2,j} - 2\phi\beta_{2,j-1} + \beta_{2,j-2})^2$ is imposed for the seasonal term following Eilers and Marx (2009), where $\phi = \cos(2\pi s/p)$, p is the period of the data (365 days) and s is the knot distance. The corresponding smoothing parameters λ_1, λ_2 need to be specified. These penalties were chosen so that in the limit (i.e. for very large smoothing parameters λ_1, λ_2) a constant (i.e. no change) and seasonal fits are obtained for the trend and a seasonality respectively. Also, the circular penalty P_2 ensures that the seasonal fit at the end of one year and the beginning of the next year join smoothly.

The model is fitted using the back-fitting algorithm (Wood (2006)) (see Section 4.4.1), so that each smooth function s_1, s_2 is estimated in turn. What follows describes how to fit the model for a single smooth function. The subindexes 1 and 2 that identify each of

the components are suppressed.

Assuming the smoothing parameter λ to be fixed, the model parameters β can be estimated (individually for each of the components) by minimizing:

$$\hat{\beta} = argmin_{\beta \epsilon R^k} \left[\sum_{i=1}^{n} \rho_\tau (y_i - B(x_i)^T \beta) + \lambda \beta^T P \beta \right]$$

where $B(x_i)$ is the B-spline basis evaluated at x_i and P is the penalty matrix. This can be translated into a penalized least squares problem (Reiss and Huang (2012)), that can be solved using penalized iterative reweighted least squares (PIRLS) (Wood (2006); Pratesi et al. (2006)), where the weights at iteration (j) are given by:

$$w_i^{(j)} = \frac{\tau - I\left[(y_i - B(x_i)^T \hat{\beta}^{(j)}) < 0\right]}{2(y_i - B(x_i)^T \hat{\beta}^{(j)})} \tag{4.8}$$

for $i = 1, \ldots$,n. A large upper bound is set for the weights to avoid residuals close to zero (Reiss and Huang (2012)), which would result in the check function not being differentiable. Truncating all weight values above the given upper bound does not have a big effect on the fitting process, as we are just creating very small residuals, whose contribution to the objective function is negligible, even smaller. The estimated parameters $\hat{\beta}$ at each iteration (j) are:

$$\hat{\beta}^{(j)} = (B^T W^{(j-1)} B + \lambda P)^{-1} B^T W^{(j-1)} y \tag{4.9}$$

where $B = (B(x_1), \ldots, B(x_n))^T$, $y = (y_1, \ldots, y_n)^T$ and W is an n×n diagonal weight matrix.

The expression for calculating the weights is obtained as follows. The parameter estimates at iteration $(j+1)$ are obtained based on the estimated $\hat{\beta}^{(j)}$ at iteration (j) by solving:

$$\hat{\beta}^{(j+1)} = argmin_\beta \left[\sum_{i=1}^{n} w_i^{(j)} (y_i - B(x_i)^T \beta)^2 + \lambda \beta^T P \beta \right]$$

Where the weights $w_1^{(j)}, \ldots, w_n^{(j)}$ satisfy

$$\sum_{i=1}^{n} 2w_i^{(j)}(y_i - B(x_i)^T\beta)(-B(x_i)) + 2\lambda P\beta = 0$$

Assuming there are no zero residuals, then:

$$\sum_{i=1}^{n} \left[\tau - I(y_i - B_i^T\beta < 0)\right](-B_i) + 2\lambda P\beta = 0$$

Re-arranging, Equation (4.8) is obtained.

The estimation algorithm can be summarized as follows:

1. Select initial estimates $\beta(\tau)^{(0)}$. These can be obtained from a mean regression.
2. At each iteration (j), calculate the residuals $e_i^{(j-1)} = y_i - B(x_i)\beta^{(j-1)}$ and associated weights $w_i^{(j-1)} = \frac{\tau - I(e_i^{(j-1)} < 0)}{2e_i^{(j-1)}}$ from previous iteration.
3. Calculate the new weighted penalized least squares estimates:

$$\beta^{(j)} = (B^T W^{(j-1)} B + \lambda P)^{-1} B^T W^{(j-1)} y$$

Iterate steps 2,3 until convergence.

4.4.1 The back-fitting algorithm

The trend and seasonal pattern are estimated individually using the back-fitting algorithm (Wood (2006)). Hence, at each step of the algorithm, model estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ are obtained following the PIRLS algorithm detailed above. The proposed model (4.7) has two additive components. In this situation, the back-fitting algorithm is relatively simple. It works as follows:

1. set $\hat{s}_i = 0$ for i=1,2.
2. For i=1,2, repeat:

3. Calculate the partial residuals $e_p^i = y - \sum_{k \neq i} \hat{s}_k$, where $\hat{s}_k$ is estimated using the PIRLS method described above.
4. Set $\hat{s}_i = e_p^i$

Steps 2 to 4 are repeated until convergence. Let S_i be the smoothing matrix of the i^{th} component. Then $\hat{s}_i \neq S_i y$; however, it is possible to find a matrix $\tilde{S}_i$ such that $\hat{s}_i = \tilde{S}_i y$. The fitted values can then be expressed as $\hat{y} = \tilde{S} y$ where $\tilde{S} = \sum_{i=1}^{p} \tilde{S}_i$. Following Giannitrapani (2006), the $\tilde{S}_i$ can be derived as follows. At the first step of the back-fitting algorithm:

$$\begin{aligned}\hat{s}_1^{(1)} &= S_1 y \\ \hat{s}_2^{(1)} &= S_2[Id - S_1]y = S_2[Id - P_2^{(1)}]y\end{aligned}$$

where Id is the identity matrix. At the second step:

$$\begin{aligned}\hat{s}_1^{(2)} &= S_1[Id - S_2(Id - P_2^{(1)})]y = S_1[Id - P_1^{(2)}]y \\ \hat{s}_2^{(2)} &= S_2[Id - S_1(Id - P_1^{(2)})]y = S_2[Id - P_2^{(2)}]y\end{aligned}$$

and so on. Giannitrapani (2006) shows that the following recursive formula can be used:

$$P_i^{(l)} = S_i \left[Id - \sum_{k>i} P_k^{(l-1)} + \sum_{k<i} P_k^{(l)} \right]$$

where $P_i^{(0)} = Id$. At step (l), $\hat{m}_i^{(l)} = P_i^{(l)} y$.

Hence, for model (4.7) the $(\mathrm{j})^{th}$ step of the algorithm becomes:

$$\begin{aligned}\hat{s}_1^{(j)} &= P_1^{(j)} y \\ \hat{s}_2^{(j)} &= P_2^{(j)} y \\ P_1^{(j)} &= S_1(Id - P_2^{(j-1)}) \\ P_2^{(j)} &= S_2(Id - P_1^{(j)})\end{aligned}$$

4.4.2 Pointwise confidence bands for fitted values

The following theorem (Searle, 1971) will be used:

Theorem 1. If $x \sim N(\mu, V)$, then $E(x^T A x) = tr(AV) + \mu^T A \mu$

Confidence intervals for fitted values were computed accounting for the temporal correlation in the residuals. The degrees of freedom associated with the error term for a least squares based regression can be approximated by (Giannitrapani (2006)):

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

where n is the total number of observations. In the case of a weighted least squares, the residual sum of squares can be expressed in matrix form as:

$$RSS = y^T(I - S)^T W(I - S)y$$

Using Theorem 1 and assuming observations y are independent with variance σ^2, it follows that:

$$E[RSS] = tr[(I - S)^T W(I - S)\sigma^2] = \sigma^2 tr[(I - S)^T W(I - S)] + \mu^T(I - S)^T W(I - S)\mu$$

Hence, the degrees of freedom of the error term can be calculated as:

$$df_{error} = tr[(I - S)^T W(I - S)] = tr[W - WS - S^T W - S^T WS]$$

Incorporating the weight matrix means that the confidence intervals get wider as we move towards extreme quantiles. The weight matrix is calculated using the expression from the PIRLS algorithm:

$$w_i = \frac{\tau - I(y - \hat{y} < 0)}{2(y - \hat{y})}$$

where $\hat{y}$ are the fitted values obtained from the last back-fitting iteration.

The standard error of the fitted values $\hat{y}$:

$$se(\hat{y}) = \sqrt{var(Sy)} = \sqrt{diag(SS^T)\sigma^2} \tag{4.10}$$

where S is the smoothing matrix obtained from combining the individual smoothing matrices from each term in the back-fitting algorithm so that $\hat{y} = Sy$. The variance σ^2 can then be estimated as:

$$\hat{\sigma}^2 = \frac{RSS}{df_{error}}$$

Expression (4.10) was derived under the assumption of independence. Hence, it has to be adjusted for correlated observations, as is the case of the river flow data. Assuming $var(y) = V\sigma^2$, where V is the correlation matrix (that can be estimated using the residuals from the quantile regression):

$$se(\hat{y}) = \sqrt{var(Sy)} = \sqrt{diag(SVS^T)\sigma^2} \tag{4.11}$$

The residual sum of squares, degrees of freedom and σ^2 estimator need to be adjusted for correlation too. For correlated errors with correlation matrix V:

$$RSS_c = y^T(I-S)^T V^{-1} W(I-S)y \tag{4.12}$$

The expected value:

$$E[RSS_c] = tr[(I-S)^T V^{-1} W(I-S)V]\sigma^2$$

And hence the associated degrees of freedom:

$$df_{error.c.} = tr[(I-S)^T V^{-1} W(I-S)V] \tag{4.13}$$

and

$$\hat{\sigma}^2 = \frac{RSS_c}{df_{error.c.}}$$

4.4.3 Results

Model (4.7) with τ=0.95 was fitted to the eight Scottish rivers presented in Chapter 2. To avoid repetition, only one of them will be discussed in detail. This corresponds to the River W. of Minnoch (gauging station 81006). Rich B-spline basis with k_1=45 and k_2=75 basis functions were selected for the trend and seasonal component respectively, with first order and circular penalties as described in Section 4.4. During the PIRLS fitting process, the upper bound for the weights was selected to be 10. Its effect on the fitted values was investigated by varying it and refitting the model but no difference was found in the fitted values.

4.4.3.1 Choice of smoothing parameters

A number of methods for choosing the smoothing parameter have been proposed. However, none of methods performs particularly well when observations are correlated. The smoothing parameters $\lambda = (\lambda_{trend}, \lambda_{season})$ were selected individually for each component based on the effective degrees of freedom of the model (Hastie and Tibshirani (1990); Bowman et al. (2009)), calculated as the trace of the hat matrix (Wood (2006)). A sensitivity analysis, where the adequacy of the model was assessed graphically, was carried out in order to decide how many degrees of freedom would be necessary for each component. These were chosen to be $df = (df_{trend} = 15, df_{season} = 45)$. A rather simple argument to aid with this decision was followed. There are about 19 years worth of data, so initial values of degrees of freedom could be based on the assumption that there are 19 parameters needed for the trend component (one for each year) and about 57 parameters for the seasonal component (three per year). The model was first fitted using these initial values. Smaller and larger values (10-25 for the trend component and 40-60 for the seasonal component) were considered and finally the values $df = (df_{trend} = 15, df_{season} = 45)$ were chosen.

The high number of degrees of freedom required can be explained by a complex non-monotonic trend and a seasonal pattern that changes from year to year. Also, estimating high quantiles of the distribution requires more flexibility than estimating the mean (there is more variation in the higher quantiles).

The resulting fitted model is shown in Figure 4.2. Goodness of fit was informally assessed based on the proportion of positive (5.33%) and negative (94.67%) residuals, very close to the expected proportions of 5% and 95% respectively. The individual trend and seasonality terms are plotted in Figure 4.3. Despite looking at extreme values, there is still a strong seasonality present. The seasonal cycle varies slightly in amplitude and phase from year to year, with a very pronounced cycle around 2001. A non monotonic decreasing trend from the beginning of the record until about 2003 can be seen, with a peak around 1998.

4.4.3.2 Residual correlation structure estimation

The quantile regression model was fitted assuming independent errors. However, this is unlikely to be the case when dealing with daily data. The MODWT (see Chapter 3) was used to estimate the Hurst parameter (see Chapter 2) from the residuals of the quantile fitted model and hence obtain an estimate of the correlation matrix. An issue with the wavelet based estimation that, to the authors knowledge, is usually not discussed in the literature, is the level of decomposition. Percival and Walden (2006) suggest choosing the maximum level of decomposition J so that the wavelet variance at the corresponding scale has a sufficiently large effective degrees of freedom. Here the maximum level of decomposition was chosen as follows. We have already accounted for variations corresponding to scales 8 (2^7=128 days) and higher with the seasonal and trend components, hence we are interested in the correlation structure of lower scales. As well, we do not expect any significant correlation at higher scales as it has already been "removed" with the model. That means that in the estimation of the Hurst parameter scales up to 7 will be considered. This yields an estimate of $\hat{H}$=0.59. The first and second scales, corresponding to one and two days, are not included in the estimation process, as they can be associated with short term dependence.

The correlation was estimated to be:

$$\hat{\rho}(k) = \frac{1}{2}\left\{(k+1)^{2\hat{H}} - 2k^{2\hat{H}} + |k-1|^{2\hat{H}}\right\}$$

where $\hat{H}$ is the Hurst parameter estimated from the residuals using the wavelet based method and assuming that the underlying process is a fGn. Under long range dependence, however, the correlation never reaches the value zero, even for observations far apart. This implies that the correlation matrix will never be sparse, with the consequent computational cost. A tolerance value had to be set for computational issues. This was set to 10^{-5}.

4.4.4 River comparison

The set of rivers described in detail in Chapter 2 are further investigated here. A quantile regression model was fitted to each of them individually, with the aim of comparing their seasonal and trend components. The model was set using 40 and 75 cubic B-splines bases

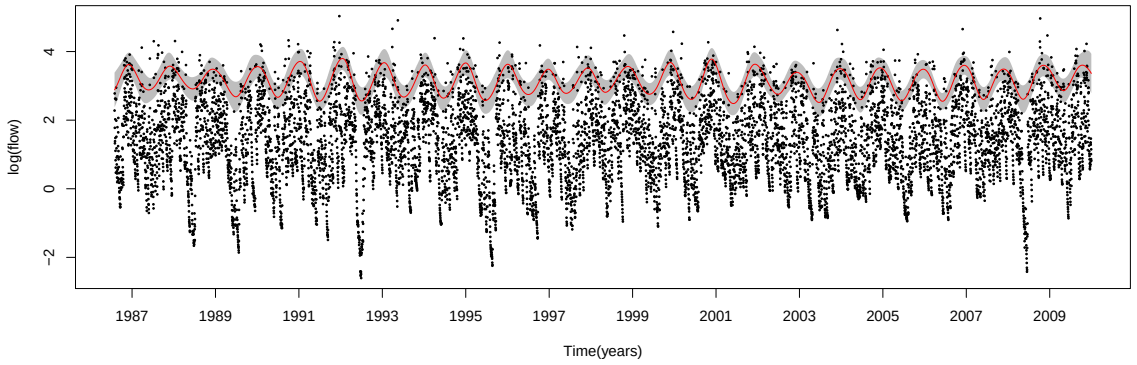

FIGURE 4.2: 95^{th} quantile fitted model (red) and approximate 95% pointwise confidence bands (grey bands) assuming long range dependence with Hurst exponent $\hat{H}$= 0.57. Units are in log($m^3/3$). Water of Minnoch (gauging station 81006)

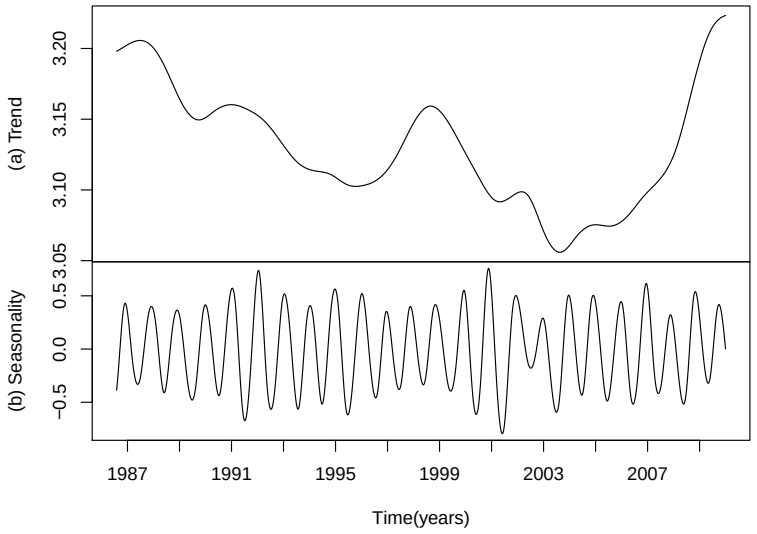

FIGURE 4.3: 95^{th} quantile trend (top) and seasonality (bottom). Units are in log($m^3/3$). Water of Minnoch (gauging station 81006)

functions for the trend and seasonal components respectively. The degrees of freedom were set to 15 and 45 respectively. For a river series with 15695 observations (Water of Leith), fitting the model (without calculating the full smoothing matrix) takes about 32 hours on a CPU with 3.3G. For longer series, such as the River Tay (20471 observations) memory storage problems appear. To reduce the computational cost slightly, analysis was restricted to the period of time in common for all eight rivers (1^{st} August 1986-31^{st} December 2005). Each resulting series has 7088 observations.

The fitted trends for each river are shown in Figure 4.5. Overall, the River Tay shows a slightly upwards trend, while the Water of Minnoch shows a decreasing one. For the

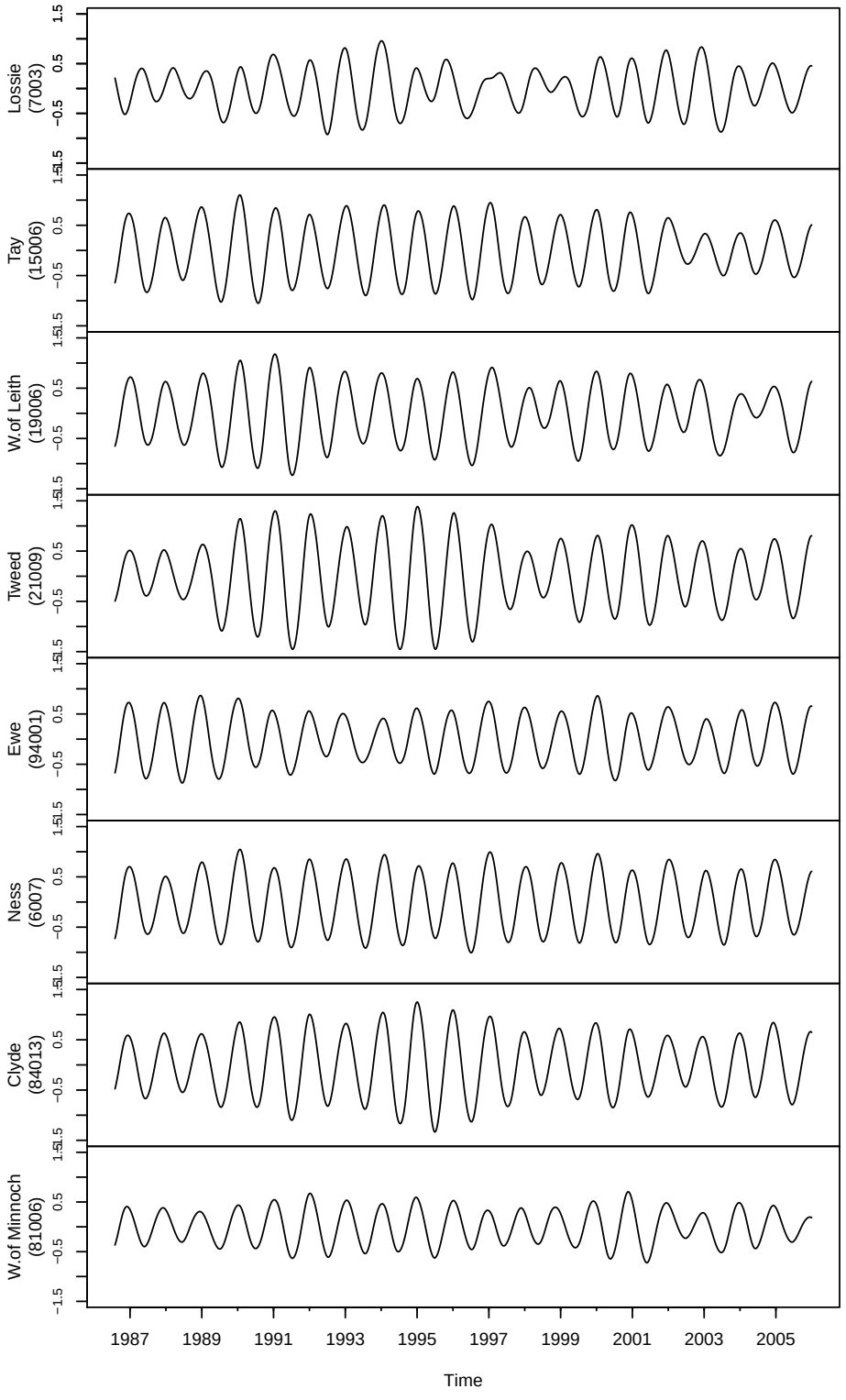

FIGURE 4.4: Seasonal component from the 95^{th} quantile fitted model for all eight rivers. Units are in $\log(\mathrm{m}^3/3)$

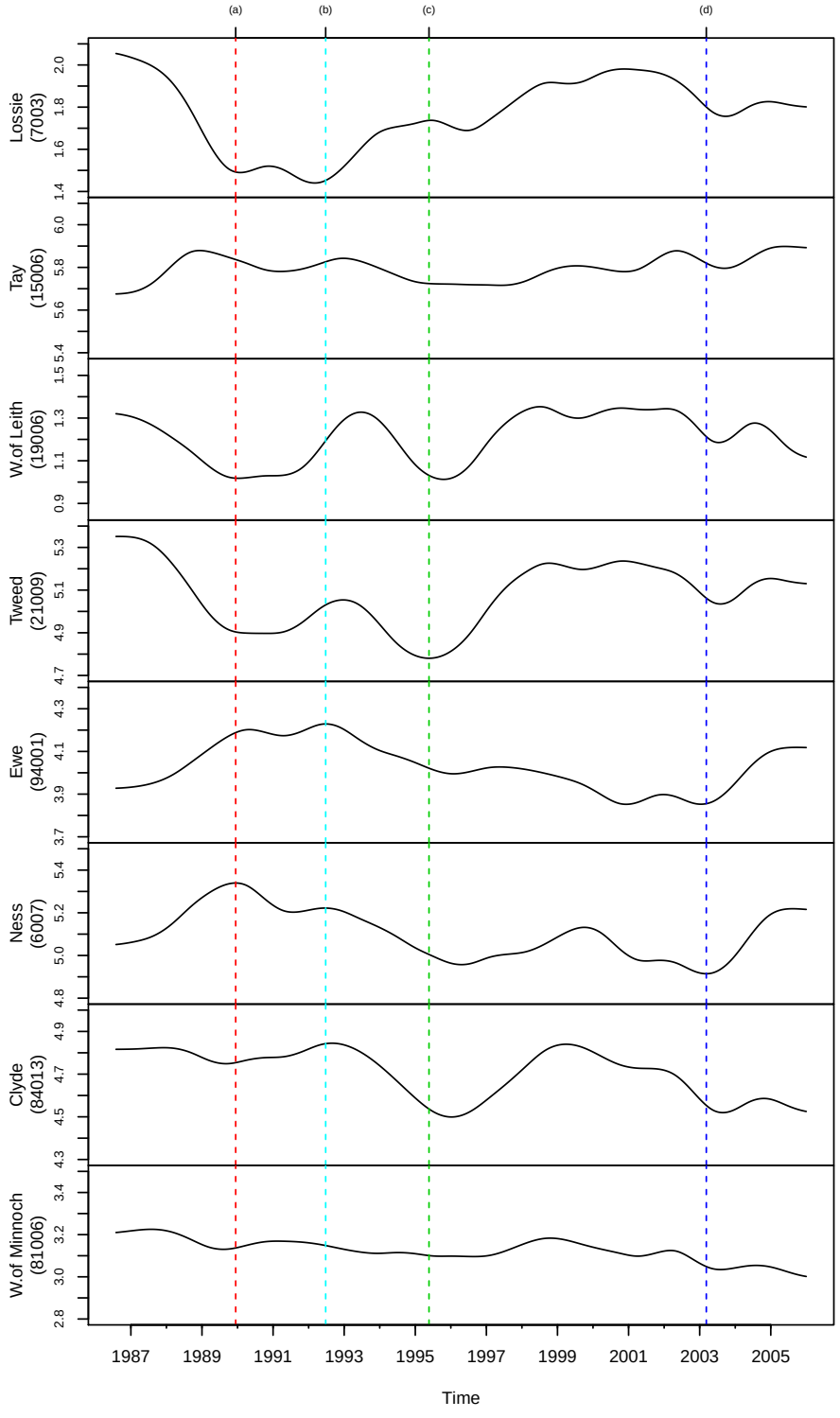

FIGURE 4.5: Trend component from the 95^{th} quantile fitted model for all eight rivers. Units are in $\log(m^3/3)$. Note the scale on the y axis varies across rivers. The vertical lines (a), (b), (c) and(d) highlight particular features of the data and are referred to in the text

remaining rivers, the fitted trend is not monotonic and includes increasing and decreasing periods. Rivers Ewe and Ness on one hand and Water of Leith and Tweed on the other show similar patterns. Interestingly, rivers in the East, namely the Water of Leith, Tweed and Lossie show a decreasing trend until about December 1989 ((a) in Figure 4.5) which flattens out for the Water of Leith and Tweed and keeps on decresing until June 1992 for the Lossie ((b) in Figure 4.5). On the other hand, the rivers Ewe and Ness, on the West, show an increasing pattern at the same time, while the trend for the Clyde is fairly constant in that time period. From 1991 onwards rivers Ewe and Ness show a steady decrease until about March 2003 ((d) in Figure 4.5), while the trend of the Lossie increases. The Water of Leith and Tweed reach a minimum around May 1995 ((c) in Figure 4.5) to then increase again and flatten out. A similar decrease can be seen for the Clyde, only slightly later in time. The River Clyde, despite being on the West, behaves more similar to rivers Tweed and Water of Leith on the East than to rivers on the West.

The fitted seasonal patterns are shown in Figure 4.4. The estimated seasonal pattern is more pronounced in large catchments than in small ones, with the exception of the Water of Leith which despite being a small catchment shows a strong pattern specially during 1990-1992. Rivers Ewe, Ness and Water of Minnoch show a seasonal pattern that is fairly constant over time, while rivers Lossie and Tweed show the greatest variability amongst all rivers. In particular, the seasonal cycle of the Lossie is different during 1997-1999, with both the phase and amplitude of the cycle changing with respect to the rest of the years. River Tweed shows an increased amplitude cycle during 1990-1996. This feature can also be seen in the river Clyde although not as pronounced and in the Water of Leith during 1990-1993. The seasonal cycle of the Tay is very stable apart from the end of the period (from 2002 onwards) when the amplitude decreases.

4.5 Summary and discussion

Quantile estimation provides a natural way of assessing variability in time series, by allowing the whole conditional distribution to be described in detail and providing added information that would not be available in a mean regression. By fitting models to high quantiles, further insight into the behaviour of extreme values can be gained. Extreme value analysis has been the main tool for producing risk estimates over years, and even

though the theory is well established it relies on a set of assumptions that might be dubious given current conditions. In particular, in Scotland, it is the magnitude and not the frequency of floods that seems to be the problem, with a small number of large floods having a profound impact (Black and Burns (2002)), specially in the West of the country. Hence it is important to assess trends in extreme values directly, so that changes in magnitude of extreme flows can be identified. Extreme value analysis aims to provide estimates of flood risk based on a given quantile of the response variable. Quantile regression allows exploration of how the corresponding quantile changes with time, hence directly relates to the problem of assessing trends in extreme values. If there is clear evidence of the quantile not being constant, then that might suggest that the extreme value based estimate is not reliable, as the assumption of stationarity is being violated.

The results from the fitted model show that the $95th$ quantile of the distribution of river flow changes over time in all eights rivers investigated, as illustrated by the trends and seasonal patterns identified. This suggests that the model parameters and return value estimates presented in Chapter 2 (Section 2.5) might not be reliable, as they were estimated based on methods for stationary series. If the stationarity does not hold, the extreme value methods described in Section 2.5 are subject to "unbounded and unquantifiable bias" (Eastoe and Tawn (2009)). A possible solution would be to remove the non-stationarity by subtracting the trend and seasonal patterns identified by means of quantile regression and then carry out a extreme value analysis.

4.5.1 Hydrological findings

The proposed model was applied to the eight rivers selected in Chapter 2. To make models comparable, the same number of basis functions and degrees of freedom were chosen for all eight rivers. The results suggest some differences in the long term trend between the East and the West, with increasing and decreasing periods being almost opposite in time. The seasonality seems to be more variable in the East than in the West, in agreement with published studies that suggest higher variability in the East, and also with rainfall patterns being more variable in the East as described in Chapter 2.

4.5.2 Statistical issues

In this chapter a quantile model for river flow is developed. The weighted least squares approximation of the objective function for a quantile regression model allows the parameters to be fit in an easy and efficient way. The problem of the check function not being differentiable at the origin is circumvented by avoiding zero residuals. Another advantage is that inclusion of covariates is straightforward. Further, the model can be easily extended to incorporate other modelling techniques already available for mean regression. By fitting the model in an additive non-parametric framework a great amount of flexibility is achieved. The use of P-splines makes extension to higher dimensional datasets feasible.

Even though application of the model has only been illustrated here for the 95^{th} quantile, the model can be directly applied for any other quantile of interest (within the centre of the distribution). This raises the issue of quantile crossing. Quantile crossing has not been directly investigated as the main objective was to evaluate changes in extreme values, rather than to model the whole distribution of river flow. Nevertheless, it is expected that crossing might happen if the smoothing parameters are not appropriate (and hence the model under or oversmooths the quantile of interest).

To date, no gold standard method for choosing the smoothing parameter λ is available, and, while a number of methods have been proposed, these are known to fail when data are correlated. Smoothing parameter selection was then based on personal judgement after visual inspection of a range of fitted models with various degrees of freedom and assessment of the proportions of positive and negative residuals. Choosing smoothing parameters or degrees of freedom are equivalent ways of deciding the amount of smoothing of a model. However, degrees of freedom are somehow more intuitive as they can be directly related to the number of parameters of the model. This means that when a full sensitivity analysis for assessing the effect of the smoothing parameter is not possible, as was the case here due to high computational cost, one can support the choice of λ based on the degrees of freedom in the model (Hastie and Tibshirani (1990); Bowman et al. (2009)). This represents a clear advantage of the developed model with respect to the fitting procedure that uses simplex methods, in which case the interpretation of

the smoothing parameter is not clear. A criteria that was used to help with the choice of degrees of freedom was the proportion of positive and negative residuals. An insufficient number of parameters (oversmoothing) caused the proportions to be far from the expected values of 5% and 95% respectively.

The model was fitted assuming independence of observations, and the temporal dependence was incorporated in the inference process. The residual correlation structure was characterized using a long memory process. This is a reasonable assumption for daily river flow data. Further, the large lag-one sample autocorrelation coefficient (close to one) suggests that it is not advisable to use ARMA(p,q) models to characterize dependence. Approximate 95% pointwise confidence bands were calculated assuming the residuals to be normally distributed. Assumption of normality seemed reasonable from histograms and qqplots, with the difference that instead of having mean zero, the 95% quantile of the residuals is zero. The structure of the residuals is an interesting topic that has not been explored in the literature, where most of the quantile regression models assume independent and identically distributed observations, but no assumption is made on the distribution of the errors other than independence. This means that the 'traditional' goodness of fit tools based on residuals available for mean regression cannot be used here, making model assessment difficult. A way of assessing if the fitted model is appropriate is to compare observed and fitted quantiles, conditioning on a value of the explanatory variable, for a range of quantiles (Lee and Neocleous (2010)). Deviations from a straight line in the plot of observed versus fitted quantiles suggests that the model is not adequate. However, this was not feasible as the computational cost of the fitted model for a single quantile is relatively high and only one observation is available for every covariate (time) value.

The model proposed in this chapter includes two additive terms, a trend and a seasonal component that is allowed to change from year to year. Both the trend and the seasonality are expressed as smooth functions of time. A large number of degrees of freedom were required to fit the proposed model. This is due to the large flexibility of the model itself and the high quantile (95^{th}) for which the model was fitted.

One possible application of the proposed model could be to use the fitted quantile model as a time-varying threshold for POT series. Some examples of this application can be found in the literature (Kyselý et al. (2010); Northrop and Jonathan (2011)). Threshold selection is an open problem, with no gold-standard rule currently available. In some cases, especially under nonstationarity, threshold choice might be problematic. Northrop and Jonathan (2011) argue that if extremes are not stationary, then there is no reason for using a constant threshold.

Traditional extreme value models (see Chapter 2, section 2.5) assume a constant distribution of extremes over time (i.e. stationarity). This assumption might no longer hold for extremes of non-stationary series. Hence, when strong patterns (such as long term trends and seasonal cycles) are present in the data, using a constant threshold might prove inadequate. Instead, the use of a time varying threshold has been suggested (Coles (2004)). As Northrop and Jonathan (2011) point out, using a constant threshold when the time series is not stationarity has some implications, especially when extremes are modelled depending on some covariates. It is possible that a threshold that is high enough for a particular covariate value might not be for another. This is particularly important when the aim is to estimate a covariate effect, in which case it is desired to have exceedances for the whole range of covariates values, something that might not happen when a constant threshold is used (Northrop and Jonathan (2011)).

Kyselý et al. (2010) use a time-dependent threshold to model extreme temperature. The threshold is set as the fitted model of a 95% quantile regression on time, where time is a linear or quadratic term. Northrop and Jonathan (2011) model hurricane-induced wave heights over space, setting a threshold at each location that is a function of the covariates (longitude and latitude). By using a quantile regression model, Northrop and Jonathan (2011) ensure that the probability of exceedance is the same across all locations. Excesses are then modelled using a point process representation (Coles (2004); Northrop and Jonathan (2011)) to avoid the problem of threshold-dependence of the GP distribution parameters.

Alternative ways of dealing with non-stationarity have been suggested. Some studies

restrict analysis of extreme values to the winter season to avoid having to deal with non-stationarity and on the basis that it is the period when most of the extremes happen. It is important though to look at the whole year as extreme values do also happen in summer months. The quantile approach allows to work with the whole year and directly model the seasonal effect. Zhou and Wu (2009) propose a way of dealing with non-stationary series, by assuming that they are locally stationary. This means that a local quantile regression would work, by applying it to sections of the data that are stationary. However, their asymptotic theory is based on the assumption that the correlation function is summable, which means that the theory fails when dealing with long-memory processes as is the case here. Eastoe and Tawn (2009) propose pre-processing the data to make it stationary (or as stationary as possible) instead of using a time-varying threshold, so that return levels below the chosen threshold can be estimated. A comparison between the standard approach, pre-processing the data and using a time-varying threshold suggest an improvement in estimation for the latter two, which show similar results. Eastoe and Tawn (2009) argue that a time varying threshold is not able to "distinguish the covariate effects that are found in the GPD parameters between those which affect the centre of the distribution and those which affect the tails". However, this drawback could be avoided by including the covariates in the quantile regression model used for selecting the time varying threshold, which can be done easily using the model proposed here.

Chapter 5

Spatial Modelling of Extreme River Flow

The results from the quantile modelling of river flow in Chapter 4 suggested some differences amongst the eight rivers considered that might be due to their geographical location. The aim of this chapter is to study the spatial pattern in extreme river flow values across Scotland. Two different approaches are explored. The first one concentrates on estimating conditional probabilities of exceedance following the work of Keef et al. (2009). The second one is a spatial quantile regression model that describes the spatial trend in extreme river flow values across Scotland, using spatial location as a covariate. A full spatio-temporal model is not provided. Instead, the proposed spatial quantile regression model is fitted independently over time on a monthly basis. The chapter starts with a review of spatial methods for modelling extremes. Then the conditional probability approach is presented, and two examples of its application discussed. The spatial quantile regression model is introduced next as an extension of the temporal quantile model presented in Chapter 4. Monthly spatial trends for the 95^{th} quantile are presented for the set of Scottish rivers available over the period 1996-2005. The chapter finishes with a section that briefly outlines how a full spatio-temporal quantile model could be constructed.

5.1 Models for spatial extremes

Most of the literature available on spatial analysis concentrates on mean regression. Geostatistical models (see, e.g. Diggle and Ribeiro Jr. (2007)) assume spatial data to be a realization of an underlying spatial Gaussian process or Gaussian random field $\{Z(s), s \in D\}$, $D \subseteq R^p$, where s represents spatial location. A common choice for s are the longitude and latitude coordinates of the location where the measurements were taken. The Gaussian assumption, however, is not realistic when modelling extreme values, whose distribution is known to be skewed. Data would need to be transformed so that the underlying spatial random field becomes a Gaussian random field and geostatistics can be directly applied. The increasing interest in extreme values, especially in environmental applications, has led to the development of new statistical models specifically designed for spatial (and spatio-temporal) extremes.

Modelling extremes over space can be seen as a multivariate extreme value exercise. However, while for the univariate case the well known GEV distribution exists and is widely used (see Chapter 2, Section 2.5), "there is no simple parametric form for the multivariate extreme value limiting distribution" (Davison et al. (2012)). A number of summary measures to characterize the dependence structure are available instead. Let D be the number of spatial locations, and assume that the random variables $Y_1, \ldots, Y_D$ each follows a standard Fréchet distribution. The joint distribution can be expressed in terms of the **exponent measure** V (Davison et al. (2012); Fuentes et al. (2012); Wadsworth and Tawn (2012)), a function that characterizes the dependence among the D sites via the expression:

$$P(Y_1 \leq y_1, \ldots, Y_D \leq y_D) = \exp\{-V(y_1, \ldots, y_D)\}$$

If all the sites are independent, then $V(y_1, \ldots, y_D)$=1/y$_1 + \ldots + 1/y_D$, while in the case of total dependence, $V(y_1, \ldots, y_D) = max(1/y_1, \ldots, 1/y_D)$. The **extremal coefficient** θ_D (Davison et al. (2012); Fuentes et al. (2012); Wadsworth and Tawn (2012)) provides a numerical measure of dependence. It can be obtained by evaluating the exponent measure at 1, θ_D=V(1,...,1), ranging from 1 (totally dependent) to D (totally independent).

For a given threshold $u \in \Re$:

$$P(max(Y(s_1), \ldots, Y(s_D)) < u) = (P(Y(s_1) < u))^{\theta_D}$$

The extremal coefficient θ_D is independent of the threshold u (Fuentes et al. (2012)) and can be interpreted as the number of independent variables involved in a D-variate distribution. However, θ_D only provides information on the degree of dependence but does not allow the joint distribution to be modelled (Davison et al. (2012)). A number of approaches that directly address modelling the joint extreme value distribution have been proposed. A good review of these models can be found in Davison et al. (2012) and Davison and Gholamrezaee (2012). Amongst these models, latent variables models, copulas and max-stable processes are first briefly introduced. For further details see Davison et al. (2012). Then the method proposed by Keef et al. (2009) for characterizing the spatial dependence of extremes is discussed in detail and applied to two sets of river flow series in Scotland. Finally, spatial quantile regression is introduced and a model for extreme river flow data in Scotland is proposed.

5.1.1 Latent variables, copula models and max-stable processes

5.1.1.1 Latent variables

One way of modelling spatial extremes is by introducing a latent spatial process $\{Z(s)\}$, usually assumed to be a Gaussian stationary process. In such case, the response variables $Y(s) = \{Y(s_1),\ldots,Y(s_D)\}$ are assumed to be conditionally independent given $\{Z(s)\}$. The latent process is introduced via the parameters of the extreme value (GEV or GP) distribution (μ, σ, ξ). Inference is performed in a Bayesian hierarchical setting (Davison et al. (2012)). The simplest case is to model each of the parameters (μ, σ, ξ) independently in terms of latent spatial processes that characterize and drive spatial dependence in extremes. For example, $\mu(s) = \beta_\mu(s) + Z_\mu(s)$, i.e. the parameter μ is expressed as a spatial trend plus a Gaussian stationary process $\{Z_\mu(s)\}$ with a known covariance function, whose parameters need to be estimated along with β_μ. Examples of environmental applications include modelling extreme daily temperature (Fuentes et al. (2012)) and precipitation (Cooley et al. (2007); Davison et al. (2012)).

5.1.1.2 Copula models

The dependence structure amongst spatial extremes can be expressed using a copula function (Fuentes et al. (2012)). A copula C is a multivariate distribution on marginally uniform random variables. Its related to the joint distribution via the expression:

$$F(y_1, \ldots, y_D) = C(F_1(y_1), \ldots, F_D(y_D))$$

where F_1,...,F_D are univariate marginal distributions (Davison et al. (2012)). A common choice is the Gaussian copula, defined as:

$$C(u_1, \ldots, u_D) = \Phi(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_D), \Omega)$$

where Φ is the cdf of a standard normal random variable, $\Phi(, \Omega)$ is the joint Gaussian distribution with covariance matrix Ω and $U_i = \Phi(Y_i)$. In this way the correlation structure of extreme values can be established by specifying the covariance function of the latent Gaussian spatial process. However, the Gaussian copula is not flexible enough for characterizing complex dependence in the tails (Fuentes et al. (2012)). An extreme value copula is the copula function of the limiting extreme value distribution G (see 2) in the multivariate case. Asymptotically, the spatial Gaussian copula can be approximated by an independent extreme value copula, for which the extremal coefficient is always equal to the number of locations, resulting in extreme values being independent (irrespective of their correlation structure), which is not realistic in many cases.

5.1.1.3 Max-stable processes

A third alternative is to use max-stable processes. The condition of max-stability holds in the univariate case, in which if the limiting distribution G exists, it must be max-stable (Davison et al. (2012)), meaning that it satisfies the equation:

$$G^D(b_D + a_D y) = G(y)$$

for sequences $\{a_D\} > 0$, $\{b_D\}$. Max-stability can be thought as a property of the GEV distribution that allows extrapolation beyond the range of the data. The idea is to extend the GEV distribution to the spatial case, so that the joint behaviour of maxima

over a spatial region can be studied. Assume the marginal variables $Y = (Y_1, \ldots, Y_D)$ to be mutually independent standard Fréchet distributed, max-stability means that $DY \stackrel{d}{=} \max(Y_1, \ldots, Y_D)$. Consider a collection of iid random fields $\mathrm{Y}_i(\mathrm{s})$, where $\mathrm{s} \in R^D$ and i varies with time. We can define a max-stable process $\{Z(s)\}$ as the limit process of the maxima of the $\mathrm{Y}_i(\mathrm{s})$'s. $\{Z(s)\}$ is assumed to be stationary. Hence we can think of a max-stable process as the limiting distribution of the pointwise maxima of independent copies of a process (Davison et al. (2012); Davison and Gholamrezaee (2012)). Max-stable processes are best characterized via their spectral representation, for which models have been proposed. In particular, the Smith (Smith (1990)), Schlather (Schlather (2002)) and Brown-Resnick (Brown and Resnick (1977)) models are widely used. The representations that these models offer are based on points of a Poisson process and random fields and were motivated by physical phenomena (Davison et al. (2012); Davison and Gholamrezaee (2012)). These models however usually have slow rates of convergence, the limiting dependence structure is restrictive and in some cases they do not produce realistic realizations (Wadsworth and Tawn (2012)). Recently Wadsworth and Tawn (2012) developed a more general class of dependence models to overcome these difficulties by superposing two processes. For max-stable models, writing down the full likelihood function is not possible. Instead, a composite marginal likelihood, built using the pairwise marginal distributions, is used for parameter estimation (Davison et al. (2012); Davison and Gholamrezaee (2012); Wadsworth and Tawn (2012)). A modified version of the AIC is available for model comparison. Examples of environmental applications can be found in Davison et al. (2012); Davison and Gholamrezaee (2012), who use a max-stable representation to model annual maximum temperature and in Wadsworth and Tawn (2012), who model wave height data.

An alternative method for analyzing spatial extremes in terms of conditioning was proposed by Heffernan and Tawn (2004) and Keef et al. (2009). This method is described in the next section, followed by an application to river flow data.

5.2 Conditional spatial dependence of extreme values

Keef et al. (2009) propose a method to investigate the spatial dependence of high flows, based on previous work by Heffernan and Tawn (2004), that diverges from 'traditional'

spatial analysis, in the sense that no correlation function or metric to measure distance is required. The main results and notation for this section have been drawn from Keef et al. (2009) and Heffernan and Tawn (2004).

Define $\triangle = \{1, \ldots, d\}$ as the set of gauged sites (rivers) under study, $X = (X_1, \ldots, X_d)$ and X_i as the data record at each gauging station $i = 1, \ldots, d$, with cdf F_{X_i}. Keef et al. (2009) propose two conditional summaries of the dependence structure amongst high river flows; first, the joint probability of a subset of rivers $C \subseteq \triangle$ having flows over a certain threshold v_p, and secondly, the expected number of sites within $C \subseteq \triangle$ that are likely to exceed that threshold v_p. Both measures are defined conditioned on one of the rivers in C exceeding the threshold v_p. Prior to modelling the dependence structure, data at each gauging station are standardized so that they all follow the same marginal distribution. This is done to account for differences in catchment properties such as catchment size, directly related to the average river flow value, and rainfall. Originally, Heffernan and Tawn (2004) suggested a semi-parametric model for the marginal distributions based on the generalized Pareto distribution (see Chapter 2):

$$\hat{F}_{X_i}(x) = \begin{cases} 1 - \{1 - \tilde{F}_{X_i}(u_{X_i})\}\{1 + \xi_i x/\sigma_i\}_+^{-1/\xi_i} & x > u_{X_i} \\ \tilde{F}_{X_i}(x) & x \leq u_{X_i} \end{cases}$$

where $\psi_i = (\sigma_i, \xi_i)$ are the parameters of a GPD for the exceedances over a threshold u_{X_i} and $\tilde{F}_{X_i}$ is the empirical cumulative distribution. The estimated distribution $\hat{F}_{X_i}$ is then used to transform each variable X_i, $i = 1, \ldots, d$, to marginally follow a standard Gumbel distribution (ie μ=0,σ=1,ξ=0) as follows:

$$Y_i = -\log\{-\log\{\hat{F}_{X_i}(X_i)\}\} = t_i(X_i, \psi_i, \tilde{F}_{X_i}) = t_i(X_i)$$

This has the 'burden' of having to choose a suitable threshold u_{X_i}, which is difficult because catchments and rainfall are different, so it is not possible to use the same threshold always and know that results will be consistent. Instead of using a GPD, Keef et al. (2009) propose a simpler transformation based on the approximation of the empirical cdf:

$$Y_i = -\log\left[-\log\left\{\frac{rank(X_i)}{n+1}\right\}\right] \tag{5.1}$$

Now all the individual river flow records Y_i follow a Gumbel standard distribution, whose upper tail distribution can be approximated by an exponential distribution:

$$\begin{aligned} F(y) = P(Y_i \leq y) &= \exp\{-\exp(-y)\} \qquad -\infty < y < \infty \\ P(Y_i > y) &= 1 - P(Y_i \leq y) \sim \exp(-y) \qquad \text{as } y \to \infty \end{aligned}$$

Denote by $Y_{i,t}$ the conditioning variable, and the remaining d-1 variables by $\mathrm{Y}_{-i,t} = (Y_{1,t}, \ldots, \widehat{Y_{i,t}}, \ldots, Y_{d,t})$, where $\widehat{Y_{i,t}}$ means that the variable $Y_{i,t}$ is excluded from the vector. Let $1-p$ be the probability of river flow exceeding v_p at each site and v_p the corresponding p^{th} quantile from the Gumbel distribution. $(\mathrm{Y}_{i,t}, \mathrm{Y}_{-i,t})$, $t = 1, \ldots, n$ is assumed to be a "segment from a multivariate stationary process with temporal dependence from day to day and independence at long time lags" (Keef et al. (2009)). The first spatial risk measure they propose is:

$$P_C(p) = P\{min_{j \in C} max_{\tau \in A_j}(Y_{j,t+\tau}) > v_p | Y_{i,t} > v_p\} \tag{5.2}$$

i.e. the probability that all rivers in area C exceed v_p at some time point t+τ, $\tau \in A_j$, given that $\mathrm{Y}_{i,t} > v_p$ at time t. $A_j = \{-L_j, \ldots, L^j\}$ $(L_j, L^j \geq 0)$ is a set of time lags between events happening in Y_i and Y_{-i}. The simplest case would be estimate $P_C(p)$ based on only two variables (Y_i, Y_j) and no time lag τ. In that case, the probability of river flow at sites Y_j and Y_i exceeding v_p at the same time can be estimated as:

$$S_j(p) = P(Y_j > v_p | Y_i > v_p)$$

The second spatial risk measure Keef et al. (2009) propose is the mean number of sites in $C \subseteq \triangle$ that exceed v_p, given that $Y_{i,t} > v_p$ at time t, calculated as:

$$\begin{aligned} &N(p) = E[\#\{j \in C / max_{\tau \in A_j}(Y_{j,t+\tau}) > v_p\} | Y_{i,t} > v_p] = \\ &\textstyle\sum_{j=1}^{|C|} P\{max_{\tau \in A_j}(Y_{j,t+\tau}) > v_p | Y_{i,t} > v_p\} \end{aligned} \tag{5.3}$$

These two spatial risk measures, when calculated empirically, do not produce reliable estimates for large values of p (p >0.995), which are those of greater interest for planning purposes, as they are based on a limited number of observations. Hence, a model that allows reliable extrapolation is needed. Keef et al. (2009) treat the problem as a

multivariate extremes problem, following the model proposed by Heffernan and Tawn (2004), where an asymptotically independent model (that also allows for dependence) is used. The idea is to model the distribution of $\mathrm{Y}_{-i,t}|Y_{i,t} = y$ when y is large. The model is fitted using semi-parametric regression, to then simulate from the joint distribution of $(\mathrm{Y}_{i,t}, Y_{-i,t})|Y_{i,t} > v_p$ for large values of p. While Heffernan and Tawn (2004) assume that the vector $(\mathrm{Y}_{i,t}, Y_{-i,t})$ is time independent, Keef et al. (2009) extend the model to allow for temporal dependence. Firstly the original model proposed by Heffernan and Tawn (2004) is presented, where temporal dependence is not considered and hence the time index t can be disregarded.

The aim is to model the distribution of $Y_{-i}|Y_i = y$ for y large. For that purpose, univariate extreme value theory can be extended to a multivariate context. Assume that there exist 'normalizing functions' (which are not unique) $a_{|i}(x)$, $b_{|i}(x)$: $\Re \to \Re^{d-1}$ / $\forall$ fixed $z \in R^{d-1}$ and any sequence of y_i values such that $y_i \to \infty$ (i.e. high enough):

$$lim_{y_i \to \infty}[Y_{-i} \leq a_{|i}(y_i) + b_{|i}(y_i) z_{|i} | Y_i = y_i] = G_{|i}(z_{|i}) \tag{5.4}$$

Denote by G_i the i^{th} marginal distribution of$G_{|i}$, a non-degenerate distribution function $\forall i \in \triangle$ and $lim_{z \to \infty}\{G_i(z)\}$=1 $\forall i$. The method assumes that the limiting distribution holds $\forall$ $y_i > u_{Y_i}$ for a suitable high threshold u_{Y_i}. When $Y_i = y_i$, with $y_i > u_{Y_i}$, the (standardized) random variable $Z_{|i}$ is defined as:

$$Z_{|i} = \frac{Y_{-i} - a_{|i}(y_i)}{b_{|i}(y_i)} \tag{5.5}$$

and the limiting distribution of $Z_{|i}$:

$$lim_{y_i \to \infty} P(Z_{|i} \leq z_{|i} | Y_i = y_i) = G_{|i}(z_{|i}) \tag{5.6}$$

Under this assumption, conditionally on $Y_i > u_{Y_i}$, as $u_{Y_i} \to \infty$, the variables $Y_i - u_{Y_i} (> 0)$ and $Z_{|i}$ are independent in the limit and their limiting marginal distributions are exponential and $G_{|i}(z_{|i})$ respectively (Keef et al. (2009)).

For each $j \in \triangle, j \neq i$ consider $G_{j|i}(z_{j|i})$, the limiting marginal distribution function of:

$$Z_{j|i} = \frac{Y_j - a_{j|i}(y_i)}{b_{j|i}(y_i)} | Y_i = y_i \text{ given } Y_i = y_i \text{ as } y_i \to \infty$$

If $G_{|i}(z_{|i}) = \prod_{j \neq i} G_{j|i}(z_{j|i})$ then we say that the elements of Y_{-i} are mutually asymptotically conditionally independent given Y_i.

The extremal dependence behaviour is characterized by $a_{|i}(y)$, $b_{|i}(y)$ and $G_{|i}$, hence estimates of the three are needed in order to derive the conditional distribution. To do so, Heffernan and Tawn (2004) propose a semi-parametric model. The parametric part involves estimating $a_{|i}(y)$ and $b_{|i}(y)$ using the (parametric) regression model:

$$Y_{-i} = a_{|i}(y) + b_{|i}(y) Z_{|i} = a_{|i} y + y^{b_{|i}} Z_{|i} \tag{5.7}$$

Specifically, $a_{|i}(y)$ and $b_{|i}(y)$ are expressed in terms of y as $a_{|i}(y) = a_{|i} y$ and $b_{|i}(y) = y^{b_{|i}}$, with the restrictions $0 \leq a_{|i} \leq 1$ and $-\infty < b_{|i} < 1$. The set of parameters $\theta_{|i} = (a_{|i}, b_{|i})$ can be estimated parametrically; assume that $Z_{|i}$ has mean $\mu_{|i}$, standard deviation $\sigma_{|i}$, then $Y_{-i}|Y_i = y$ has mean $\mu_{|i}(y) = a_{|i}(y) + \mu_{|i} b_{|i}(y)$ and standard deviation $\sigma_{|i}(y) = \sigma_{|i} b_{|i}(y)$. The parameters to estimate are now $(\theta_{|i}, \lambda_{|i}) = (a_{|i}, b_{|i}, \mu_{|i}, \sigma_{|i})$, which are parameters of a "multivariate regression model with non-constant variance and unspecified error distribution" (Heffernan and Tawn (2004)). Only $\theta_{|i}$ is of interest, hence $\lambda_{|i}$ can be considered as nuisance parameters when maximizing the function:

$$Q_{|i}(\theta_{|i}, \lambda_{|i}) = -\sum_{j \neq i} \sum_{k=1}^{n_{u_{Y_i}}} \left[log\{\sigma_{j|i}(y_{i|i,k})\} + \frac{1}{2} \left\{ \frac{y_{j|i,k} - \mu_{j|i}(y_{i|i,k})}{\sigma_{j|i}(y_{i|i,k})} \right\}^2 \right] \tag{5.8}$$

where $n_{u_{Y_i}}$ is the number of observations for which $Y_i > u_{Y_i}$. Note that this is actually a POT approach as the model is fitted using solely the observations above the chosen threshold. Maximizing equation (5.8) involves estimating $4(d-1)$ parameters for each i. The parameters $a_{|i}$ and $b_{|i}$ provide information on the relationship between Y_i and Y_j; $a_{j|i}$ describes the overall strength of the dependence, with high values indicating strong dependence and $b_{j|i}$ describes how the dependence changes with increasing values of Y_i, with positive values indicating an increase in the variance of $Y_j|Y_i = y$ (as y increases)

(Keef et al. (2009)).

The non-parametric part involves estimating $G_{|i}$, the joint distribution of $Z_{|i}$. Assuming that $\hat{\theta}_{|i}$ has been estimated, $G_{|i}$ can be estimated non-parametrically by using the empirical (or kernel smoothed) distribution of replicates of the random variable $\hat{Z}_{|i}$:

$$\hat{Z}_{|i} = \frac{Y_{-i} - \hat{a}_{|i}(y_i)}{\hat{b}_{|i}(y_i)}$$

for $Y_i = y_i > u_{Y_i}$.

Based on the dependence model $Y_{-i} = a_{|i}(y_i) + b_{|i}(y_i)Z_{|i}$ (for $Y_i = y_i > u_{Y_i}$), 3 classes of dependence can be identified:

a) $a_{j|i}$=1, $b_{j|i}$=0: (Y_i, Y_j) asymptotically dependent; i.e. the quantiles grow at the same rate

b) $0 < a_{j|i} < 1$ or $b_{j|i} > 0$: (Y_i, Y_j) positive extremal dependence

c) $a_{j|i}$=0, $b_{j|i} \leq 0$: (Y_i, Y_j) extremal near dependence

5.2.1 Incorporating temporal dependence

The conditional distribution that results from incorporating temporal dependence is:

$$\{Y_{j,t+\tau}/j \in \triangle, \tau \in A_j\}|Y_{i,t} = y_t$$

for $\mathrm{y}_t > u_{Y_i}$, for large u_{Y_i}, and $\mathrm{A}_j = \{-L_j, \ldots, L^j\}$ (Keef et al. (2009)).

The difference is that now for each component of Y_{-it} $(\{\mathrm{Y}_{1,t}, \ldots, \widehat{Y_{i,t}}, \ldots, Y_{d,t}\})$ we consider river flow values not just at time t but also at lagged times $t + \tau$, $\tau \in A_j$. The vector $\{Y_{j,t+\tau}/j \in \triangle, \tau \in A_j\}$ has Gumbel marginals and is an extension of $Y_{-i,t}$ (Keef et al. (2009)), so the method of Heffernan and Tawn (2004) can be directly applied (given that there is no missing data) at each time lag. It works as follows:

1. Estimate the distribution of $Y_{j,t+\tau}|Y_{i,t} = y_{i,t}$ separately for each time lag $\tau \in A_j$, $j \in \triangle$. Denote the parameters of these marginal components as $a_{|i}^{(\tau)}$,$b_{|i}^{(\tau)}$.

2. Estimate the distribution of $Z_{|i}^{(\tau)}$. At time t, $y_{i,t} > u_{Y_i}$ and observed value $y_{j,t+\tau}$, the corresponding standardized variable:

$$z_{j,t|i}^{(\tau)} = \frac{y_{t,t+\tau} - \hat{a}_{j|i}^{(\tau)} y_t}{y_t^{\hat{b}_{j|i}^{(\tau)}}}$$

   It is possible to calculate the standardized variable $Z_{|i}^{(\tau)}$ even in the presence of missing data. See Keef et al. (2009) for details on how to deal with this.

3. As in the simpler case (with no time lag), samples can be generated from the conditional distribution $\{Y_{j,t+\tau}/j \in \triangle, \tau \in A_j\}|Y_{i,t} = y_{i,t}$ at each $\tau \in A_j$ and the corresponding $P_C(p)$ and $N(p)$ can be estimated following expressions 5.2 and 5.3 respectively.

5.2.2 Uncertainty

The proposed model has three sources of uncertainty:

1. Estimation of the semiparametric marginal models (i.e. the models that allow the transformation into Gumbel distribution). This can be ignored since ranks are used (following equation (5.1)) to derive the marginal empirical distribution.

2. Parametric normalization functions $a_{|i}()$, $b_{|i}()$

3. Non-parametric model for $G_{|i}$

Standard errors of the estimates of $P_C(p)$ and $N(p)$ can be calculated using standard semiparametric block bootstrap methods (Davison and Hinkley (1997)). This is done in four steps:

1. Obtain a non-parametric block bootstrap sample (with replacement) from the data (marginally transformed to follow a Gumbel distribution), where blocks are defined as hydrologic years.

2. For each bootstrap sample $k = 1, \ldots, K$, fit a dependence model model, i.e. estimate $a_{|i}^k$, $b_{|i}^k$ and $G_{|i}^k$.

3. For each fitted model, generate pseudo-observations and calculated estimated values for $P_c(p)^k$ and $N(p)^k$, $k = 1, \ldots, K$.

4. Take the 2.5^{th} and 97.5^{th} ordered $P_c(p)$ and $N(p)$ values to obtain a 95% CI.

5.2.3 The model in practice

The method can be applied to a set of river flow data records by the following steps:

1. Fit a marginal model (GPD) for each X_i, $i = 1, \ldots, d$; i.e. estimate the parameters $\psi_i = (\sigma_i, \xi_i)$ for each river flow series individually. Note that to fit the GPD a threshold u_{X_i} (not necessarily the same for all rivers) must be chosen. Alternatively, use the empirical cdf $\tilde{F}_{X_i}$. The latter is adopted here.

2. Transform the data into Gumbel margins following:

$$Y_i = -\log[-\log\{\hat{F}_{X_i}(X_i)\}]$$

3. Model the dependence structure:

   a) Select a suitable common threshold $u_{Y_i} = u \forall i$ to fit the dependence model. To decide on the threshold value, a series of diagnostic plots can be used. A good idea is to perform a sensitivity analysis; i.e. fit the model for a range of threshold values, generate estimates for $P_C(p)$ and $N(p)$ and assess the stability of these estimates. Also, a plot of $Z_{|i}$ versus Y_i, is useful, to check the assumption of these two being independent for $Y_i > u_{Y_i}$, for each i. Ideally, we want a threshold as low as possible; high enough so that independence is reached, but small enough so that there is enough data above the threshold to reliably fit the model (Keef et al. (2009)).

   b) Once the threshold u has been selected, fit the corresponding conditional models:

$$\begin{aligned} Y_{-i} &= a_{|i}(y) + b_{|i}(y) Z_{|i}, \quad Y_i = y > u \\ a_{|i}(y) &= a_{|i} y \qquad\qquad\qquad 0 \leq a_{|i} \leq 1 \\ b_{|i}(y) &= y^{b_{|i}} \qquad\qquad\qquad -\infty \leq b_{|i} \leq 1 \end{aligned}$$

c) Consider possible simplifications of the dependence structure, like pairwise exchangeability (i.e. $(\hat{a}_{j|i}, \hat{b}_{j|i}) = (\hat{a}_{i|j}, \hat{b}_{i|j})$) and independence of the components of $Z_{|i}$. Pairwise exchangeability would mean that the influence of river j on river i is the same as the influence of river i on river j, while independence of the components of $Z_{|i}$ would mean that the limiting distribution $G_{|i}$ could be factorized as the product of the marginal distributions, avoiding the estimation of the multivariate distribution $G_{|i}$.

4. Generate pseudo-samples

   (a) Pick a threshold of interest $v_p (> u_{Y_i})$ and simulate Y_i from a Gumbel distribution conditional on its exceeding v_p:

$$F_u^G(x) = 1 - \exp\{1 - \exp\left(\tfrac{x-u}{\sigma}\right)\} \quad x > u$$

   where σ is the scale parameter and u is the threshold we condition on. Alternatively, use the exponential distribution shifted by u:

$$f_u(x) = \exp(-(x-u)) \quad x > u$$

   Here threshold v_p was chosen corresponding to p=0.95. This means that the dependence model will be fitted for extreme values in the conditioning river corresponding to probabilities of exceedance of 0.05 and lower.

   (b) Sample $Z_{|i}$ from $\hat{G}_{|i}$ independently of Y_i

   (c) Calculate $Y_{-i} = \hat{a}_{|i}(Y_i) + \hat{b}_{|i}(Y_i) Z_{|i}$

   (d) (optional) Transform $Y = (Y_{-i}, Y_i)$ back to the original scale. Note that this step is only possible if the marginal model has been fitted using a GPD

5. Evaluate uncertainty by calculating 95% CIs using block bootstrapping and model check. In particular, check that the assumption of independence between the standardized variables $Z_{|i}$ and the conditioning variable holds and that empirical and model based estimates agree.

5.2.4 A more general framework: Estimating functionals of the joint tails of $\mathbf{X}=(X_1, \ldots, X_d)$

The method described in Keef et al. (2009) is an adaption of a more general method proposed by Heffernan and Tawn (2004) designed to study the probability $P(X \in C)$, where C is an extreme set, so that $\forall x \in C$, at least one component of $x = (x_1, \ldots, x_d)$ is extreme; i.e. $x_i > u$ for at least one i, for u large enough (and hence at least $x_{max} > u$). The extreme set C can be expressed as the union of d subsets $C = \bigcup_{i=1}^{d} C_i$, where C_i is "the part of C for which X_i is the largest component of X" (Heffernan and Tawn (2004)). C is said to be an extreme set if all x_i values in a non-empty C_i fall in the upper tail of F_{X_i}; i.e. if the random variable V_{X_i} is defined as $V_{X_i} = inf_{x \in C_i}(x_i)$ (i.e. the smallest component of x), then $F_{X_i}(v_{X_i})$ is close to 1. Note that the maximum value that F_{X_i} takes is 1, so $F_{X_i}(v_{X_i})$ being close to 1 means that v_{X_i} is very high up in the upper tail of the distribution. The probability of X being in the extreme set C can be re-written as:

$$P(X \in C) = \sum_{i=1}^{d} P(X \in C_i) = \sum_{i=1}^{d} P(X \in C_i | X_i > v_{X_i}) P(X_i > v_{X_i})$$

In order to calculate $P(X \in C)$, estimates are needed for both $P(X_i > v_{X_i})$, for which a marginal extreme value model is needed, and $P(X \in C_i | X_i > v_{X_i})$, for which an extreme value model for the dependence structure needs to be defined. The marginal extreme value model is just a GPD (by construction). The dependence structure is modelled as described previously in Section 5.2.3.

5.2.4.1 Return levels

A return level for an extreme event with probability p is defined as the value v_p that satisfies:

$$P(Y \in C(v_p)) = p$$

where:

$$C^m(v_p) = \{y \in \Re^m / \sum_{i \in M} y_i > v_p\}$$

$m = 2, \ldots, d$; i.e. $C^m(v_p)$ is formed of those observations y such that the sum of their components $(y_1, \ldots, y_d)$ is over the threshold v_p. Note that now subvectors of size $m =$

$2, \ldots, d$ of the transformed variable Y are being considered, indexed by $M \subseteq \{1, \ldots, d\}$. The idea is to produce a plot of v_p versus p for high values of p. Instead of working with the individual variables, linear combinations of them (in the Gumbel scale) are used. The probability $P(Y \in C(v_p))$ depends on both the conditional dependence model and the marginal model, as it can be expressed as:

$$P(Y \in C(v_p)) = \sum_{i=1}^{d} P(Y \in C(v_p)|y_i > v_{Y_i})P(Y_i > v_{Y_i})$$

where Y_i, as before, represents the chosen conditioning variable. Hence, $P(Y \in C(v_p))$ can be estimated using the fitted conditional model; $P(Y \in C(v_p))|y_i > v_{Y_i})$ is the long run proportion of the simulated samples that fall in $C(v_p)$. Estimation of $P(Y_i > v_{Y_i})$ is straightforward as Y_i follows a standard Gumbel distribution (by construction).

5.3 Case Study 1: Northern Scotland and Glasgow Area

The dependence structure amongst two sets of Scottish rivers was investigated by means of the two spatial risk measures $P_C(p)$ and $N(p)$ proposed by Keef et al. (2009). The first set comprises the rivers Dulnain, Lossie, Ewe and Ness (Figure 5.1(a)) in the Northern part of Scotland (see Chapter 2 for details on the rivers Lossie, Ewe and Ness), while the second one is centered around Glasgow. The Glasgow area is of special interest, being the largest city of Scotland and the one where most of the population is concentrated. Rivers Clyde (see Chapter 2), Kelvin, Irvine and Glazert were investigated here. Their spatial location is shown on Figure 5.1(b).

Each data record was individually transformed according to expression (5.1) to follow a standard Gumbel distribution. A range of thresholds (corresponding to probabilities 0.75-0.90) $u_{Y_i} = u$ was considered for fitting the dependence model. After a sensitivity analysis, where the stability of extrapolations (Figure 5.2) as well as the assumption of independence between the conditioning variable Y_i and the standardized variables $Z_{|i}$ (Figures 5.3 and 5.4) were checked, a threshold $u_{Y_i} = u$ such that $P(Y_i < u)$=0.8 was found to be adequate. This corresponds to u=1.50 $\log(m^3/s)$ (in the standard Gumbel scale). The dependence model was then fitted using the `optim` function in R,

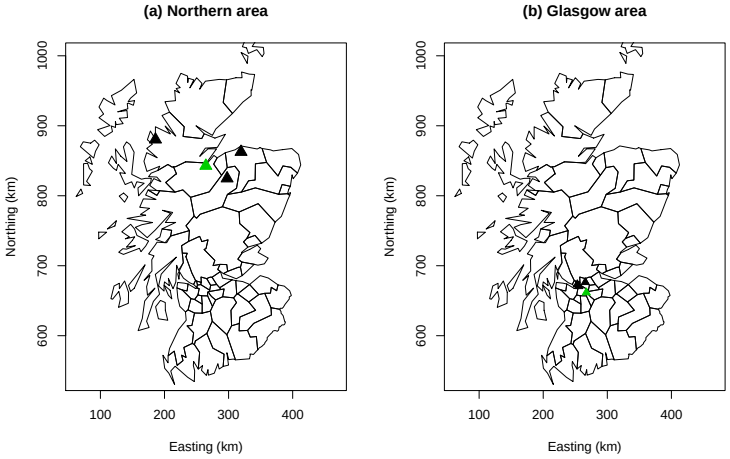

FIGURE 5.1: Location of gauging stations, (a) Northern area and (b) Glasgow area. The green one indicates the conditioning river ((a) Ness and (b) Clyde)

conditioning on each of the four rivers in turn. The resulting estimates $a_{|i}$ and $b_{|i}$ are summarized in Tables 5.1 and 5.2 for the North and Glasgow areas respectively.

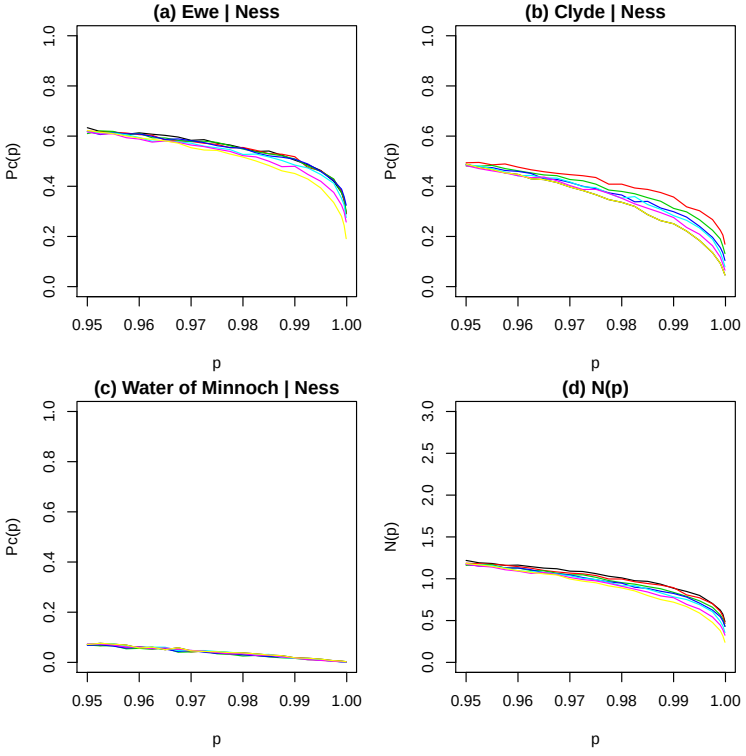

FIGURE 5.2: Estimated values for $P_c(p)$ and $N(p)$ for thresholds 0.75-0.90 (Northern Area)

All pairs (apart from Ewe|Lossie in the Northern set, for which there is extremal near

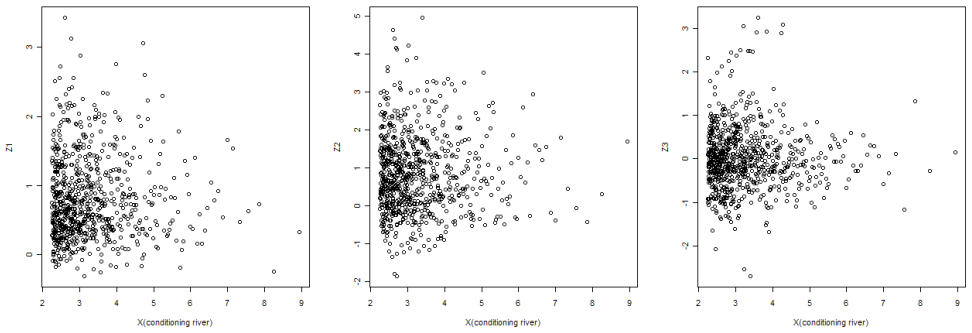

FIGURE 5.3: Standardized variables vs conditioning variable - Northern area

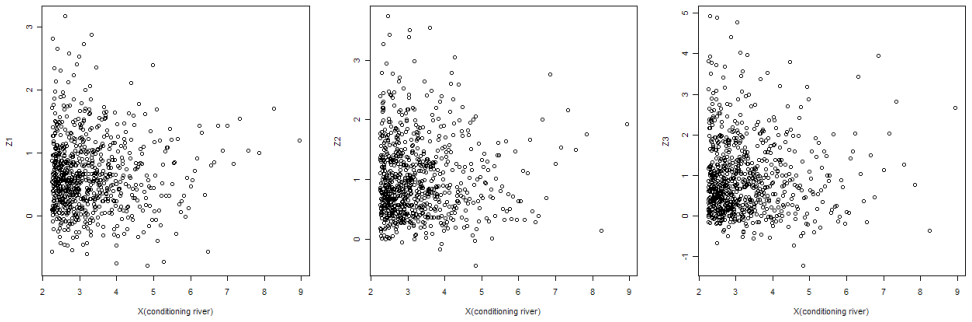

FIGURE 5.4: Standardized variables vs conditioning variable - Glasgow area

Rivers Conditioning on	Ness	Dulnain	Lossie	Ewe
Ness		$\hat{a}$=0.275 $\hat{b}$=0.591	$\hat{a}$=0.163 $\hat{b}$=0.143	$\hat{a}$=0.897 $\hat{b}$=0.181
Dulnain	$\hat{a}$=0.461 $\hat{b}$=0.382		$\hat{a}$=0.103 $\hat{b}$=0.442	$\hat{a}$=0.309 $\hat{b}$=0.543
Lossie	$\hat{a}$=0.046 $\hat{b}$=0.010	$\hat{a}$=0.559 $\hat{b}$=0.479		$\hat{a}$=0.000 $\hat{b}$=-0.035
Ewe	$\hat{a}$=0.808 $\hat{b}$=0.746	$\hat{a}$=0.252 $\hat{b}$=0.521	$\hat{a}$=0.104 $\hat{b}$=0.290	

TABLE 5.1: Estimated parameters $a_{|i}$ and $b_{|i}$ of the conditional probability model for a threshold u=1.50 (corresponding to probability 0.8) - Northern Area

independence) exhibit positive extremal dependence and pairwise exchangeability cannot be assumed.

Estimates of $P_C(p)$ and $N(p)$ are shown here only for the models conditioned on the

Rivers Conditioning on	Clyde	Kelvin	Irvine	Glazert
Clyde		$\hat{a}$=0.599	$\hat{a}$=0.340	$\hat{a}$=0.470
		$\hat{b}$=0.403	$\hat{b}$=0.454	$\hat{b}$=0.191
Kelvin	$\hat{a}$=0.413		$\hat{a}$=0.254	$\hat{a}$=0.538
	$\hat{b}$=0.619		$\hat{b}$=0.589	$\hat{b}$=0.650
Irvine	$\hat{a}$=0.438	$\hat{a}$=0.465		$\hat{a}$=0.540
	$\hat{b}$=0.394	$\hat{b}$=0.452		$\hat{b}$=0.504
Glazert	$\hat{a}$=0.428	$\hat{a}$=0.642	$\hat{a}$=0.625	
	$\hat{b}$=0.358	$\hat{b}$=0.476	$\hat{b}$=0.441	

TABLE 5.2: Estimated parameters $a_{|i}$ and $b_{|i}$ of the conditional probability model for a threshold u=1.50 (corresponding to probability 0.8) - Glasgow area

River Ness and the River Clyde respectively. These two were chosen to be the conditioning rivers within each set as they are the largest in each area, hence it seemed reasonable to investigate their influence on the rest of the rivers. A large catchment integrates effects of weather events of different types and scales, while small catchments are more likely to be influenced by local, smaller events. Before generating samples from the fitted model, the assumption of independence between the standardized variables $Z_{|\text{Ness}} = (Z_{1|\text{Ness}}, Z_{2|\text{Ness}}, Z_{3|\text{Ness}})$ and the conditioning river (Ness), and $Z_{|\text{Clyde}} = (Z_{1|\text{Clyde}}, Z_{2|\text{Clyde}}, Z_{3|\text{Clyde}})$ and the conditioning river (Clyde) was checked. Figures 5.3 and 5.4 suggest that independence is a reasonable assumption in both models as there is no apparent systematic pattern in either plot.

10000 pseudo-samples were obtained from the fitted model in each case in order to estimate $P_C(p)$ and $N(p)$. Confidence intervals were calculated using block bootstrap methods. The resulting estimates are shown on Figures 5.5 and 5.6. Figures 5.5(a), (b) and (c) show the conditional probabilities (y axis) of the rivers Dulnain, Lossie and Ewe being above a range of thresholds v_p associated with probabilities p (x axis), while Figures 5.6(a), (b) and (c) show the conditional probabilities of the rivers Kelvin, Irvine and Glazert being above a range of thresholds v_p associated with probabilities p. The blue lines represent the 95% confidence intervals calculated using block bootstrap (based on 100 samples). In all cases, as p increases (the flow event becomes more extreme) the conditional probability decreases. This is expected, as very extreme events tend to be more localized in space.

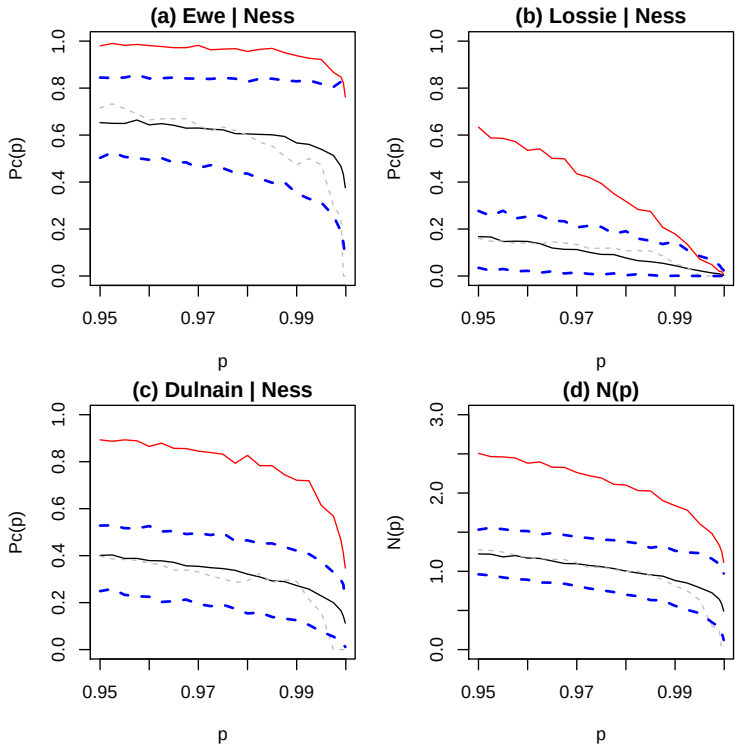

FIGURE 5.5: Estimated values for $P_C(p)$ (Figures (a), (b), (c)) and $N(p)$ (Figure (d)) for a range of probabilities of exceedance $1-p$ (Northern Area). In each of the plots, the solid black line represents the model estimates, the blue dashed lines represent 95% block bootstrap confidence intervals, the grey dashed line represents the empirical estimates and the red solid line corresponds to the model estimates when a lag of up to three days is included

Figure 5.5(d) shows the expected number of rivers (out of the 3) whose river flow would be over the corresponding threshold. For p >0.95, on average, one of the 3 rivers in the area (as well as the River Ness) is expected to be over the same threshold, probably the Ewe.

The Ewe, despite being the furthest away from the Ness amongst the three rivers considered, shows the strongest dependence, with relatively high (>0.4) conditional probabilities even for very extreme events. This can be explained by the fact that both are influenced by westerly events, related to increased precipitation, as described in Chapter 2. Also, they both belong to the same catchment cluster according to Acreman and Sinclair (1986). The Lossie, on the other hand, shows conditional probabilities which are fairly low, and it can be considered independent of the Ness for very extreme events, which might be explained by the two rivers being influenced by different weather patterns.

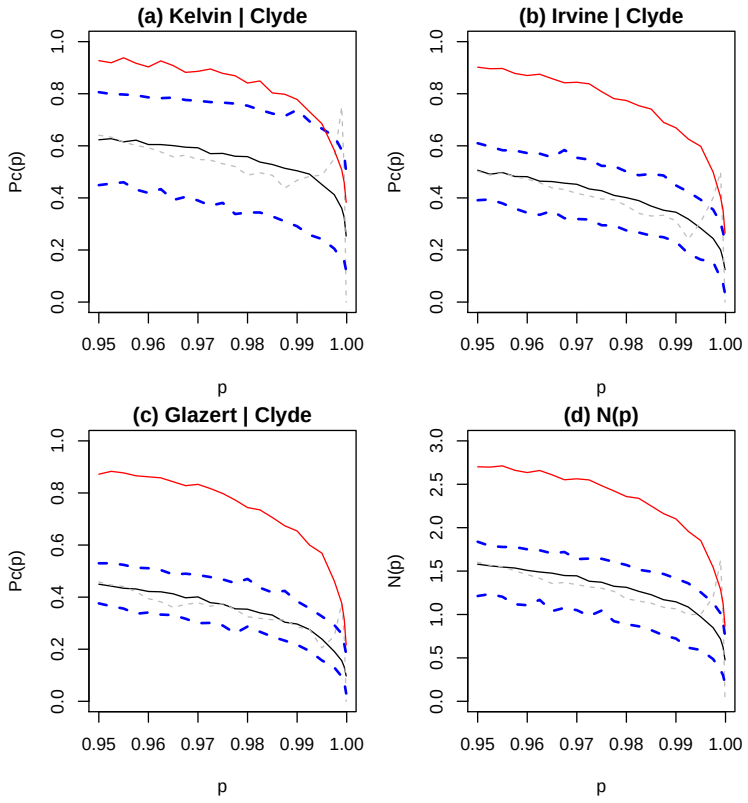

FIGURE 5.6: Estimated values for $P_C(p)$ (Figures (a), (b), (c)) and $N(p)$ (Figure (d)) for a range of probabilities of exceedance $1-p$ (Glasgow Area). In each of the plots, the solid black line represents the model estimates, the blue dashed lines represent 95% block bootstrap confidence intervals, the grey dashed line represents the empirical estimates and the red solid line corresponds to the model estimates when a lag of up to three days is included

No great differences are observed amongst the dependence on the River Clyde of the three rivers considered. The conditional probabilities are slightly higher for the River Kelvin, which is a tributary of the Clyde and hence a stronger relationship between the two is expected than for the rivers which are not directly connected to the Clyde.

An interesting extension of Keef et al. (2009) is that of investigating time lags. This allows for events lasting more than one day to be explored, e.g. if an extreme flow takes place in the river Clyde today, the model estimates the probability that river flow is also extreme in rivers Kelvin, Irvine and Glazert either today or a few days earlier/later. Lag ranging from -3 to 3 days were considered. This range of lags was chosen as dependence between extreme daily values more than three days apart is very unlikely. To derive

estimates for the conditional probabilities in this case, separate conditional models were fitted for each time lag τ as described in Section 5.2.1. The resulting estimates for $P_C(p)$ and $N(p)$ are shown on Figure 5.5 and Figure 5.6 (red solid line).

The lagged results (red solid lines in Figures 5.5 and 5.6) are similar qualitatively to the non lagged version, the main difference being that the estimated conditional probabilities are higher. This might be explained by the fact that adding the time lag offsets the catchment size effect to some extent; larger catchments (e.g. Clyde) take longer to respond than the smaller catchments, so the same event might be a day or so later in the Clyde. The incorporation of the time lag in the model may also allow for weather systems that move West to East or viceversa across the region. For the rivers Dulnain and Ewe the rate at which the corresponding conditional probabilities decrease is roughly the same as in the non lagged case, whereas for the River Lossie the situation is slightly different; here, for small values of p (0.95-0.98) the conditional probability is much higher than when no lags were considered, while for values of p close to 1 (i.e. very extreme events) the conditional probability is still very close to zero.

Figures 5.5(d) and 5.6(d) show the expected number of sites that will be over the corresponding threshold. In the lagged case, even for a very extreme event in the Clyde, we would still expect one other river to exceed that same extreme flow and two or three for most flows $p > 0.95$.

5.3.1 Conditional return level

Following Section 5.2.4.1, conditional return levels were estimated for a) Dulnain+ Lossie+ Ewe|Ness, b) Dulnain+ Lossie|Ness and c) Ewe|Ness in the Northern set, and (a) Kelvin+ Irvine+ Glazert|Clyde, (b)Irvine+ Glazert|Clyde and (c) Kelvin|Clyde in the Glasgow area. The resulting return levels are shown on Figure 5.7 and Figure 5.8 plotted against their exceedance probability. For a fixed return level, the exceedance probability should decrease as the event in the conditioning river becomes more extreme. For each of the three combinations in each set, return levels were calculated conditioning on the River Ness and Clyde exceeding a threshold corresponding to probabilities of 0.85,0.90,0.95,0.98 and 0.995. Each of these scenarios is represented with a different

colour on Figure 5.7 and Figure 5.8. As expected, as the return level increases, the exceedance probability decreases. As the event in the conditioning river becomes more extreme, the return level graph becomes steeper. This suggests that the relationship between a return level and its associated probability varies due to the strength of dependence on the conditioning river, hence highlighting the importance of taking the dependence structure (amongst rivers) into account when estimating flood risk. For low return levels, the exceedance probability decreases as the event in the conditioning river becomes more extreme; however, this 'pattern' changes as we move towards higher return levels, revealing the complicated dependence structure amongst extreme river flows.

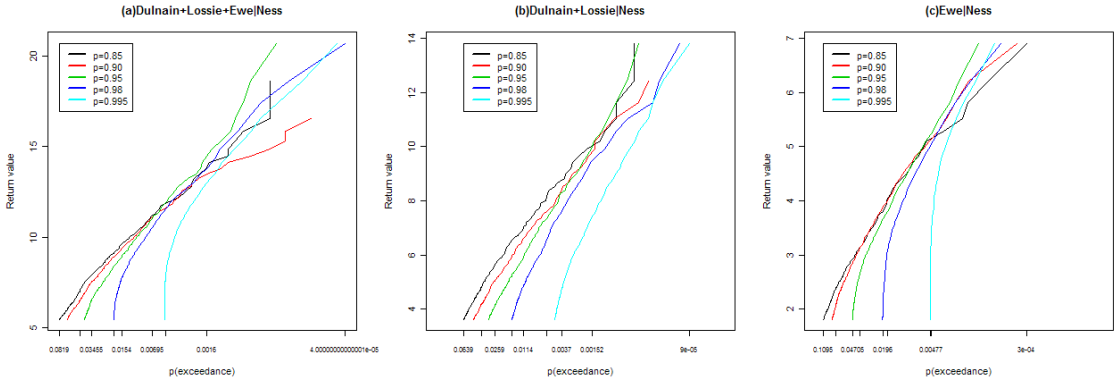

FIGURE 5.7: Conditional return values (standard Gumbel scale) - Northern area. Note that Figures (a) and (b) refer to return levels of more than one river. The different colored lines correspond to different threshold values

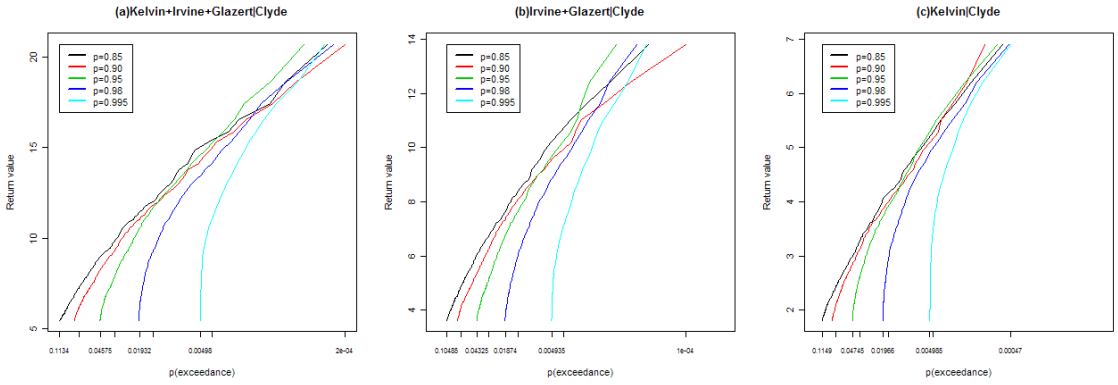

FIGURE 5.8: Conditional return values (standard Gumbel scale) - Glasgow area. Note that Figures (a) and (b) refer to return levels of more than one river. The different colored lines correspond to different threshold values

5.4 Case Study 2: Application on the selected eight rivers

To further investigate the degree of dependence between extreme river flow within the East and West of Scotland, the method was applied to the eight rivers selected in Chapter 2. A separate model was fitted for eastern and western rivers. The estimated conditional probabilities for the non-lagged and lagged version (with a time lag of (-3,3) days) are shown in Figure 5.9 for rivers in the East and Figure 5.10 for rivers in the West. Rivers Tweed and Ness were chosen as the conditioning rivers respectively. As expected, based on the results obtained in the previous chapter, the dependence between the Lossie and the Tweed is lower that between the remaining eastern rivers, although once the time lag is incorporated in the model, the conditional probabilities for the Lossie increase considerably. In this case, the inclusion of the time lag may allow for weather systems that move North to South or viceversa across the region. Confidence intervals are wider for the River Tay. On the other hand, there is strong dependence between the rivers Ewe and Ness, while the Water of Minnoch could be considered nearly independent of the Ness. As before, once the time lag is included in the model, the conditional probabilities increase.

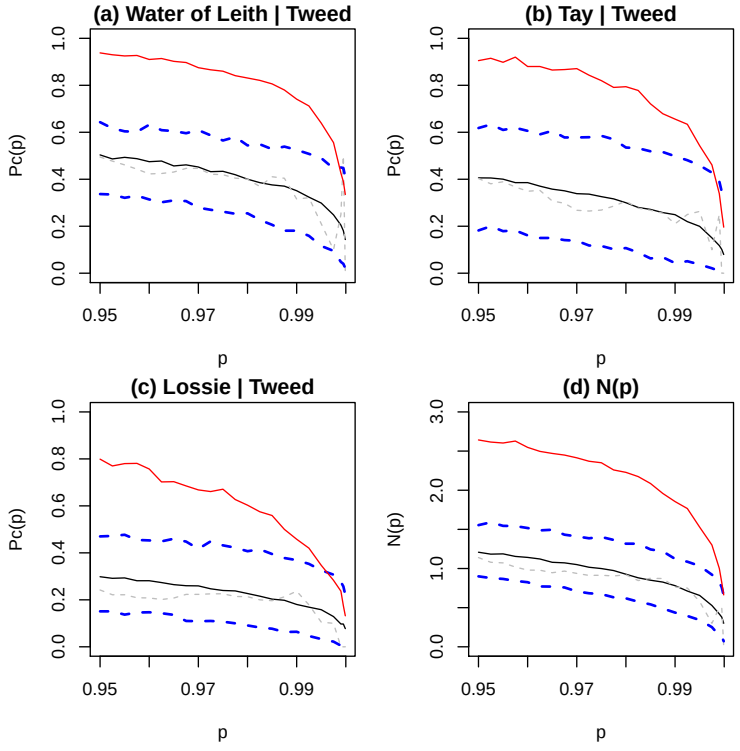

FIGURE 5.9: Estimated values for $P_C(p)$ and $N(p)$ (lagged version) - Eastern rivers

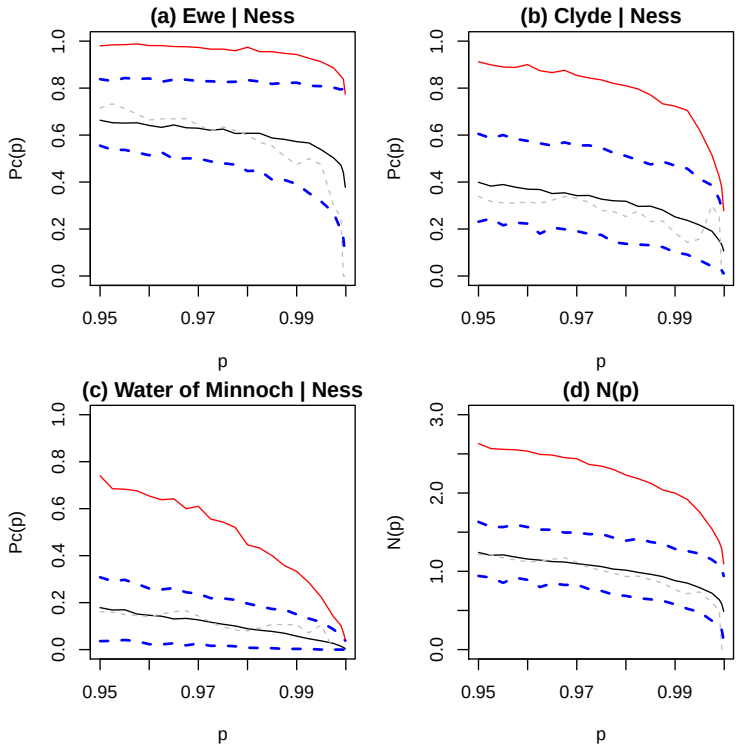

FIGURE 5.10: Estimated values for $P_C(p)$ and $N(p)$ (lagged version) - Western rivers

The model proposed by Keef et al. (2009) and described in Section 5.2 fits the data well but it is limiting as it does not allow the spatial dependence amongst all 119 gauging stations to be characterized. That would involve fitting a model that includes all 119 stations, which is not computationally feasible. Also, a conditioning river would need to be chosen. An alternative way of applying the model to characterize the spatial dependence in the whole of Scotland would be to build a model at each gauging station, using the gauging station itself as the conditioning variable, and include in the model all the gauging stations within a certain distance. This approach was followed by Keef et al. (2009) to investigate spatial dependence in the UK.

The models described in Section 5.1.1 have the aim of identifying the distribution of extreme values in space. "A typical goal in applications is the estimation of a high quantile of the distribution of the response variable Y" (Davison et al. (2012)). Hence a quantile model might be preferred, as a way of directly obtaining the quantile without having to worry about the whole distribution.

5.5 Spatial quantile regression

It is recognized that environmental processes are highly variable and as such investigation of the average spatial trend might not be informative enough, especially when the interest is in extreme values. Despite quantile regression being now widely used in environmental applications, its application to spatial data is fairly recent and as such there is not much literature available. In its simplest form, it can be seen as an extension of quantile regression (see Chapter 4), with the difference that now the predictor is bi-variate, defined by the longitude and latitude (or similarly Northing and Easting) coordinates.

A number of approaches for spatial quantile regression are available. He et al. (1998) extend the 1-dimensional regression problem, which can be solved using linear programming methods, into a 2-dimensional context. The objective function can be re-written in terms of bi-variate smoothing splines (Koenker (2005); He et al. (1998)) and then minimized using linear programming methods. Penalization and fitting of splines coefficients becomes considerably more complex though. Alternatively, a local linear regression nonparametric approach was suggested by Hallin et al. (2009), in which the regression coefficients are allowed to vary spatially. On the other hand, Lee and Neocleous (2010); Reich et al. (2011) and Reich (2012) have proposed fitting spatial quantile regression models in a Bayesian framework. In particular, Lee and Neocleous (2010) apply quantile regression to count data to model the number of hospital admissions related to respiratory disease in Scotland and investigate its relationship with air pollution. A spatial autoregressive term is incorporated in the predictor, built as a weighted average of the response, in which non zero weights are assigned to neighboring areas. The incorporation of this covariate successfully removes residual spatial correlation. Reich et al. (2011) develop a Bayesian spatial quantile additive model for ozone with a number of meteorological variables as covariates. Observations are spatio-temporal, and the model is reparameterized using basis functions. The spatial structure is introduced by setting spatial priors on the basis coefficients. However, it is not feasible for large data sets, in which case an approximation is suggested (Reich et al. (2011)). Reich (2012) proposes a Bayesian hierarchical spatio-temporal quantile model, in which the quantile function at each spatial location is defined as a linear function of time. By making the quantile

function vary with time and space, the whole response distribution can be modelled and various quantiles can be fitted simultaneously. The quantile function, linear in time, is expressed in terms of piecewise Gaussian basis functions with spatially varying coefficients. The spatial structure is accounted for by specifying the covariance function of the Gaussian spatial processes associated with the parameters of the model (intercept and basis coefficients). The parameters are estimated using MCMC and any residual spatial correlation is modelled via a spatial copula. Results from modelling temperature data using the proposed model are presented and compared to a traditional quantile regression model. The results from both models agree, but the Bayesian model performs better as it provides a smoother spatial surface and increased power for identifying significant temporal trends at individual locations. This is achieved by borrowing information across spatial locations.

The models proposed by Lee and Neocleous (2010); Reich et al. (2011) and Reich (2012) were developed with the aim of modelling the whole distribution of the response variable. This thesis focuses on extreme values and hence modelling the whole distribution distribution is not really the ultimate goal, but modelling quantiles high up in the tail instead (e.g. 90^{th}, 95^{th} quantile). Hence, a simpler model is proposed in the next section. Even though it does not allow fitting various quantiles simultaneously, it is relatively fast to implement and therefore can be fitted for a range of quantiles independently, to investigate, for instance, how the spatial distribution of extreme values changes with respect to the mean. In the next section, a model for spatial extreme values is developed, followed by an application to the data available. The data set consists of 119 daily river flow records across Scotland over the period 1^{st} January 1996 - 31^{st} December 2005. See Appendix A for a table of the gauging stations included in the analysis.

5.6 A spatial quantile model for river flow

The approach adopted for the temporal quantile model, that considers the minimization problem as a weighted least squares problem, is extended to build a spatial quantile model. The proposed model is:

$$Q_{log(flow)_i}(\tau|x,y) = s(x,y) + \varepsilon_i \tag{5.9}$$

where $Q_{log(flow)}(\tau|x,y)$ is the τ^{th} quantile of the (conditional) distribution of log(flow) and $s(x,y)$ is a bi-variate smooth function of the easting and northing coordinates. At this point the errors ε_i are assumed to be independent, but no distributional assumption is imposed. The smooth function $s(x,y)$ can be expressed in terms of the tensor product of the marginal B-splines basis (Wood (2006); Eilers and Marx (2003, 2004); He et al. (1998)). Denote the marginal B-spline basis as $B_k(x)$, $k = 1,\ldots,n_x$, $B_j(y)$, $j = 1,\ldots,n_y$, where n_x, n_y are the number of basis functions for each variable x and y. The corresponding marginal smooth components can be expressed as:

$$S_x(x) = \sum_{k=1}^{n_x} B_k(x)\gamma_k$$
$$S_y(y) = \sum_{j=1}^{n_y} B_j(y)\gamma_j$$

The bi-variate function $s(x,y)$ can be expressed as:

$$s(x,y) = \sum_{k=1}^{n_x}\sum_{j=1}^{n_y} B_k(x)B_j(y)\gamma_{kj}$$

where $\Gamma = [\gamma_{kj}]$ is a $n_x \times n_y$ matrix of parameters. The model can be expressed in matrix notation as:

$$s(x,y) = B\gamma$$

where $\gamma = vect(\Gamma)$ is a $(n_x \times n_y) \times 1$ vector of coefficients and $B = B_x \times B_y$ is a $n \times (n_x \times n_y)$ matrix, where $\times$ is the tensor product. Matrix B can be constructed as $B = (B_x \bigotimes 1_{n_y}) \bigodot (1_{n_x} \bigotimes B_y)$, where $\bigodot$ represents element-wise multiplication and $\bigotimes$ is the Kronecker product. By expressing the model as a linear model, the parameters can be estimated easily using efficient matrix-vector operations. A penalty term needs to be added to control the amount of smoothness. The penalty is constructed in a similar way as in the univariate case, with the difference that now penalties are set individually on the rows and columns of the matrix B. The penalty term becomes then:

$$\lambda_x||P_x\gamma||^2 + \lambda_y||P_y\gamma||^2$$

where λ_x, λ_y are smoothing parameters corresponding to x and y respectively. The individual penalties P_x, P_y are constructed based on the penalty in the univariate case

(Eilers and Marx (2003)) following:

$$P_x = (D_d^T D_d) \bigotimes I_{n_y}$$
$$P_y = I_{n_x} \bigotimes (D_d^T D_d)$$

where D_d is a difference matrix of order d and I_n is the identity matrix. The (approximate) objective function for a spatial quantile regression becomes:

$$||W(y - B\gamma)||^2 + \lambda_x ||P_x \gamma||^2 + \lambda_y ||P_y \gamma||^2 \tag{5.10}$$

where W is the matrix of weights calculated following Equation (4.8). The vector of parameters γ can be estimated using PIRLS. At iteration (j):

$$\gamma^{(j)} = (B^T W^{(j-1)} B + \lambda_x P_x + \lambda_y P_y)^{-1} B^T W^{(j-1)} y \tag{5.11}$$

5.6.1 Choice of smoothing parameters

In the temporal model, visual inspection allowed the desired degrees of freedom of the model to be determined and hence an appropriate smoothing parameter λ. However, in two dimensional smoothing visual assessment is somewhat difficult. He et al. (1998) suggest choosing smoothing parameters $\lambda = (\lambda_x, \lambda_y)$ to minimize the SIC:

$$\text{SIC}(\lambda) = \log \left[\frac{1}{n} \sum_{i=1}^{n} \rho_\tau (y_i - \hat{g}(x_i)) \right] + \frac{1}{2n} df_\lambda \log n$$

The SIC approach was adopted here, despite not being optimal for correlated data (Opsomer et al. (2001)).

5.6.2 Accounting for spatial correlation

In order to model residual spatial correlation, the error terms ε_i in Model (5.9) are assumed to be realizations of an underlying spatial process $Z(x)$ with quantile τ zero, variance σ^2 and correlation structure $\rho(h)$. The process $Z(x)$ is said to be (weakly) stationary if the expected value does not depend on x and the covariance $Cov(Z(x), Z(x+h))$ only depends on h. The spatial process $Z(x)$ can be modelled through the semivariogram

function:

$$\gamma(x_i, x_j) = \frac{1}{2}\text{Var}(Z(x_i) - Z(x_j))$$

The process $Z(x)$ is said to be isotropic if $\gamma(x_i, x_j)$ only depends on $||h|| = ||x_i - x_j||$. Assuming the spatial process $Z(x)$ to be stationary and isotropic, the correlation structure of the residuals can be informally assessed using the empirical semi-variogram (Diggle and Ribeiro Jr. (2007)):

$$\gamma(h) = \frac{1}{2|N_h|}\sum_{h\in N_h}(Z(x_i + h) - Z(x_i))^2 \tag{5.12}$$

where N_h is the paired points whose distance lies within a given neighborhood of h.

5.6.3 Uncertainty: standard error estimation

Model (5.9) is fitted assuming independent observations. If that was the case and observations were actually independent, standard errors for the fitted values could be calculated in exactly the same way as described in Section 4.4.2 of Chapter 4. A summary of the equations needed is included below (Section 5.6.3.1). To adjust for dependence, the formula for calculating the standard errors need to be modified to incorporate the correlation matrix V, which can be estimated using the semi-variogram (Equation (5.12)). The equations in the case of correlated data are summarized in Section 5.6.3.2.

5.6.3.1 Case 1: independent fitted values

$$se(\hat{y}) = \sqrt{diag(SS^T)\sigma^2}$$

where S is the smoothing matrix at the last iteration:

$$S = B(B^TWB + \lambda_x P_x + \lambda_y P_y)^{-1}B^TW$$

and W is the weight matrix at the last iteration. The variance σ^2 can be estimated as:

$$\hat{\sigma}^2 = \frac{RSS}{df_{error}}$$

$$RSS = Y^T(I - S)^TW(I - S)Y$$

$$df_{error} = tr((I - S)^T W (I - S))$$

5.6.3.2 Case 2: correlated fitted values

$$se(\hat{y}) = \sqrt{diag(SVS^T)\sigma^2}$$

where S is the smoothing matrix at the last iteration:

$$S = B(B^T W B + \lambda_x P_x + \lambda_y P_y)^{-1} B^T W$$

W is the weight matrix at the last iteration and V is the correlation matrix. Both V and σ^2 are estimated using the empirical semi-variogram.

The aim of spatial models is often not only to obtain fitted values at the observed locations but also make predictions at locations were data was not available. This is usually referred to as kriging (Diggle and Ribeiro Jr. (2007)). For Model (5.9), predicted values at unobserved locations can be obtained by setting up a grid of values for the spatial coordinates which is then used to create a matrix of B-spline basis B_{grid}. The number of basis functions has to be the same as the number of basis functions used to fit the model. The corresponding equations for uncorrelated and correlated data are summarized in Section 5.6.3.3 and Section 5.6.3.4 respectively.

5.6.3.3 Case 3: independent fitted surface

$$se(\hat{y}) = \sqrt{diag(S_{grid} S_{grid}^T)(\sigma^2 + 1)}$$

where S_{grid} is the smoothing matrix:

$$S_{grid} = B_{grid}(B^T W B + \lambda_x P_x + \lambda_y P_y)^{-1} B^T W$$

and S and W are the model smoothing and weight matrices at the last iteration. The variance σ^2 can be estimated as before:

$$\hat{\sigma}^2 = \frac{RSS}{df_{error}}$$

$$RSS = Y^T (I - S)^T W (I - S) Y$$

$$df_{error} = tr((I - S)^T W (I - S))$$

5.6.3.4 Case 4: correlated fitted surface

$$se(\hat{y}) = \sqrt{diag(S_{grid} V S_{grid}^T)(\sigma^2 + 1)} \tag{5.13}$$

where S_{grid} is the smoothing matrix

$$S_{grid} = B_{grid}(B^T W B + \lambda_x P_x + \lambda_y P_y)^{-1} B^T W$$

S and W are the model smoothing and weight matrices at the last iteration and V is the correlation matrix. Both V and σ^2 are estimated using the empirical variogram.

5.7 Application to daily river flow data in Scotland

Daily river flow data are available for 119 gauging stations over the period January 1996-December 2005. At each individual location, the mean flow was removed as a way of standardizing the data. This is done to account for differences in flow values due to catchment size. The proposed spatial model (5.9) does not have a time component and hence it must be fitted for every time point individually. However, this implies that only 119 observations are available, which is too small a number for fitting a quantile model, especially for extreme values of τ. To overcome this problem and increase the number of observations, a whole month's data were considered at each location. The model was fitted independently for each month (January-December) and year (1996-2005), resulting in 120 models in total. A value of τ=0.95 was chosen to investigate the spatial pattern of extreme river flow. 15 B-spline basis functions were used for each marginal basis ($n_x = n_y$=15), with a total number of 225. A second order penalty was set on each margin. Smoothing parameters in each direction were chosen to be λ_x=2.632 and λ_y=1.054 based on the SIC criteria. In order to make the fitted models comparable, the same smoothing parameters were used for all 120 of them, even though in some cases the chosen values did not minimize SIC. The fitted surfaces can be seen in Figures 5.11-5.20. As a way of informally assessing whether the model was appropriate or not, the proportion of residuals above and below the fitted surface was calculated. The proportions of positive residuals ranged from 93.98% to 94.71% (mean 94.45%) and the negative ones

from 5.29% to 6.02% (mean 5.55%), very close to the theoretical expected values of 95% and 5% respectively.

Figures 5.11-5.20 show considerable variability across months and years. Overall, there is an East-West difference but there are some months in which the pattern is not so well defined (e.g. January 1996, December 2002). On occasions, the pattern is more pronounced along the North-South gradient (e.g. September 2002). In general, values tend to be lower (blue colour) during the summer months (June, July, August), as expected, with a clear East-West gradient in most cases. However, there are a few time points (August 1996, July and August 1998, July 1999, August 2000, 2001 and 2005) at which small areas on the South West show considerably higher flow values (yellow colour), comparable to the values reached in some of the winter months. In particular, July 2002 and August 2004 show very different features with respect to the remaining summer months, with a large area of high values in the North-East during July 2002 and high values over the whole of Scotland during August 2004. During the winter months, particularly high values across the whole of Scotland can be seen in February 1997, December 1999, winter of 2000 and 2002, January and October 2004 and January 2005. In general, over the winter months, there is a clear East-West gradient, with Western values usually higher than those in the East, although on occasions the gradient is reversed, e.g. in April 1998, April 2000 and October 2002.

Residual spatial correlation was investigated using the empirical semi-variogram (Equation (5.12)). Since there is more than one observation per spatial location the empirical semi-variogram has to be modified. Assuming that the residual spatial correlation is constant over time, i.e. within each month, the pooled variogram can be calculated as (Pebesma and Duin (2005)):

$$\gamma(h) = \frac{1}{2\sum_{t=1}^{n_{month}} N_{h(t)}} \sum_{t=1}^{n_{month}} \sum_{h=1}^{N_{h(t)}} (Z(s,t) - Z(s+h,t))^2$$

where n_{month} represents the number of days in the corresponding month and $N_{h(t)}$ the available number of point pairs with separation distance close to h for day t. The pooled empirical variograms for 1996 can be seen in Figure 5.21. For most of the months, the

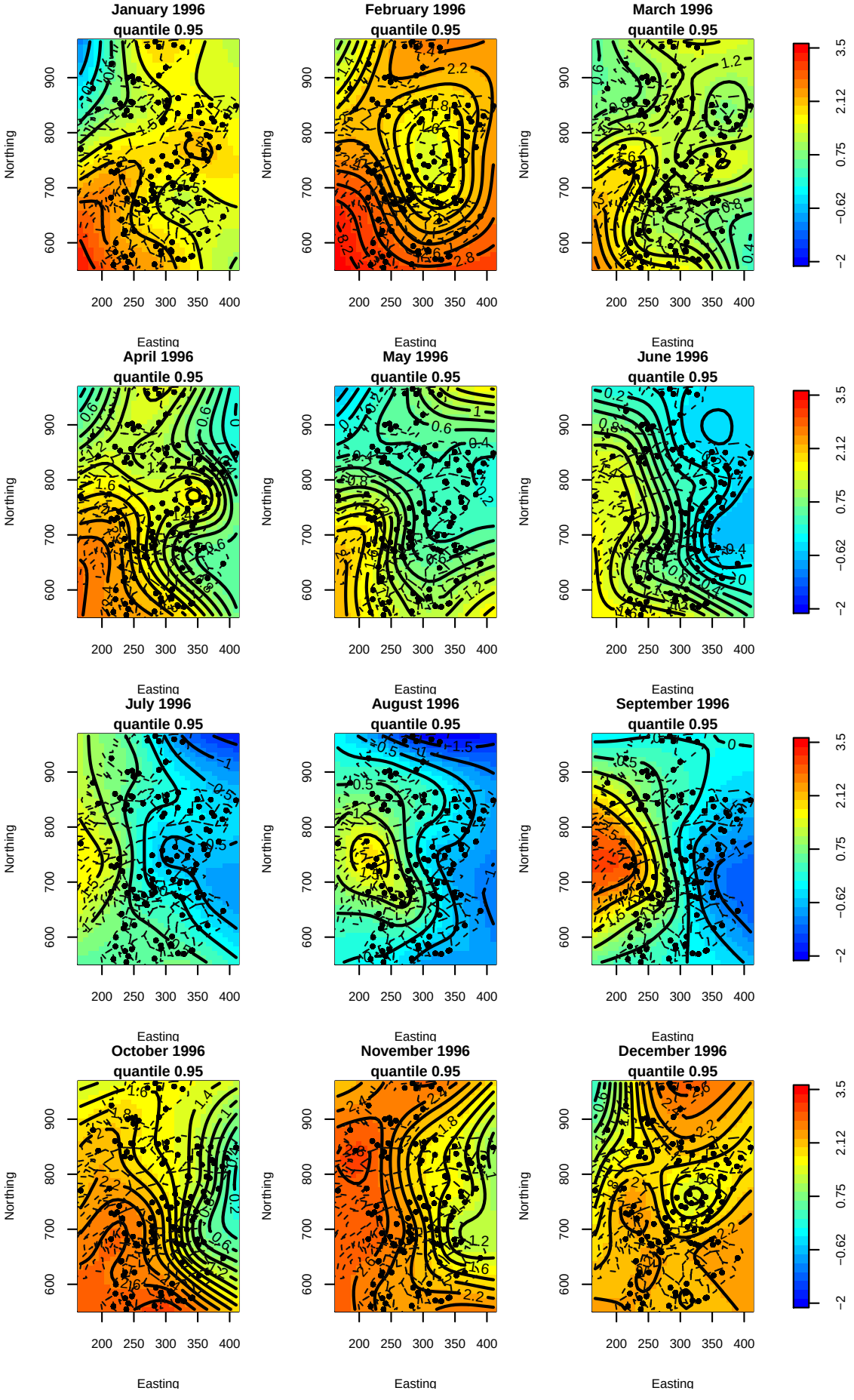

FIGURE 5.11: Fitted 95^{th} quantile surfaces (1996)

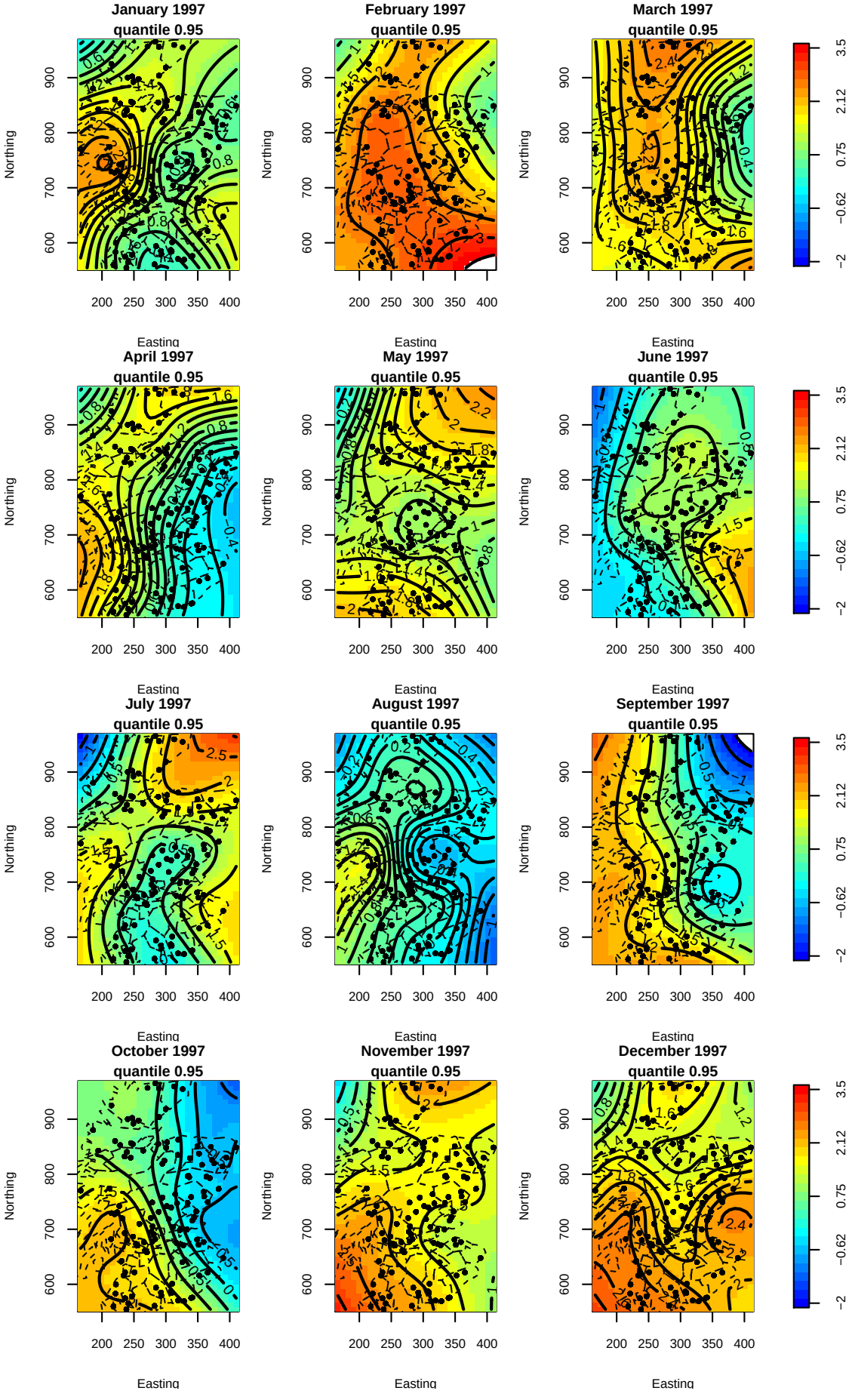

FIGURE 5.12: Fitted 95^{th} quantile surfaces (1997)

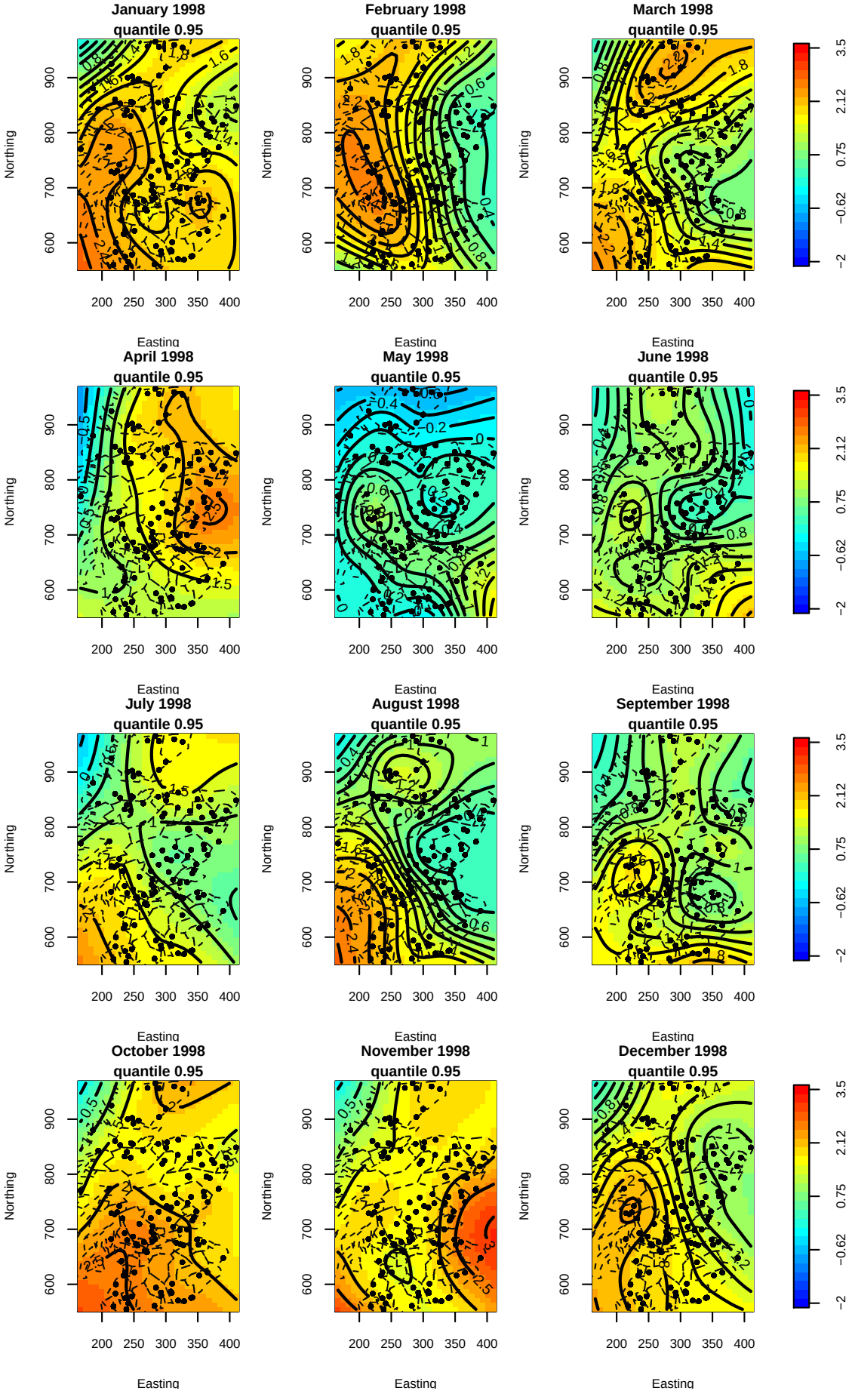

FIGURE 5.13: Fitted 95^{th} quantile surfaces (1998)

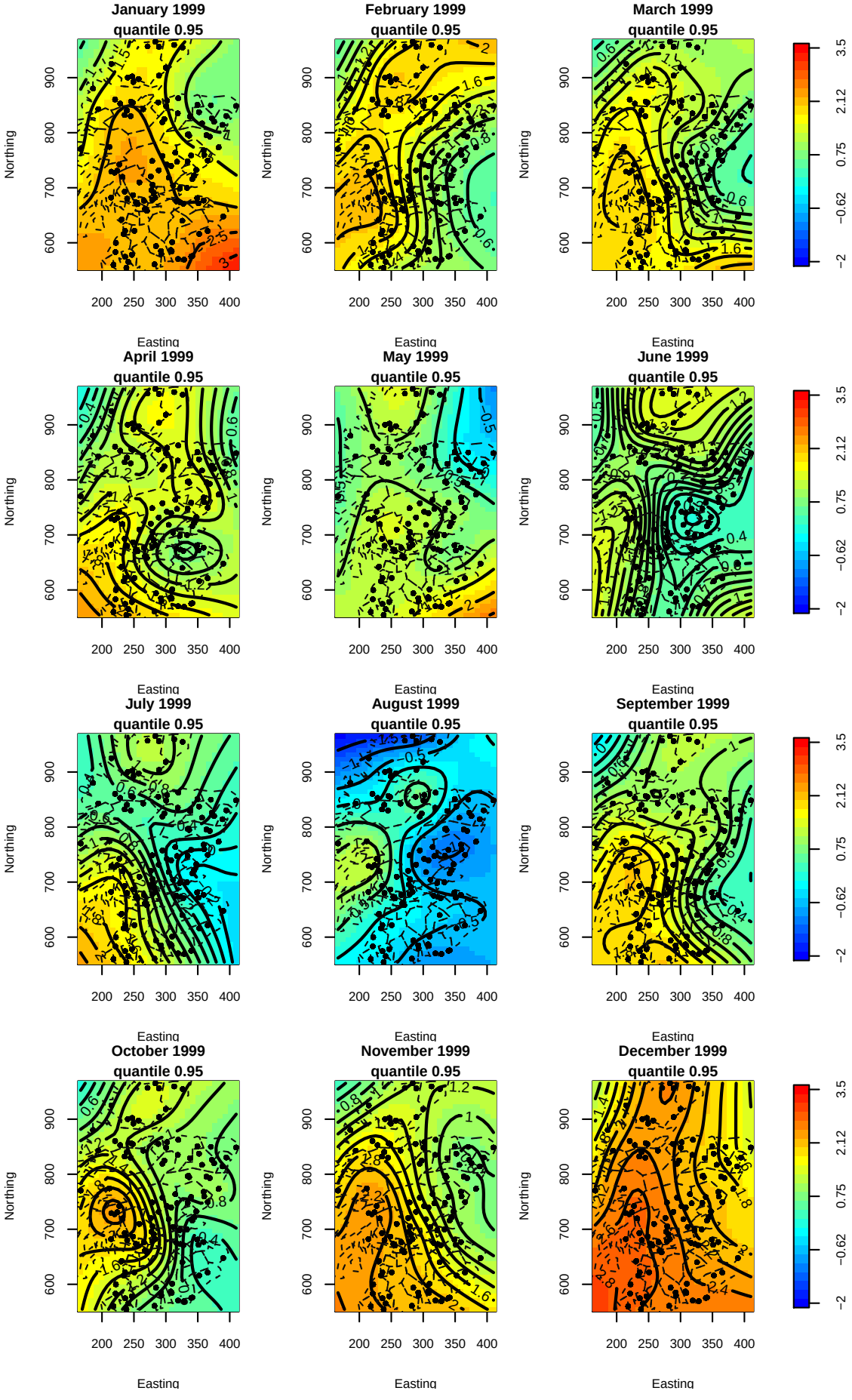

FIGURE 5.14: Fitted 95^{th} quantile surfaces (1999)

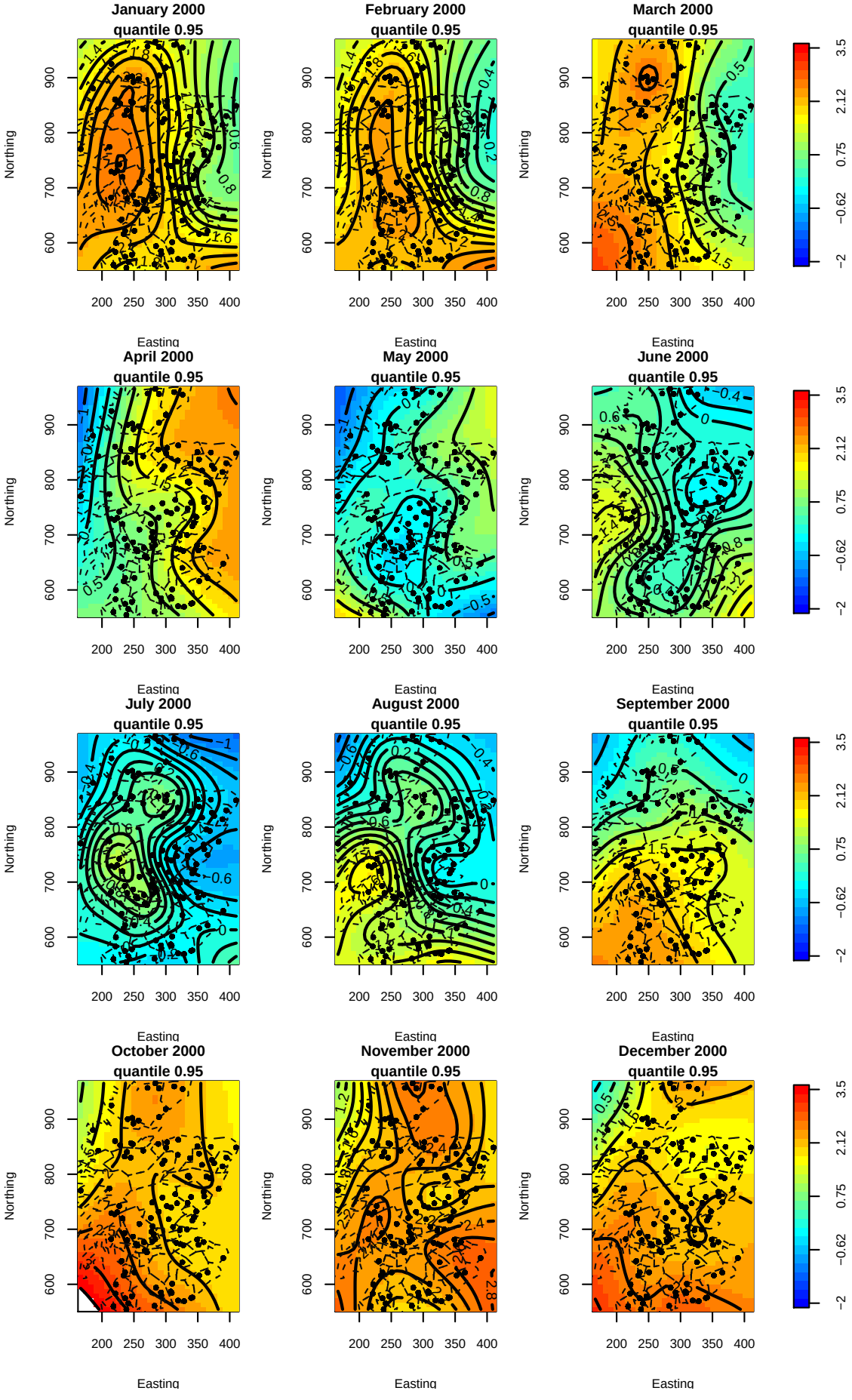

FIGURE 5.15: Fitted 95^{th} quantile surfaces (2000)

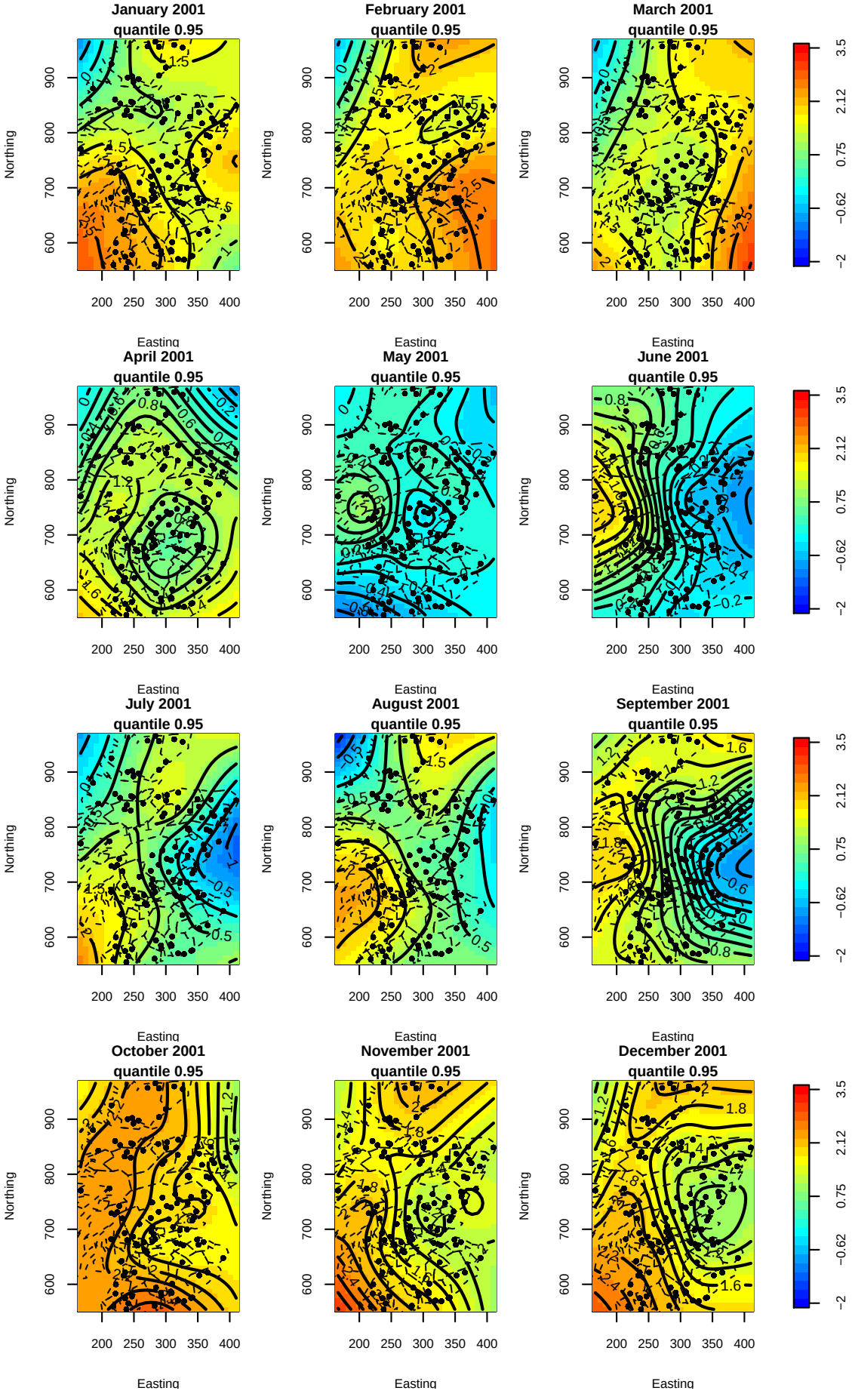

FIGURE 5.16: Fitted 95^{th} quantile surfaces (2001)

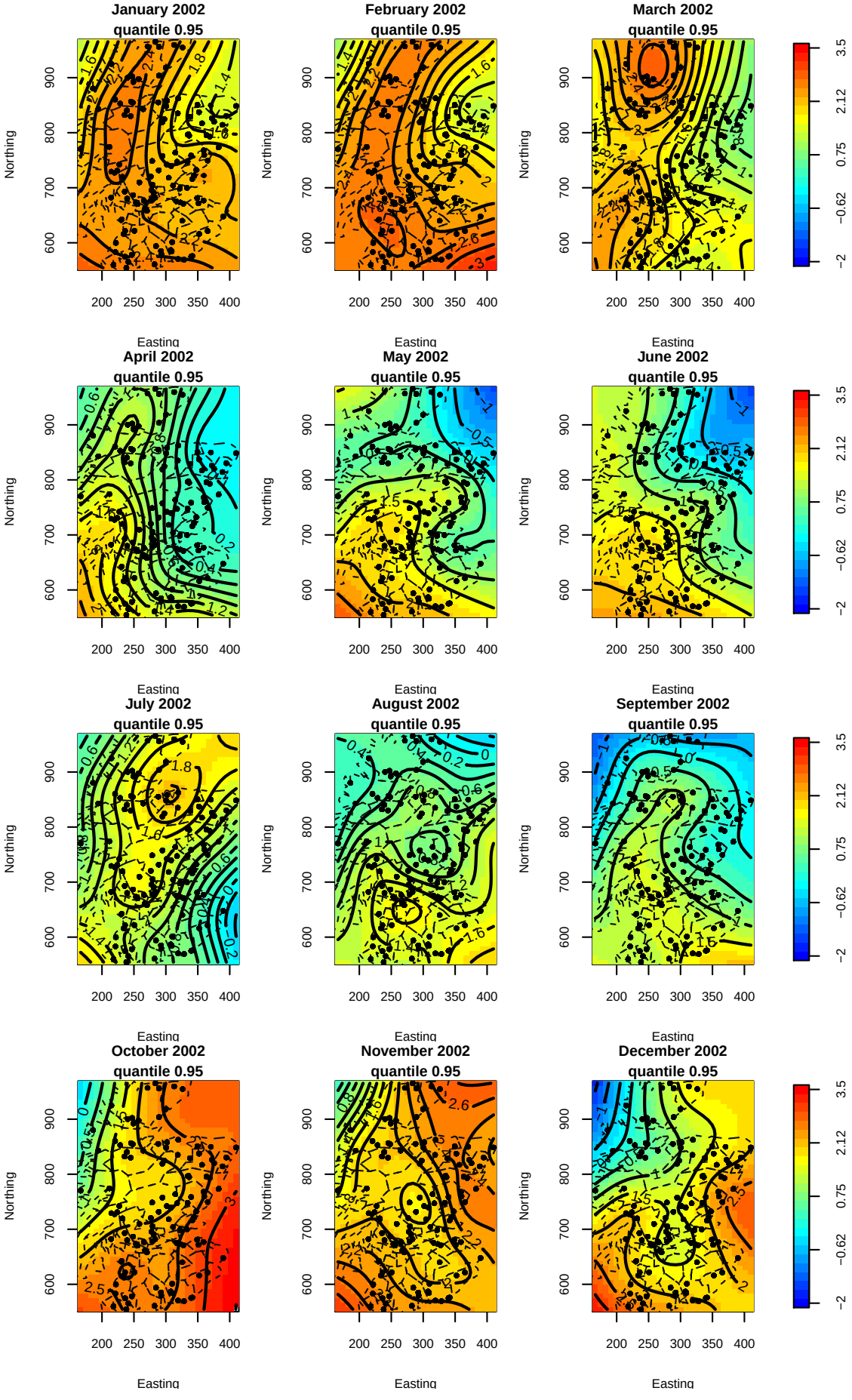

FIGURE 5.17: Fitted 95^{th} quantile surfaces (2002)

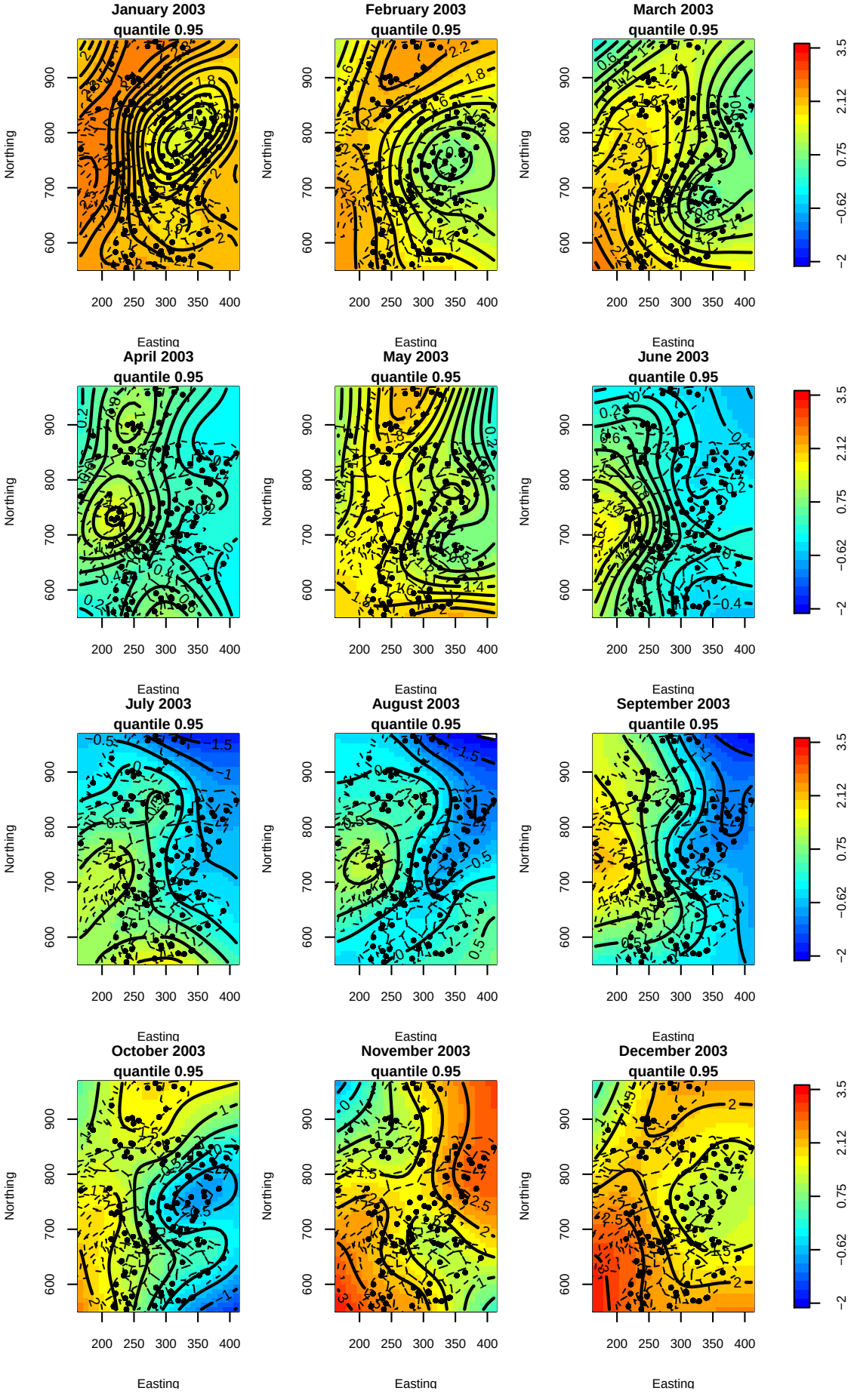

FIGURE 5.18: Fitted 95^{th} quantile surfaces (2003)

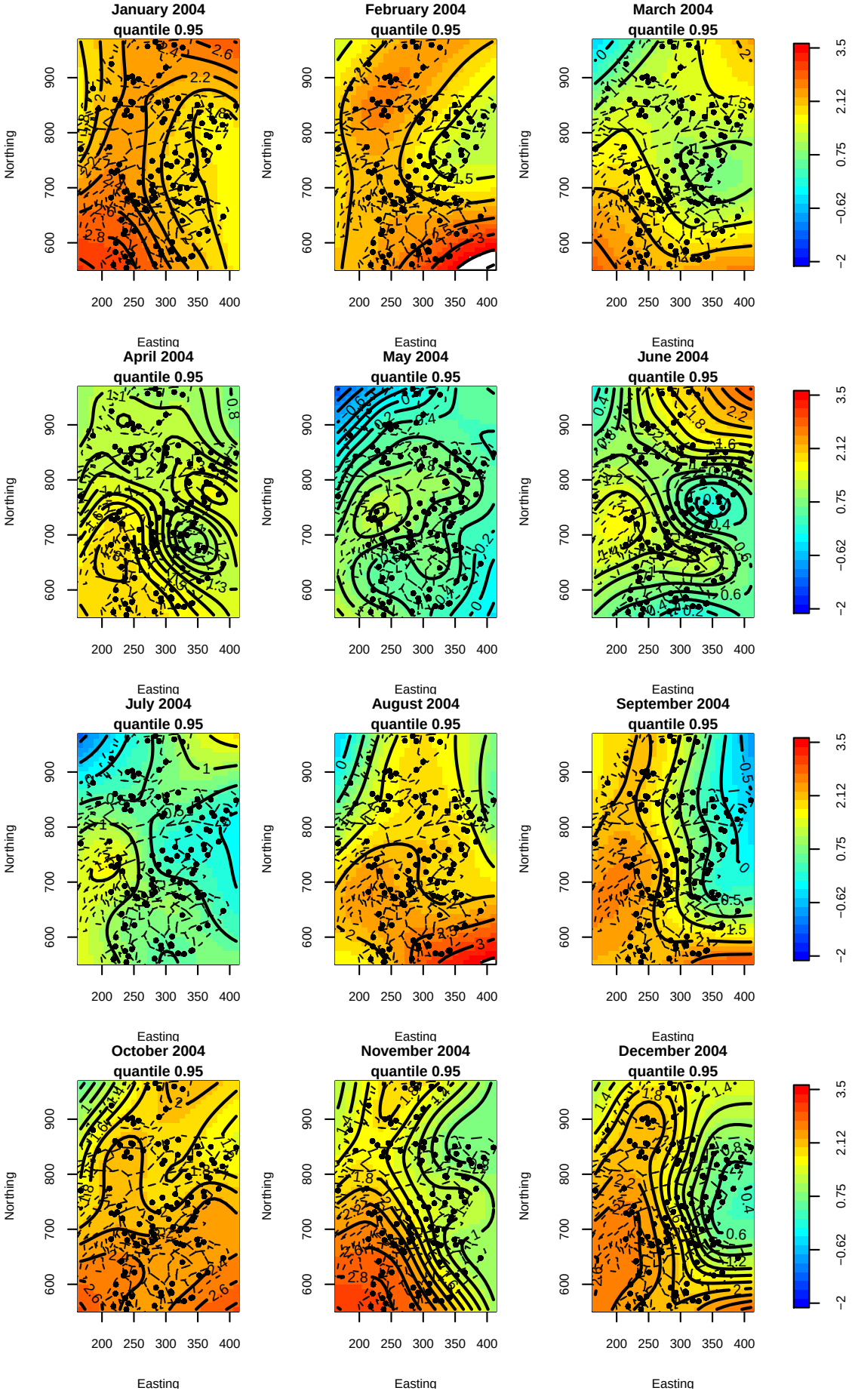

FIGURE 5.19: Fitted 95^{th} quantile surfaces (2004)

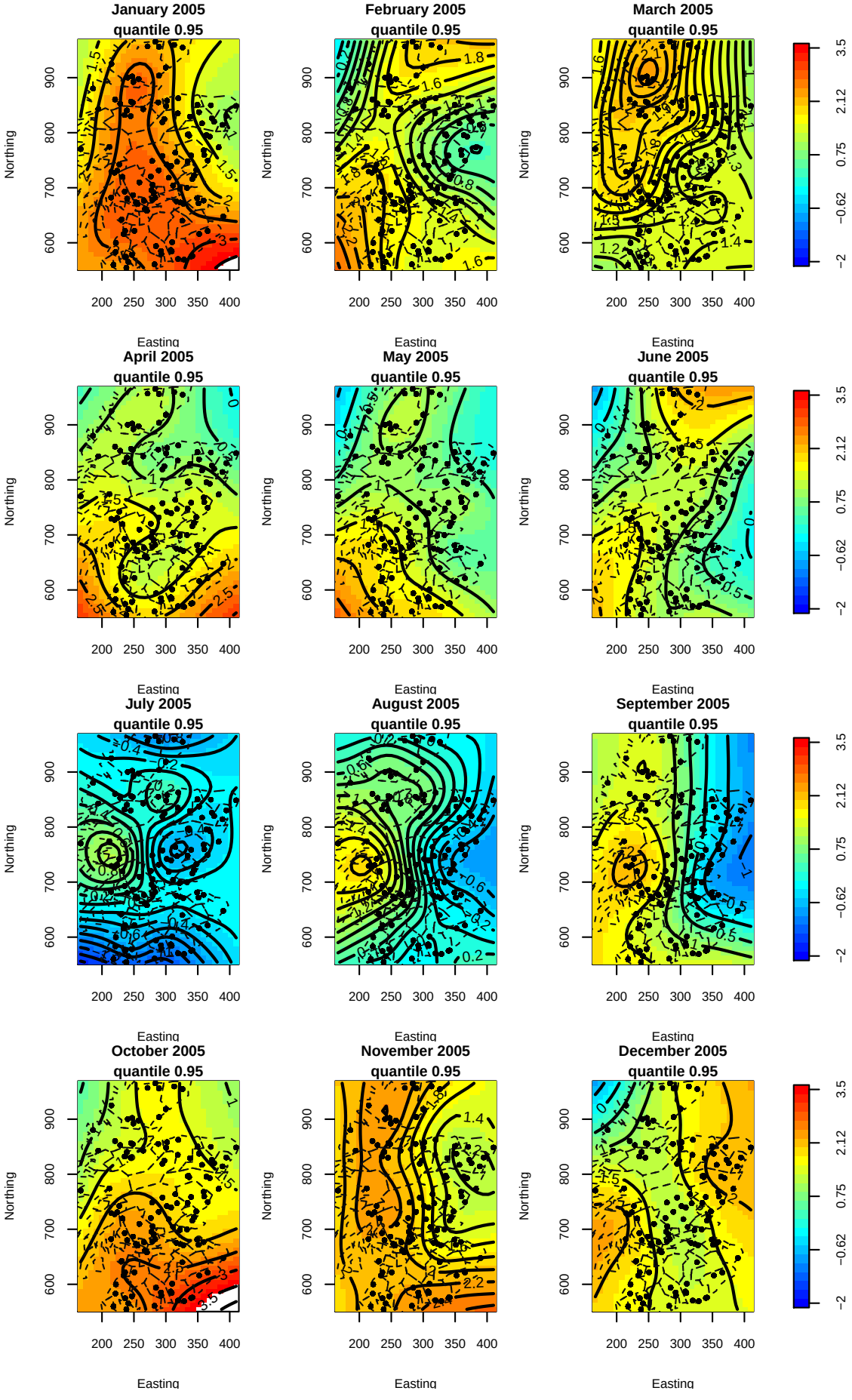

FIGURE 5.20: Fitted 95^{th} quantile surfaces (2005)

empirical pooled variogram is upper-bounded.

A theoretical model was then fitted to the empirical pooled semi-variogram for every month and year. A range of models have been proposed in the literature. The functions that characterize them are always negative definite. The Matern family (Equation (5.7)) is widely used and provides a great range of flexibility (Diggle and Ribeiro Jr. (2007)). It is characterized by a set of parameters: ϕ >0 is the range parameter, σ^2 is the partial sill and $k > 0$ is the order (and determines the smoothness of the function). The order parameter is usually chosen a priori. The range ϕ is defined as the distance beyond which locations will be spatially independent. When the variogram reaches its sill asymptotically, the practical range is defined as the distance at which the variogram reaches 95% of its sill (Cressie (1993)). τ^2 is the nugget and can be associated with measurement error. The Matern family can be represented as:

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ \tau^2 + \sigma^2 \left(1 - \left\{2^{k-1}\Gamma(k)\right\}^{-1} \left(\frac{h}{\phi}\right)^k \kappa_k(h/\phi)\right) & h > 0 \end{cases}$$

where κ_k is a modified Bessel function (Diggle and Ribeiro Jr. (2007)). Two special cases of the Matern family are the exponential (k=0.5) and the Gaussian ($k \to \infty$):

- Exponential model

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ \tau^2 + \sigma^2 \left(1 - \exp\left(-\frac{h}{\phi}\right)\right) & h > 0 \end{cases}$$

- Gaussian model

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ \tau^2 + \sigma^2 \left(1 - \exp\left(-\left(\frac{h}{\phi}\right)^2\right)\right) & h > 0 \end{cases}$$

Another commonly used model is the spherical model:

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ \tau^2 + \sigma^2 \left(\frac{3}{2}\frac{h}{\phi} - \frac{1}{2}\left(\frac{h}{\phi}\right)^3\right) & 0 \le h \le \phi \\ \tau^2 + \sigma^2 & h > \phi \end{cases}$$

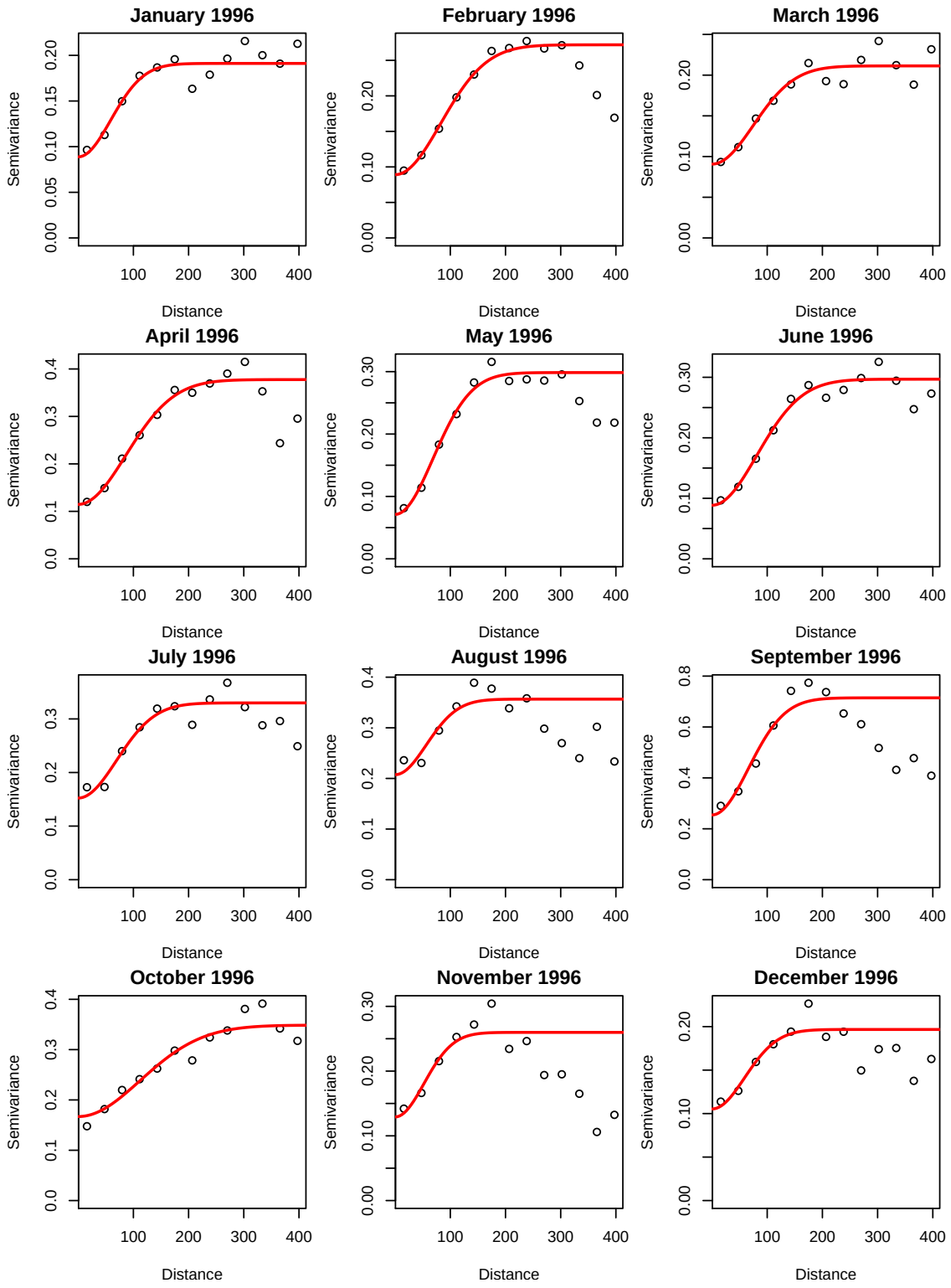

FIGURE 5.21: Emprical variograms (points) and model based variogram (red solid line) based on the residuals from the 95^{th} quantile fitted model. Note the scale on the y axis changes across graphs

Month	$\hat{\tau^2}$	$\hat{\sigma}^2$	$\hat{\phi}$	Practical Range
January	0.089	0.102	83.40	144.36
February	0.089	0.183	116.72	202.03
March	0.091	0.121	105.69	182.93
April	0.115	0.263	122.14	211.41
May	0.071	0.227	97.19	168.21
June	0.089	0.208	114.47	198.13
July	0.152	0.178	98.01	169.64
August	0.207	0.149	79.84	138.19
September	0.254	0.460	92.84	160.69
October	0.167	0.182	158.63	274.56
November	0.129	0.131	73.84	127.81
December	0.105	0.091	83.63	144.74

TABLE 5.3: Summary of fitted parameters for variogram, 1996, quantile 0.95

The exponential, Gaussian, Matern (k=2.5) and spherical models were investigated, the Gaussian one providing the best fit for most of the months and years. Hence, a Gaussian model was adopted for the residual correlation. The parameters were fitted using curve fitting methods, in particular least squares were used. Denote by θ the vector of parameters, estimates of θ are obtained by minimizing the function:

$$S_n(\theta) = \sum n_{h_i}(\hat{\gamma}(h_i) - \gamma(h_i, \theta))^2$$

where n_{h_i} is the number of points in each bin (Diggle and Ribeiro Jr. (2007)). A weighted version was proposed by Cressie (Cressie (1993)):

$$S_c(\theta) = \sum n_{h_i} \left(\frac{\hat{\gamma}(h_i) - \gamma(h_i, \theta)}{\gamma(h_i, \theta)} \right)^2$$

The estimated parameters for 1996 are summarized in Table 5.3. The resulting fitted variograms are shown in red in Figure 5.21. Visual inspection of Figure 5.21 suggests that the model is a good fit for most of the months, although there are some cases in which the empirical variogram does not fit well.

Using the fitted variogram model and Equation (5.13), standard errors accounting for spatial correlation for the fitted surfaces from Figures 5.11-5.20 were estimated. These are shown in Figure 5.22 at three different time points. Since the fitted parameters of the theoretical variogram (Table 5.3) showed considerable variation across months, a pooled estimate was not considered to be appropriate. Hence standard errors for each

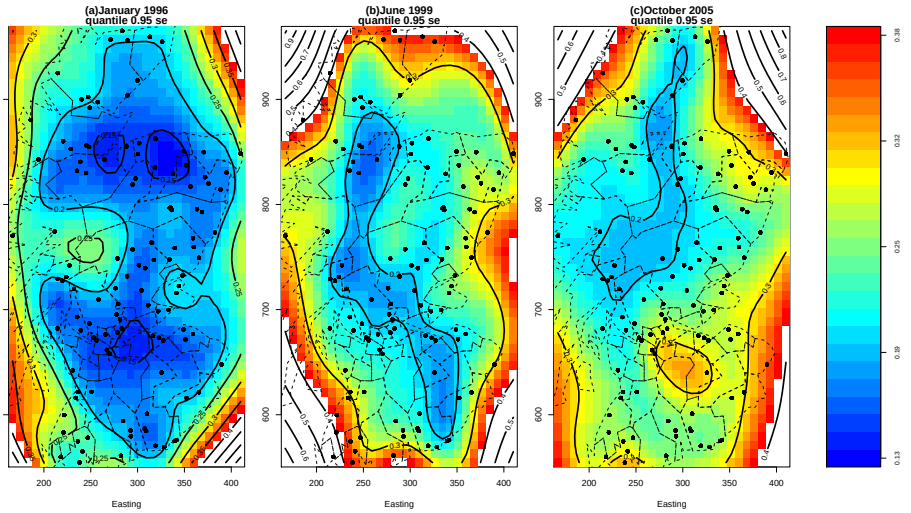

FIGURE 5.22: Standard errors for the fitted 95% quantile river flow surface for (a) January 1996, (b) June 1999 and (c) October 2005

individual month were estimated using the month specific estimated parameters. The resulting standard errors are higher close to the border of Scotland as there are less observations available and lower where there are a few stations close together. The pattern is not exactly the same for every month as slightly different models were used in each case. Standard errors in January 1996 (Figure 5.22(a)) seem to agree well with the data locations in the sense that they have lower values (blue colour) where a large number of observations are found close together. The pattern is slightly different in June (Figure 5.22(b)), when standard errors are lower in the West than in the East, and October (Figure 5.22(c)), when standard errors tend to be lower in the South than in the North. These differences are due to the differences in the model parameters among the three months, as shown in Table 5.3. The estimated variability in June and October is higher than in January, and so is the range parameter.

The fitted 95^{th} quantile spatial model suggest that the spatial pattern has changed over the years. This means that to appropriately model the data, a more complex model that incorporates both time and space should be used.

5.8 Spatio-temporal modelling: a first approach

The models presented in Chapter 4 and Chapter 5 can be combined to obtain a spatio-temporal quantile model. The simplest one is a fully additive model:

$$Q_{\log(flow)_i}(\tau|x_i, y_i, z_i) = s_1(x_i) + s_2(y_i) + s_3(z_i) + \varepsilon_i \tag{5.14}$$

where ε_i, are assumed to be independent and $s_1(x)$, $s_2(y)$ and $s_3(z)$ are smooth functions of time, day of the year and space that represent the temporal, seasonal and spatial trends respectively. The seasonal component $s_2(y)$ has been simplified to be constant over the years at this point, while in the quantile temporal model proposed in Chapter 4 it was allowed to change from year to year. The reason for this simplification is that the spatio-temporal model was not computationally feasible otherwise. Each of the smooth functions can be re-expressed as a linear combination of a set of B-spline basis functions:

$$\begin{aligned} s_1(x) &= \textstyle\sum_{j=1}^{k_1} B_{1,j}(x)\beta_{1,j} \\ s_2(y) &= \textstyle\sum_{j=1}^{k_2} B_{2,j}(y)\beta_{2,j} \\ s_3(z) &= \textstyle\sum_{j=1}^{k_3} B_{3,j}(z)\beta_{3,j} \end{aligned}$$

The model can be expressed in matrix form as:

$$y = B\beta$$

where $B = [B_1 B_2 B_3]$ is the matrix that results from combining the individual matrices B_1, B_2 and B_3 and $\beta^T = (\beta_1, \beta_2, \beta_3)^T$. Penalties can be added in a similar way as was done in Chapters 4 and 5 to control the amount of smoothness, and the model parameters β estimated using the penalized iterative weighted regression approach introduced in Section 4.4.

The results from the spatial quantile model fitted in Section 5.7 show a spatial pattern highly variable from year to year. Hence it might be desirable to include an interaction term to model explicitly how the spatial structure changes over time. Model (5.15) can be easily extended to include a space-time interaction term $s_4(x, z)$:

$$Q_{\log(flow)_i}(\tau|x_i, y_i, z_i) = s_1(x_i) + s_2(y_i) + s_3(z_i) + s_4(x_i, z_i) + \varepsilon_i \tag{5.15}$$

The smooth function $s_4(x_i, z_i)$ can be expressed in terms of the tensor product of the marginal B-spline basis of x_i and z_i as detailed in Section 5.6.

Model (5.15) was fitted to the data set of 119 rivers, with 3703 daily observations (just over 10 years) for each river. The trend ($s_1(x)$), seasonal ($s_2(y)$) and spatial ($s_3(z)$) components of the model were built as smooth functions of x=year (1995 to2005), y=day of the year (1 to 365) and z=(Easting, Northing), respectively. The number of basis functions were chosen to be k_1=10 for the trend component, k_2=6 for the seasonal component and $k_3 = 8^2$ for the space component. That means that the spatio-temporal interaction $s_4(x, z)$ will have $8^2 \times 10$ basis functions. The smoothing parameters were set to λ=0.1 in all terms. The trend, seasonal and spatial main effects are shown in Figure 5.23. The interaction effect is shown in Figure 5.24.

The overall estimated trend (Figure 5.23, top left) shows a steady increase until 1999 and a sharp decrease in 2003 to then rise again. There is a strong seasonal effect, as expected. The overall estimated seasonal pattern (Figure 5.23, bottom left) shows lower values during the summer (reaching its minimum in July) and higher values during the winter months. The estimated spatial pattern (Figure 5.23, right) suggests a pronounced South-West to North-East gradient. The interaction term (Figure 5.24) represents the adjustment that needs to be made to the overall trend from Figure 5.23 at every individual year. These adjustments are more pronounced close to the coast of Scotland and vary from year to year, although in some cases they seem to be persistent over 3-4 years.

The full model includes 720 parameters. Such a complex model raises a number of computational issues. With these specifications, running the model took 1.9 hours on a CPU with 3.3G. While time-wise the model is feasible, problems were found in memory requirements. Using a larger number of basis functions was not possible as that resulted in matrix storage issues. Even if the model was fitted for k_1=10, k_2=6 and $k_3 = 8^2$, the hat or smoothing matrix cannot be stored either. This is due to the large size of the data set, which contains 119×3703=440657 observations in total. Not being able to store the smoothing matrix means that calculating degrees of freedom or estimating standard errors is not possible.

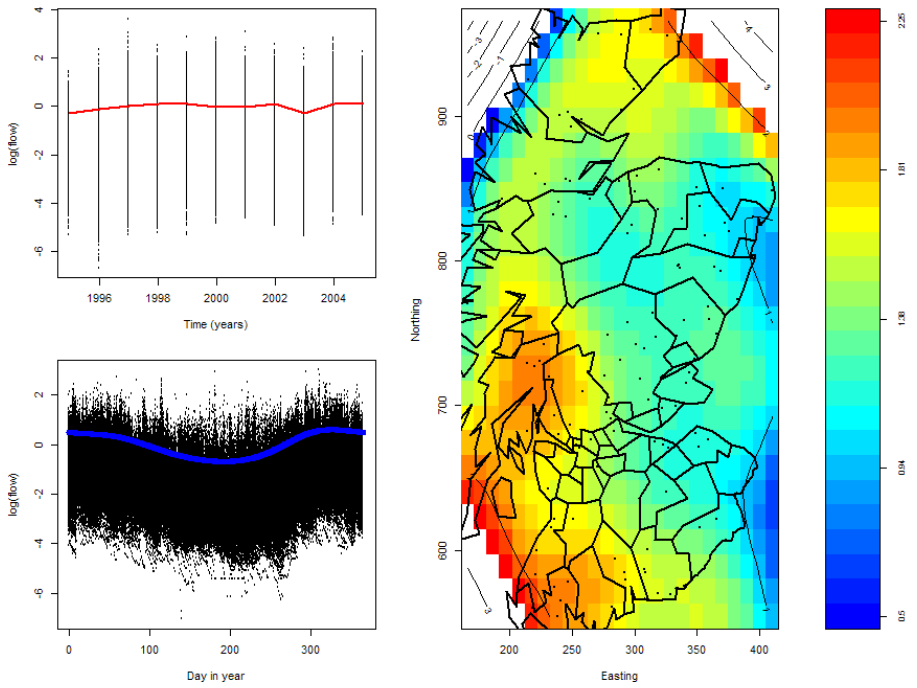

FIGURE 5.23: Estimated main effects for the 95^{th} quantile fitted model: trend (top left), seasonality (bottom left) and spatial component (right). Units are in $\log(m^3/s)$

5.9 Summary and discussion

This chapter presents two different approaches to model spatial extremes and their application to extreme river flow data in Scotland. The first approach consists of a conditional probability model and concentrates on small subsets or rivers. The second approach is a spatial quantile regression.

The two approaches can be related to each other. At a given gauging station, the first approach estimates the (conditional) probability associated with a particular return value (or quantile), while the second one provides the quantile associated with a fixed probability, taking into account the spatial structure by borrowing information across locations. However, the conditional probability model assumes stationarity, very unlike to be the case given the results obtained in previous chapters. On the other hand, the spatial quantile regression approach directly models the spatial trend without imposing any assumptions. Once the estimated spatial trend is removed, it is reasonable to assume the residuals to be stationarity.

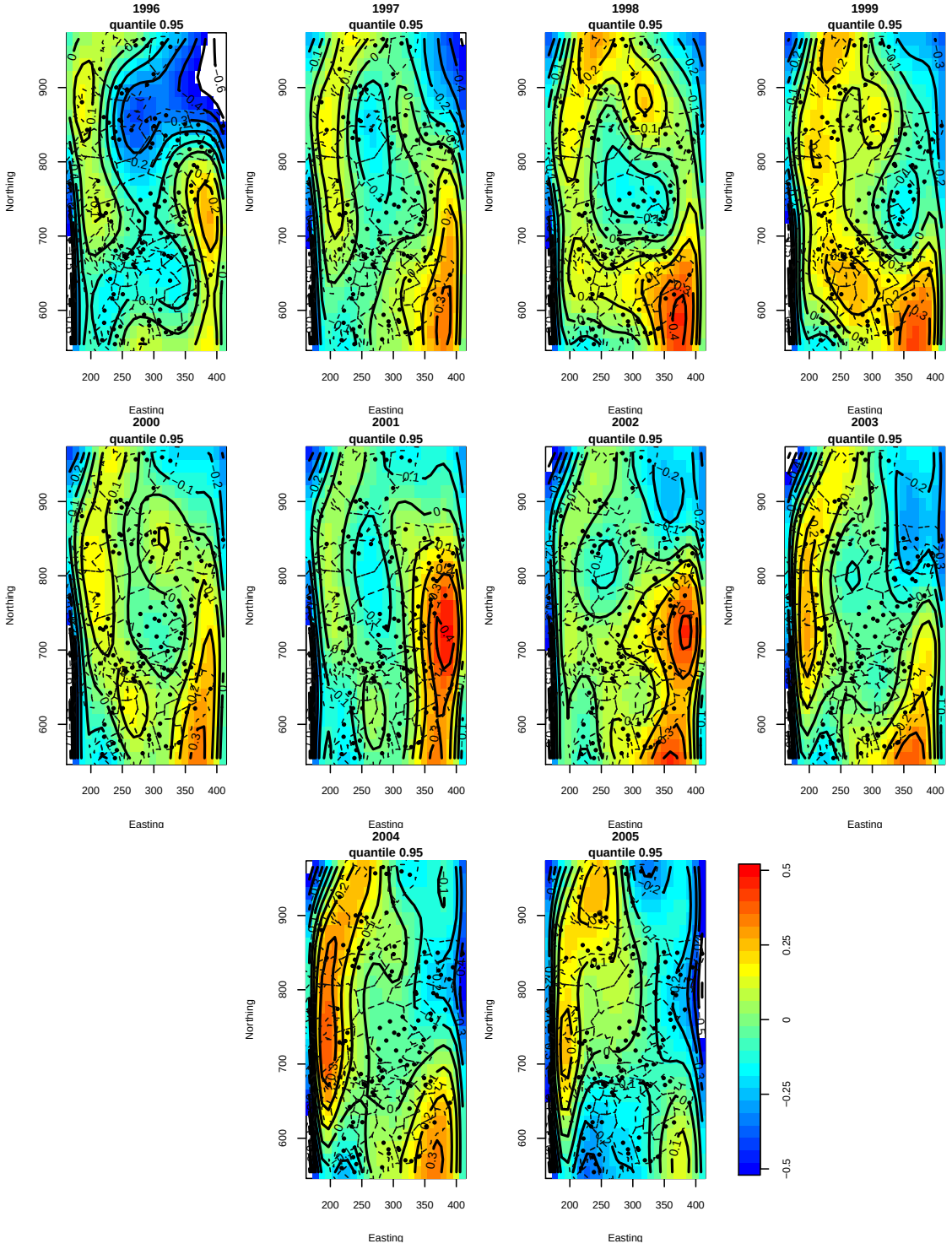

FIGURE 5.24: Estimated interaction between year and space ($\hat{s}_4(x, z)$) for the 95^{th} quantile fitted model. Units are in $\log(\mathrm{m}^3/\mathrm{s})$

5.10 Hydrological findings

The conditional probability model is applied to two sets of Scottish rivers, one in the North and one around the city of Glasgow. The model was also applied to the eight rivers followed throughout the thesis, divided into Eastern and Western rivers. The estimated conditional probabilities seem to reflect the different catchment characteristics and dominant precipitation generating events.

The fitted 95^{th} quantile surfaces show that the spatial pattern of extreme river flow in Scotland has changed over the period 1996-2005. Even within a year, the seasonal pattern changes across months, in agreement with previous studies that suggest changes in extreme rainfall and river flow to be seasonally dependent, as described in Chapter 1.

5.11 Statistical issues

The conditional probability model can be seen as a peak-over-threshold approach, meaning that a suitable threshold needs to be chosen in advance to fit the model. A sensitivity analysis, in which the stability of extrapolations and the assumption of independence were assessed, was run to choose the threshold. Here a threshold u=1.50 $\log(m^3/s)$ (that corresponds to a probability of 0.8 in the standard Gumbel scale) was chosen. Once the model is fitted, the conditional probability of a set of rivers exceeding a certain value at the same time point can be estimated. The estimated conditional probabilities seem to reflect the different catchment characteristics and dominant precipitation generating events. The model also allows inclusion of a time lag, so that flow levels in the rivers included in the model are allowed to exceed a certain value either at the same time point or up to three days apart. The conditional probabilities for the lagged model increase with respect to the non-lagged one. Adding the time lag might offset catchment size effects and allow for weather systems to move across Scotland.

The results obtained from the first approach suggest that the conditional method may yield improved estimates of extreme events on a regional level, but a full exploration of this approach for the whole of Scotland would be time consuming for two main reasons.

One is that computational issues were found when attempting to fit the model to more than four locations at at time. The second one is that a conditioning river needs to be chosen to fit the model. The number of possible combinations of groups of rivers and conditioning rivers is very large. In the application shown here, the conditioning river was chosen as the largest river within each of the sets of rivers explored. Keef et al. (2009) applied the conditional probability model to investigate spatial dependence in extreme precipitation and river flow across the UK. Each individual location was taken as the conditioning variable in a model that included all the neighbouring locations within a certain distance. In addition to the two measures of spatial risk estimated here, the joint conditional distribution can be used to investigate the dependence structure in the upper tail of the distribution and estimate quantities such as (conditional) return levels. However, this model does not consider the fact that the non-conditioning rivers might include extreme flow values at times when the flow in the conditioning river is not extreme.

The second approach considered is a quantile spatial model. Spatial quantile regression has mostly been applied to temperature or wave height data in the literature. The work presented in this chapter represents a comprehensive/extensive spatial analysis of extreme river flows in Scotland. The fitted 95^{th} quantile surfaces show that the spatial pattern of extreme river flow in Scotland has changed over the period 1996-2005. Even within a year, the seasonal pattern changes across months, in agreement with previous studies that suggest changes in extreme rainfall and river flow to be seasonally dependent, as described in Chapter 1.

The proposed model was developed as a natural extension of the temporal quantile model from Chapter 4, with the difference that now the quantile function is allowed to vary with spatial location. The spatial surface is built using bi-variate B-splines, which can be constructed as the tensor product of individual marginal B-spline basis functions. Penalties to control the smoothness of the surface are included, and smoothing parameter selection is carried out based on the SIC. The model is fitted assuming independence of the observations, and the spatial correlation structure is assessed based on the residuals and incorporated in standard errors estimation. Others methods of spatial analysis of extreme values are available, as summarized in Section 5.1.1. The quantile regression approach was chosen as the main aim was to directly assess trends in extreme

river flow. Further, it has the advantage that inclusion of covariates is straightforward, as is the extension to a space-time model. Inclusion of covariates is important as it can help identify the conditions that lead to extreme river flow.

The quantile spatial model was applied to river flow data coming from 119 gauging stations across Scotland. The data cover the time period from January 1996 to December 2005, and the model was fitted independently for each month and year. B-spline bases with 15 basis functions were used on each marginal variable (Easting and Northing), with smoothing parameters set to 2.63 and 1.05 respectively. These were chosen based on SIC. For comparison purposes, the same number of basis functions and smoothing parameters were used to fit all the models. At each month, a bi-variate smooth surface dependent on longitude and latitude is used to describe the spatial trend. Data for the whole month were used at each location. This was done to increase the number of observations available to fit the spatial model, as 119 observations are not enough for reliably estimating a quantile surface. An alternative would have been using the whole data set, but that would not allow to explore the changes in the spatial trend over time. The resulting fitted spatial trends vary considerably over time, but in general values tend to be higher in the West than in the East of Scotland, and lower during the summer months. However, 119 spatial locations might not be enough to fully capture the spatial structure of extreme river in Scotland.

Spatial dependence was assessed based on the residuals and a Gaussian correlation function was found to be appropriate in most cases. The model assumes the correlation to be isotropic, however, in some cases the empirical and fitted variogram differed suggesting that the assumption of isotropy might not be always suitable. This should be further investigated. The parameters of the correlation function varied across months, suggesting that pooling the specific month estimates may not be appropriate and hence standard errors for each month were calculated using their own correlation function.

The results from the spatial analysis indicate changes in the spatial trend over time, suggesting that a full spatio-temporal model is needed to appropriately model the data. To the author's knowledge, quantile spatio-temporal modelling has only been done in a

Bayesian context (Reich et al. (2011); Reich (2012)). The extension of the spatial quantile approach proposed here to build a fully spatio-temporal model is briefly outlined. By including a time-space interaction in the model, changes in the spatial pattern of extreme river flow can be examined. However, fitting the spatio-temporal model raises a number of issues that need to be addressed. In order to reduce the computational cost, a constant seasonal pattern is set, as opposed to what was done in Chapter 4, where the seasonal pattern was allowed to change from year to year, which in turn implies a large number of parameters needed. Additional terms, such as an interaction between year and day of the year can be included to account for changes in the seasonal pattern over the years, but this increases the number of model parameters and hence the computational cost. Memory storage was found to be limiting when trying to fit the model, and the smoothing matrix could not be stored, which makes standard error estimation not possible. The model outlined here is only preliminary and opens up an area of future research.

Chapter 6

Discussion and Main Conclusions

This thesis presents and discusses a number of approaches for modelling extreme river flow values over time and space and their application to a set of Scottish rivers. Overall, the results obtained provide a further insight into the trends and variability in extreme river flows in Scotland, assessing how these have changed over time and space. Many previous studies have avoided dealing with non-stationarity by restricting the time period to winter months. The methods included in this thesis directly address the seasonality of the data and the results obtained suggest departure from the expected seasonal conditions. A wide variety of methods for analyzing extreme values are available, making it difficult to choose the most appropriate one. Those included in this thesis have been selected to address different aspects of the data, namely the variability, temporal trend, seasonality and spatial dependence. Each of the methods can be seen as complementary to the others for a deeper understanding of extreme value behaviour. Wavelet analysis provides a complete and detailed picture of the sources of variability in a time series as well as changes over time and scale. Quantile regression allows direct modelling of the trend and seasonality of extreme river flow values over time and space, gaining information about extreme behaviour or river flow rather than about the mean.

Data were logged to stabilize the variability and improve normality and all the results presented in the thesis are on the logged scale. Data were considered in a slightly different time and scale resolution through the thesis depending on the methodology applied. River hydrology reflects a range of scales of processes, with catchment size exerting

some control over river flow behaviour. In general, four scales of hydrological variability can be recognised: (1) event-scale, which may be a few hours in a small catchment up to a few days in the largest catchments; (2) weather system scale, which is usually between one day and one week; (3) seasonal; and, (4) annual. Analysis of trends over each of these scales depends on filtering or averaging the data at appropriate scales.

Chapter 2 is mainly exploratory and for that the daily data were used. Even though the main aim is to study extreme events, it was considered that understanding the average behaviour of river flow was worthwhile before fully exploring extreme values. Also, it provides a reference to which trends in extreme values can be compared. Monthly maxima was preferred to daily data for studying river flow variability using wavelet analysis. This is due to two main reason. First, to directly study variability in 'extreme' values, and second, for computational reasons, as it was not possible to apply the continuous wavelet transform to daily data due to data volume. Monthly maxima was preferred to annual maxima as the latter does not have enough temporal resolution for a meaningful wavelet analysis and it would not have allowed to explore the seasonality. A POT approach was not possible either as it would result in irregular spacing in time, making wavelet calculations not possible. Daily data were used for fitting the 95% quantile regression models, both in time and space. The daily resolution was chosen here as a large number of observations are needed for reliably estimating the parameters of a quantile model, especially when the aim is to estimate very extreme quantiles. For the spatial model, a way of standardizing the data to account for catchment sizes differences, directly related to average flow levels, was needed. The conditional probability model approach proposed by Keef et al. (2009) directly standardizes the data by marginally transforming each individual series to follow a standard Gumbel distribution and then uses observations above a certain threshold to fit the model. Hence the daily data were used to fit this model. On the other hand, the quantile spatial model was also fitted to the daily data, but prior to fitting the model each individual series was standardized to the mean to make quantile levels across locations comparable. This way extreme events are compared adjusting for the differences in the distribution at each site.

6.1 Summary of results and main findings

6.1.1 Wavelet Analysis

The river flow series studied were characterized by non-stationarity, mainly due to changes in the variability of the seasonal cycle. Wavelet analysis is presented here as a powerful tool for exploring the variability and cross-correlation of non-stationary time series. Differences between the East and the West of Scotland have been identified in river flow maxima over the last 40 years, both in the long term trend and seasonal pattern (see Figure 3.9 and Figure 3.10). The latter appears to be subject to great variability, being slightly higher in the East than in the West, with a common time point around 1986 for all the rivers, when the variability is minimum. The difference in variability between the East and the West can be explained by weather conditions being more stable in the West than in the East, as described in Chapter 2. Periods of greater seasonal variability (and hence a more pronounced seasonal pattern) coincide with flood rich periods. In an effort to investigate how current climate change might impact on the occurrence of extreme events, relationships between NAO and AMO and river flow maxima were explored (see Figures 3.11-3.14). Relationships between NAO and river flow have been reported in previous studies (Shorthouse and Arnell (1997); Labat (2010); Kingston et al. (2009); Bouwer et al. (2008)) while literature relating AMO to river flow or rainfall is less abundant. The strength of the correlation between river maxima and these two indices varies temporally, appearing to be stronger, specially for the AMO, on the annual river flow cycle in the second half of the record (from about 1987 onwards). The influence of AMO and NAO not only changes over time but also the phase difference, which relates to the lag between events in both series, is highly variable, and is different depending on location (East/West) but also on the size of the catchment. In particular, NAO's influence on rivers Ewe and Ness is particularly strong, which is explained by their geographical location, being more affected by changes in pressure systems. Both the NAO and the AMO seem to have little or no significant influence in the River Lossie. Peak flows in the Lossie catchment are mostly associated with frontal storms of long duration (Cameron (2006)).

Trends identified from monthly maxima series using wavelet analysis resemble the trends that were obtained on the daily series using the stl decomposition (see Figure 2.15). While the overall pattern is more or less the same, there are some subtle differences that might seems small but that might actually make a difference when talking about extreme events. In particular, Rivers Ewe and Ness show similar patterns in both their long term trends and annual variability changes based on the monthly maxima series, although the variability is slightly greater in the Ewe. These two rivers had already been identified to be very similar in Chapter 2 for being exposed to similar weather patterns. The greater variability in the Ewe is probably reflecting a catchment effect; the Ewe has a smaller catchment size and the influence of lakes (present in both the Ewe and Ness catchments, see Chapter 2) is lesser than on the Ness.

Modelling river flow data and in particular extreme river flow values presents a number of challenges. The first issue identified was the correlation structure, better characterized by a long-memory or long-range dependence model rather than the commonly used ARMA models. It is important to identify and account for long-range dependence as this is a correlation structure that causes observations far apart in time to be still significantly correlated. This has an effect on uncertainty estimates, making confidence intervals wider than if the dependence was short-range. Wavelet analysis was used to estimate the Hurst parameter that characterizes long-range dependence. For the eight Scottish rivers investigated, the estimated Hurst parameter ranges between 0.53 and 0.72 depending on the river (see Table 2.5).

A number of decisions had to be made when applying wavelet analysis to the river flow data. In the discrete case, the maximum overlap discrete wavelet transform was preferred to the discrete wavelet transform to avoid the sample size limitations. Even though the maximum overlap discrete wavelet transform is not an orthogonal transformation, this was not a concern as the aim of the analysis was mainly descriptive. For other purposes the discrete wavelet transform may be more suitable. The filter used (LA(8)) was chosen so that events in the original time series and in the wavelet decomposition could be aligned in time. Smaller and greater values of the filter width either introduced artifacts in the data or did not show any difference in the decomposition, respectively. The level of decomposition was set to four as further decomposition did not prove useful. In the

continuous case, the Morlet wavelet was chosen to perform the analysis for being the one recommended in the literature. The effects of using a different wavelet functions were not explored. Finally, when the wavelet method is used for estimating the Hurst parameter, the level of decomposition chosen was justified based on a physical explanation. The first two scales, corresponding to 1 and 2 days, were not included in the estimation as they can be related to short memory behaviour rather than long range dependence (P. Craigmile, personal communication).

6.1.2 Quantile regression

Quantile regression is used here as an alternative way of modelling extreme events. In particular, it is useful when the main interest is to directly model the temporal changes in extreme value behaviour. The quantile temporal regression model proposed in Chapter 4 allows the trend and seasonal components of extreme river flow to be identified. The model was applied to eight Scottish rivers. The results suggest some differences in the long term trend between the East and the West, with increasing and decreasing periods being almost opposite in time, and the seasonality seems to be more variable in the East than in the West (see Figure 4.4 and Figure 4.5). The identified trends differ to some extent from the trends identified for mean value in the previous chapters. Even though the differences are small, they can have a significant impact in flood risk assessment. Quantitatively, the trends estimated for the $95th$ quantile are closer to extreme values than those estimated for the mean. Qualitatively, the time at which the trend peaks can have a significant impact on flooding; even it the difference in peak time between the average trend and the quantile based trend is small, it can mean, for example, that the peak happens during the autumn in one case and during the winter in the other case.

Most of the studies found in the literature overcome the seasonality issue by just looking at a portion of the data, usually over the winter months, as it is the period where most extreme values are expected to take place. Black and Werritty (1997) highlight the importance of investigating flood seasonality as it is not constant across Scotland. Based on data covering the period 1959-1988 and from 156 gauging stations, Black and

Werritty (1997) found four different flood seasonal regimes, where floods were considered as events in peak-over-threshold series. The southwest of Scotland has the earliest seasonality with peaks usually occurring in August-November, while the northeast of Scotland and the Lothian region have a weak seasonality with peaks more spread over the year. The East coast, in particular the river Tweed, has a late seasonality, with peaks concentrated in February-March. Black and Werritty (1997) also found a catchment size effect, with most small catchments dominated by early flood seasonality while large inland catchments tend to have a winter flood seasonality.

The individual fitted models can help understand the hydrological processes related to extreme value occurrence in the individual catchments. By accounting for the seasonality, it is possible to look at extreme values independently of the season during which they occur. This is important because despite floods mainly happening during the winter months, recent years have seen a rising number of summer related floods. By setting a 95% quantile regression fit as the threshold and identifying all observations over the fitted model as extremes, the probability of any of the excesses happening is constant and equal to 0.05, independently of when in the year. That means that we are able to take into account events that in the current series happened during the summer, which might be disregarded using a constant POT threshold approach. It can be argued that the events localized in the summer months could as well happen during the winter given the current climate change, as their probability of occurrence is the same, regardless of the season.

Choosing an appropriate threshold is important as it might have an impact on posterior modelling. For some of the rivers, choosing a suitable constant threshold for fitting a GP distribution in Chapter 2 was particularly difficult as the parameters estimates were highly variable dependending on the chosen threshold. A number of authors have questioned whether a constant threshold should be used for determining peak-over-threshold series, especially when the original series is clearly nonstationary. Instead, using a time varying threshold has been suggested (Northrop and Jonathan (2011); Kyselý et al. (2010)). The temporal quantile model proposed in this thesis can be used to aid with threshold selection.

The PIRLS approach used for fitting the parameters in a quantile regression model provides a strong alternative to linear programming methods, as the model can be easily used in combination with any nonparametric smoothing method available and extended to analyze high dimensional data. This makes it a very flexible and easy to implement approach. Also, because it is a regression-like model, it is more intuitive than other extreme value models and inclusion of covariates is straightforward. The two additive components of the model were estimated using the back-fitting algorithm. This was main chosen for computational reasons, as the full design matrix was too large given the large number of basis functions used for both components. Time in a continuous scale was chosen as the explanatory variable for both the trend and seasonality, with a periodic constraint set on the latter to ensure continuity at the beginning and end of each year. This is a very flexible model that allows the seasonal pattern to change over the years, and was motivated by the results obtained from the wavelet analysis that suggested a non-constant annual cycle.

6.1.3 Spatial analysis

A number of methods are available in the literature for the analysis of extreme values in a spatial context. Here two different approaches were applied to extreme river flow data from Scotland. The first approach investigated, based on a conditional probability model, extends extreme value theory to a multivariate setting and allows fitting of an extreme value model for a number of rivers at the same time, for which a conditioning river needs to be decided in advance. The model performed well when applied to different sets of Scottish rivers, as the results suggested a dependence structure that can be explained based on the physical and geographical characteristics or the rivers. The conditioning river was chosen as the largest river within each of the sets of rivers explored. This model can be very useful for assessing flood risk on a regional level. However, it can be limiting for studying the dependence structure between extreme flows in the whole of Scotland. Even if the model could be fitted for all 119 gauging stations at a time, a conditioning river still needs to be chosen, resulting in a large number of possible models.

The second approach investigates Scotland as a whole by means of a spatial quantile regression model. This model is built as an extension of the temporal quantile model.

As the model is set in a regression context, covariates can be easily incorporated. This is essential to investigate the main drivers of extreme river flow in Scotland, and may prove useful in studies aiming to investigate the effects of climate change. Even though the model here only includes spatial location as a covariate, it can be used to assess which factors might be more influential on the occurrence of extreme events. The model was applied to 119 gauging stations from Scotland. The resulting fitted spatial trends vary considerably over time, but in general values tend to be higher in the West than in the East of Scotland, and lower during the summer months (see Figures 5.11-5.20). The fitted quantile spatial model might prove useful for planning purposes or decision making. While the results cannot be directly interpreted in terms of return periods as there is no underlying probabilistic model to reliably extrapolate, they provide detailed information on extreme flow values over space associated with a fixed probability. Hence they can be used as a first tool for assessing which areas might be more vulnerable than others.

6.2 Limitations

The methods and analysis presented in this thesis have a number of limitations. The main ones are detailed below.

The wavelet analysis concentrated on the trend and seasonal components. The remaining components from the multiresolution analysis were not investigated in detail. Traditional time series ARMA(p,q) models could not satisfactorily reflect the complex correlation structure of the data. Instead, a long memory correlation structure was preferred. The wavelet method was used to estimate the Hurst parameter that characterizes the long-range dependence of the data. This was preferred to other methods available in the literature as it has been shown to be the most robust. However, the optimal way of estimating the Hurst parameter would include a measure of variability to assess whether H is significantly greater than 0.5 (P. Craigmile (oral communication)). Probability based methods using approximate likelihood are available but these have not been explored here (Beran (1994); Taqqu et al. (1995)).

The approach taken to deal with the residuals from a quantile regression, either in time or space might be questionable. Common practice is to use bootstrap methods to calculated confidence intervals. However, this was not computationally feasible given the iterative nature of the fitting process. Errors in both cases were assumed to be independent, to then incorporate the dependence structure in standard error estimation. However, no distributional assumption was made on the errors. This is common practice in quantile regression and part of the reason for its great flexibility. Here, a normal distribution was assumed to calculate approximated pointwise confidence bands in the temporal case and standard errors in the spatial case. This choice should be investigated further.

In particular, the quantile regression model proposed in Chapter 4 has some limitations. It can only be fitted for one quantile at a time, while the general fitting procedure using linear programming methods (and implemented in R in the `quantreg` package) allows for multiple quantiles to be fitted simultaneously. Computationally, fitting the model was very costly, especially for long time series. This can probably be improved with the help of R packages now available to deal with large matrices in an smart and efficient way. No distributional assumption was made, which means the model cannot be used to extrapolate and hence derive return level estimates.

Smoothing parameter selection for the temporal quantile model was based on personal judgement after visual inspection of a range of fitted models with various degrees of freedom. Automatic selection procedures were not feasible given the high computational cost of the model, neither was a detailed exploration of the effect of the smoothing parameters on the fitted model, as only a few values were tried.

None of the spatial methods investigated here consider the case of having more than one gauging station per river. Including connectedness of gauging stations within a river network needs to be explored, as this will allow spatial scale effects to be directly estimated. The results from the spatial quantile model represent a marginal analysis. Ideally, a full spatio-temporal model should be used. In order to have enough data to fit the spatial model for a particular month and year, all the values in that month were

used at every spatial location. The temporal correlation within each month was neither investigated nor accounted for in the model. On the other hand, the theoretical model used for the spatial correlation did not provide a perfect fit for some of the months and years, and indications of a possible anisotropy were found in some of the plots.

6.2.1 Data limitations

A large data set, both spatially and temporally, was initially available to illustrate the application of the different methods. However, the length of the individual river flow series was highly variable and when all records were combined to a common length in time the resulting series only cover about 10 years. A longer time period would be desirable to appropriately assess trends in extreme river flow. Given the large number of spatial locations (119) it was not possible to present results for every individual river flow series. A subset of eight rivers were selected in Chapter 2 and each of the methods illustrated for the same eight rivers. These rivers were chosen to include a range of different geographical locations and catchment sizes so that differences due to these two factors could be explored. A different selection of rivers might have led to slightly different conclusions. Finally, there are a number of rivers whose flow is affected to some degree by the presence of hydro-power stations, reservoirs or public water abstractions. These have not been taken into account. Finally, the work presented concentrates on fluvial risk and does not take into account how the extreme events were generated. Most of them will be related to extreme rainfall events, but some of them might be exacerbated by snowmelt.

6.3 Future Work

The work presented in this thesis leaves scope for further developments. In the wavelet analysis, it would be interesting to re-sample the data to get maximum flows in every 1/16 of a year, i.e. every 23 days instead of every month, so that component D_3 in the wavelet decomposition corresponds exactly to variations over three months. The conditional probability model for studying the spatial dependence in extreme river flow was fitted conditioning on the largest river within each area. It would be interesting to investigate how much the results vary when the smallest river is used as the conditioning

one instead. The quantile regression models only include time and space as explanatory variables. Incorporation of other covariates, such as rainfall data or catchment characteristics is a possible extension that could be easily achieved because of the way the model is built. Another interesting area of future research could be fitting the quantile temporal model to all 119 river flow series. Functional clustering analysis methods could then be used to identify clusters of rivers with similar trend and seasonal patterns. Also, the performance of fitting extreme value models to the excesses determined using the fitted models as a time-varying threshold should be investigated.

In particular, the main area of future research is the development of a fully spatio-temporal quantile model. A brief outline of how to extend the model is included in Chapter 5. However, the spatio-temporal model raises a number of statistical challenges. In particular, three points to work on have been identified. The first one is that while the spatio-temporal model with interactions as it stands can be fitted using the PIRLS algorithm, it is not possible to calculate the hat or smoothing matrix due to memory storage limits. This step is vital for the inference process, as the hat matrix is needed for estimating confidence intervals. Sparse matrix algorithms need to be used to decrease storage requirements. The second point is that the model is fitted assuming independent observations. Appropriate time-space correlation structures for extreme values need to be investigated and incorporated in the modelling process. Finally, as the number of additive components in the model increases, so does the number of smoothing parameters to be chosen. Automatic selection procedures such as cross-validation or BIC might not be computationally feasible when the number of smoothing parameters is large. Ways of efficiently selecting smoothing parameters need to be investigated further. The way the model is built, however, means that it can benefit from advances made in nonparametric smoothing, a current area of active research.

The ultimate aim of analyzing extreme river flow values is to provide flood risk estimates. However, the relationship between extreme river flow and flood is complex as flooding duration and depth depends on river channel and floodplain geometry. Further, the economic and social impacts of flooding depend on land use, existing adaptation strategies and interactions between natural and artificial drainage networks. Further research is needed so that the information obtained from the different methods proposed here can

be used to improve flood risk estimation.

Finally, the methods presented here have been applied to river flow data, but they are of general application and can easily be transferred to any other environmental areas where extremes are of interest such as air pollution or temperature.

Appendix A

Appendix A

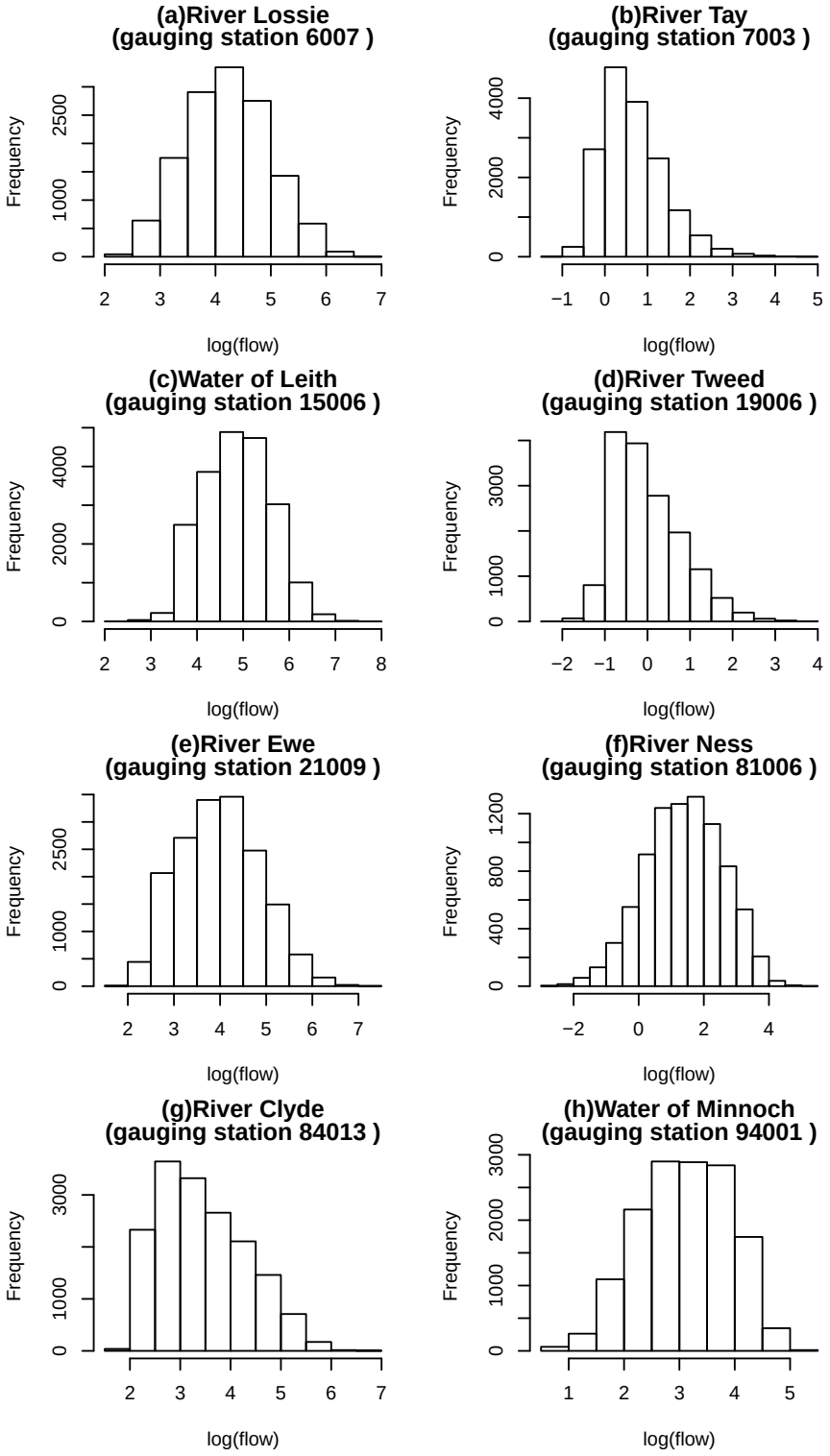

FIGURE A.1: Frequency distributions. Units are in $\log(m^3/s)$

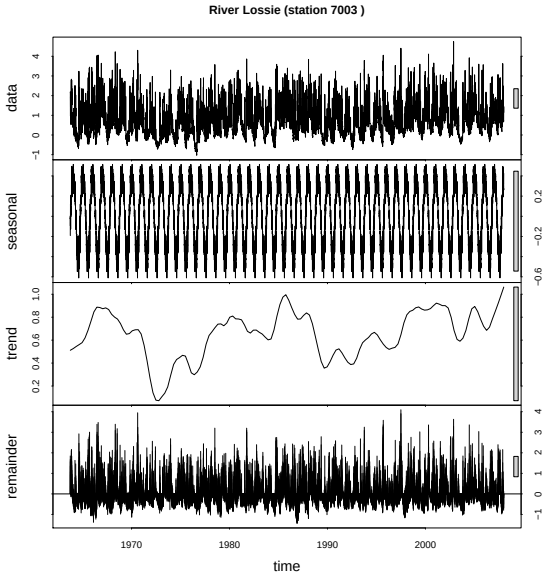

FIGURE A.2: STL decomposition of daily river flow ($\log(m^3/s)$). River Lossie (Station 7003)

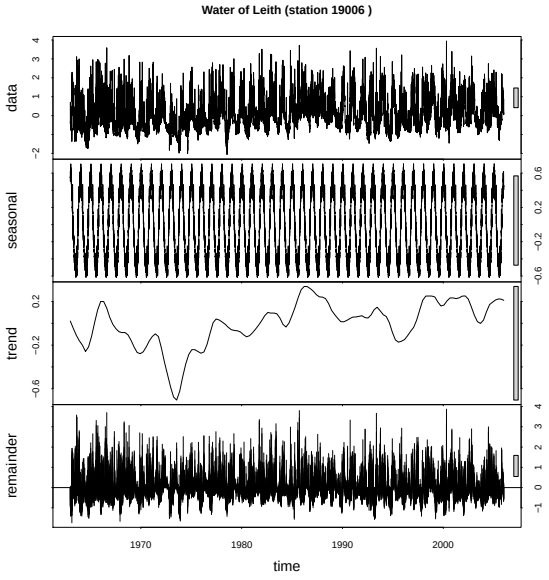

FIGURE A.3: STL decomposition of daily river flow ($\log(m^3/s)$). Water of Leith (Station 19006)

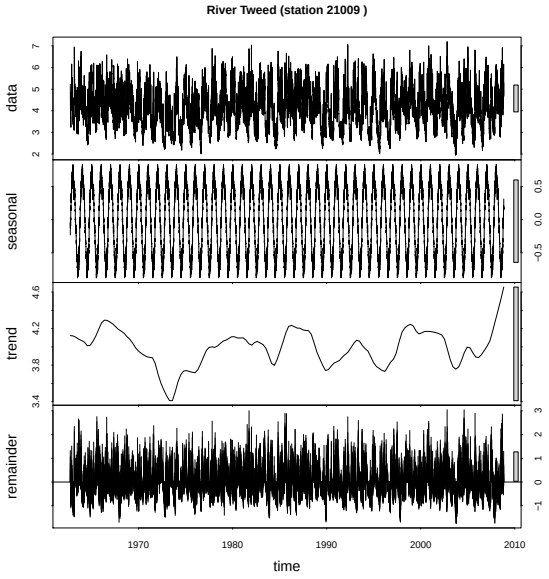

FIGURE A.4: STL decomposition of daily river flow ($\log(m^3/s)$). River Tweed (Station 21009)

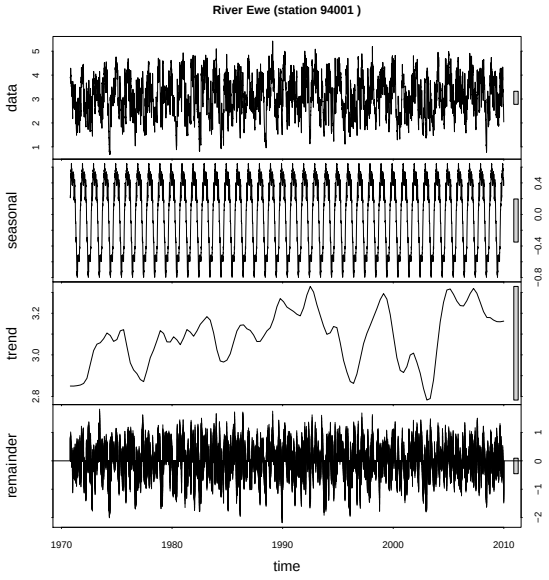

FIGURE A.5: STL decomposition of daily river flow ($\log(m^3/s)$). River Ewe (Station 94001)

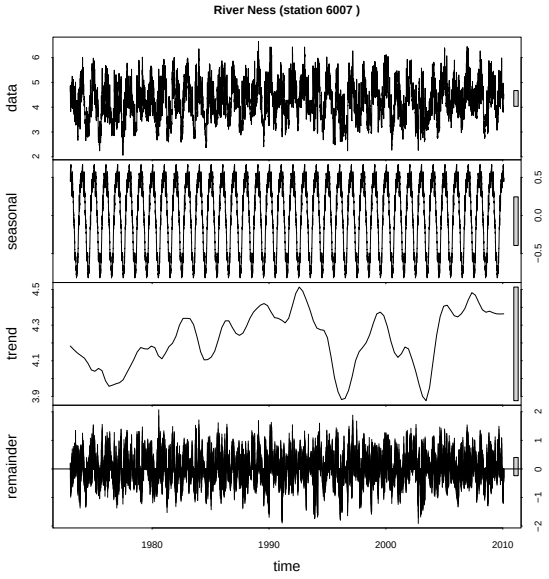

FIGURE A.6: STL decomposition of daily river flow ($\log(m^3/s)$). River Ness (Station 6007)

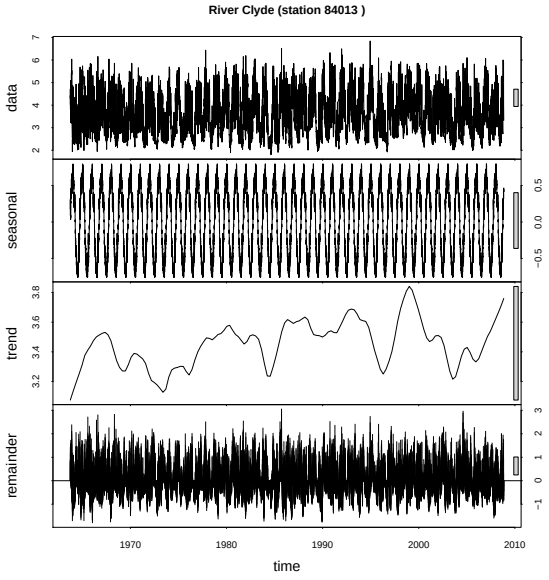

FIGURE A.7: STL decomposition of daily river flow ($\log(m^3/s)$). River Clyde (Station 84013)

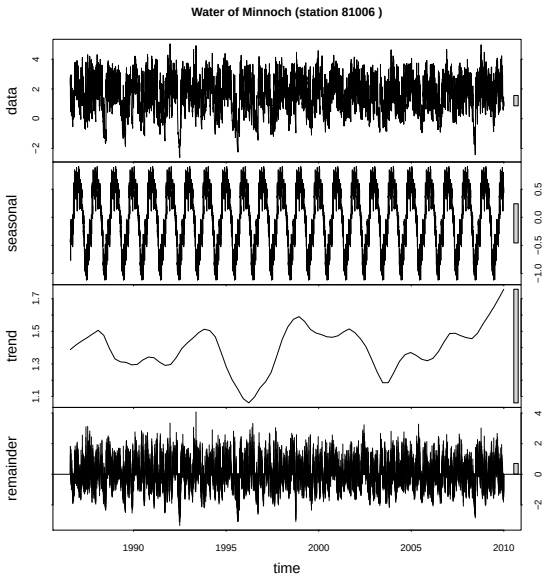

FIGURE A.8: STL decomposition of daily river flow ($\log(m^3/s)$). Water of Minnoch (Station 81006)

Station	River	Catchment Area (km^2)	Time Period (years)	Easting	Northing
1001	Wick	161.9	1995-2009	326.2	954.9
2001	Helmsdale	551.4	1975-2009	299.7	918.1
2002	Brora	434.4	1993-2009	289.2	903.9
3002	Carron	241.1	1974-2009	249.0	892.1
3003	Oykel	330.7	1977-2009	240.3	900.1
3004	Cassley	187.5	1979-2009	247.2	902.2
3005	Shin	575	1981-2009	257.4	897.4
4001	Conon	961.8	1947-2009	248.2	854.7
4004	Blackwater	336.7	1981-2009	245.5	856.3
4005	Meig	120.5	1986-2009	228.6	852.8
4006	Bran	116.1	1989-2009	220.5	860.2
5002	Farrar	311.3	1986-2009	239.0	840.5
5003	Glass	481.8	1988-2009	235.4	832.1
6007	Ness	1839.1	1973-2008	264.5	842.7
6008	Enrick	105.9	1979-2009	245.0	830.0
7002	Findhorn	781.9	1958-2006	301.8	858.4
7003	Lossie	216	1963-2007	319.4	862.6
7004	Nairn	313	1979-2009	288.2	855.1
7005	Divie	165	1977-2009	300.5	848.0
8004	Avon	542.8	1952-2006	318.6	835.2
8005	Spey	1267.8	1951-2006	294.6	819.0
8009	Dulnain	272.2	1952-2006	297.7	824.7
9002	Deveron	954.9	1960-2009	370.5	849.8
9003	Isla	176.1	1969-2009	349.4	850.6
9004	Bogie	179	1980-2009	351.9	837.3
9005	All Deveron	67	1948-2009	337.8	829.1
10002	Ugie	325	1971-2009	410.1	848.5
10003	Ythan	523	1983-2009	394.7	830.3
11001	Don	1273	1969-2006	388.7	814.2
11002	Don	787	1969-2009	375.6	820.1
11003	Don	499	1973-2009	356.6	817.0
11004	Urie	198	1984-2009	372.1	826.0
12001	Dee	1370	1929-2008	363.5	795.6
12003	Dee	690	1975-2009	334.4	796.5
12005	Muick	110	1976-2009	336.4	794.7
12006	Gairn	150	1978-2009	335.3	797.1
12008	Feugh	229	1985-2009	368.7	792.8
13001	Bervie	123	1979-2006	382.6	773.4
13004	Prosen	104	1985-2009	339.6	758.6
13005	Lunan	124	1981-2009	365.5	749.4
13008	South Eask	488	1983-2009	360.0	759.6
13009	West Water	127.2	1985-2006	359.2	768.0
14001	Eden	307.4	1967-2009	341.5	715.8
14002	Dighty	126.9	1969-2009	347.7	732.4
14005	Motray Water	60	1984-2009	344.1	722.4
15003	Tay	3210	1947-2006	308.2	739.5
15006	Tay	4857.1	1952-2008	314.7	763.6
15008	Dean Water	177.1	1958-2008	334.0	747.9

Station	River	Catchment Area (km^2)	Time Period (years)	Easting	Northing
15011	Lyon	391.1	1958-2009	278.6	748.6
15012	Tummel	1670	1973-2006	294.7	757.4
15015	Almond	84	1986-2009	288.8	731.6
15023	Braan	210	1983-2009	301.4	742.2
15025	Ericht	432	1985-2009	317.4	747.2
16003	Ruchill Water	99.5	1970-2009	276.4	720.4
16004	Earn	782.2	1972-2006	304.4	718.2
17001	Carron	122.3	1969-2009	283.2	682.0
17004	Ore	162	1972-2009	333.0	699.7
17012	Red Burn	22	1986-2009	278.8	678.0
17015	North Queich	23.1	1987-2009	311.4	704.2
17016	Lochty Burn	14	1986-2009	322.0	698.5
18001	Allan Water	161	1957-2005	279.2	705.3
18008	Leny	190	1973-2009	258.5	709.6
18011	Forth	1036	1981-2008	277.5	695.5
18014	Bannock Burn	23.7	1986-2009	281.2	690.8
19001	Almond	369	1957-2009	316.5	675.2
19006	Water of Leith	107	1963-2005	322.8	673.1
19010	Braid Burn	16.2	1969-2007	327.3	670.7
20001	Tyne	307	1961-2009	348.9	681.1
20002	West Peffer	26.2	1966-2008	339.7	575.1
20007	Giford Water	64	1973-2005	351.1	671.8
21003	Tweed	694	1959-2006	325.7	640.1
21009	Tweed	4390	1962-2008	389.8	647.8
21012	Teviot	323	1963-2006	352.2	615.9
21024	Jed Water	139	1971-2006	365.5	621.4
77002	Esk	495	1962-2009	339.7	575.1
77003	Liddel Water	319	1973-2009	341.5	575.9
77004	Kirtle Water	72	1979-2009	328.5	569.3
78003	Annan	925	1967-2006	319.1	570.4
78004	Kinnel Water	76.1	1963-2009	307.7	586.8
78005	Kinnel Water	229	1979-2009	309.1	584.5
78006	Annan	217	1983-2006	310.0	601.0
79002	Nith	799	1957-2008	292.3	585.1
79004	Scar Water	142	1963-2009	284.5	594.0
79005	Cluden Water	238	1963-2009	292.8	579.5
80001	Urr	199	1963-2009	282.2	561.0
80003	White Laggan Burn	5.7	1980-2009	246.8	578.1
81002	Cree	368	1963-2008	241.2	565.3
81003	Luce	171	1967-2009	218.0	559.9
81004	Bladnoch	334	1977-2009	238.2	554.5
81006	Water of Minnoch	141	1986-2009	236.3	574.6
82001	Girvan	245.5	1963-2009	221.7	599.7
82002	Doon	323.8	1974-2009	233.8	616.0
82003	Stinchar	341	1973-2009	221.1	583.2
83004	Lugar	181	1972-2009	250.8	621.7
83006	Ayr	574	1976-2009	236.1	621.6
83009	Garnock	183.8	1978-2006	230.7	642.3

Station	River	Catchment Area (km^2)	Time Period (years)	Easting	Northing
83010	Irvine	72.8	1977-2006	253.2	673.3
84001	Kelvin	335.1	1948-2007	255.8	670.6
84005	Clyde	1704.2	1958-2008	270.3	657.9
84013	Clyde	1903.1	1963-2008	267.2	661.7
84020	Glazert	51.9	1968-2006	265.6	676.4
85001	Leven	784.3	1963-2009	239.4	680.3
85002	Endrick Water	219.9	1963-2009	248.5	686.6
85003	Falloch	80.3	1970-2006	232.1	719.7
85004	Luss Water	35.3	1976-2009	235.6	692.9
86001	Little Eachaig	30.8	1968-2007	214.3	682.1
89002	Linne nam Beathach	50.5	1981-2009	227.2	742.2
89003	Orchy	251.2	1977-2006	223.9	731.0
89004	Strae	36.2	1978-2009	214.6	729.2
89005	Lochy	47.7	1978-2009	219.7	727.4
90003	Nevis	69.2	1982-2009	211.6	774.2
92001	Shiel	256	1995-2009	166.6	770.2
93001	Carron	137.8	1979-2009	194.2	842.9
94001	Ewe	441.1	1970-2008	185.9	880.3
95001	Inver	137.5	1977-2009	214.7	925.0
96001	Halladale	204.6	1976-2009	289.1	956.1
96002	Naver	477	1977-2009	271.3	956.8
96003	Strathy	111.8	1985-2009	283.6	965.2
97002	Thurso	412.8	1972-2009	313.1	959.5

TABLE A.1: Data set

Appendix B

Appendix B

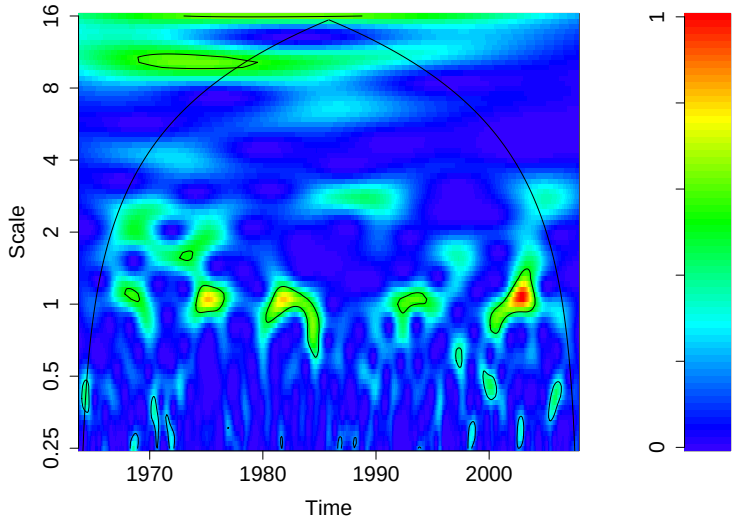

FIGURE B.1: Wavelet power spectrum of monthly maxima series - River Lossie (gauging station 7003). Both time and scale axis are in years. The thick black line highlight areas of significant variability at the 90% level

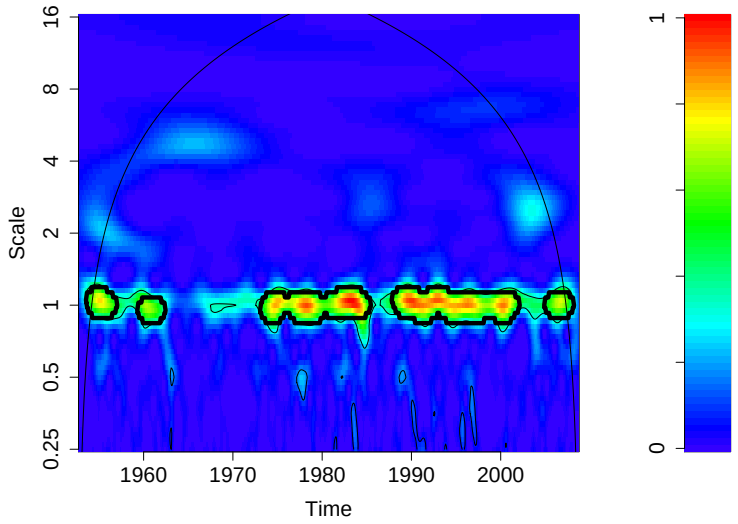

FIGURE B.2: Wavelet power spectrum of monthly maxima series - River Tay (gauging station 15006). Both time and scale axis are in years. The thick black line highlight areas of significant variability at the 90% level

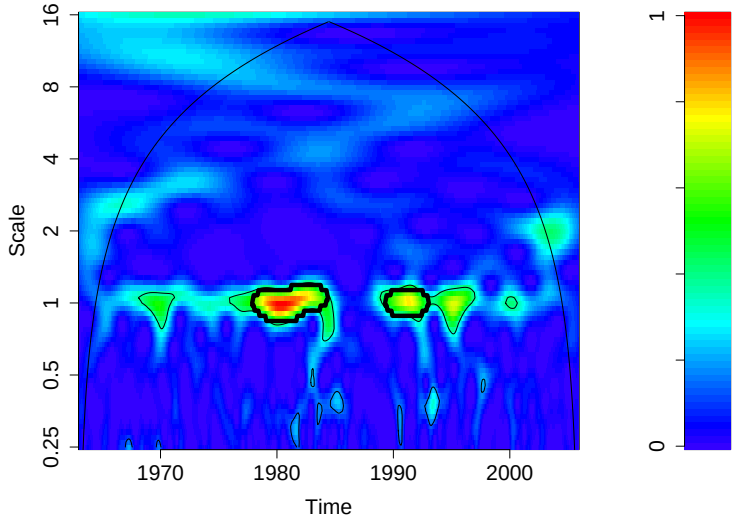

FIGURE B.3: Wavelet power spectrum of monthly maxima series - Water of Leith (gauging station 19006). Both time and scale axis are in years. The thick black line highlight areas of significant variability at the 90% level

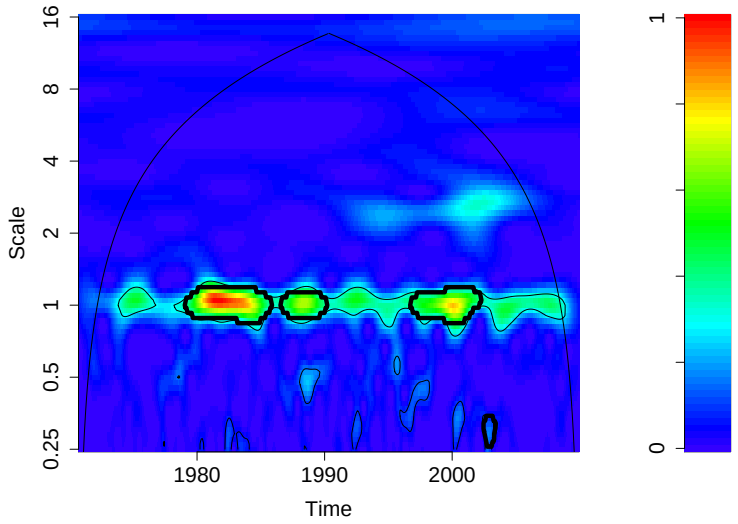

FIGURE B.4: Wavelet power spectrum of monthly maxima series - River Ewe (gauging station 94001). Both time and scale axis are in years. The thick black line highlight areas of significant variability at the 90% level

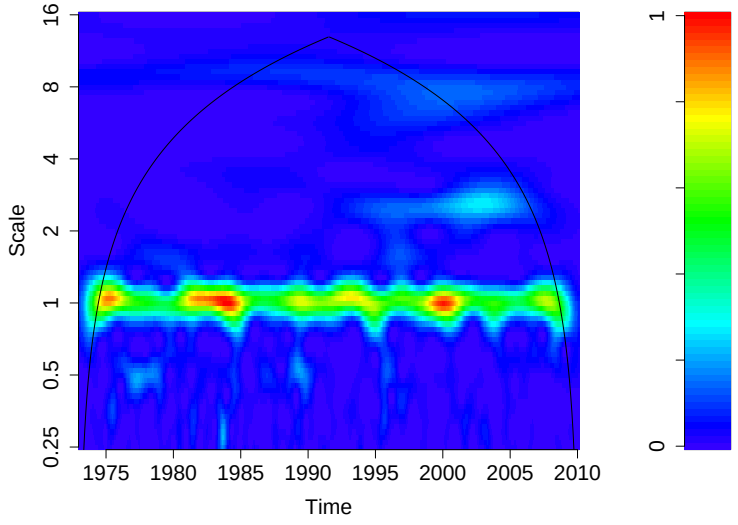

FIGURE B.5: Wavelet power spectrum of monthly maxima series - River Ness (gauging station 6007). Both time and scale axis are in years. The thick black line highlight areas of significant variability at the 90% level

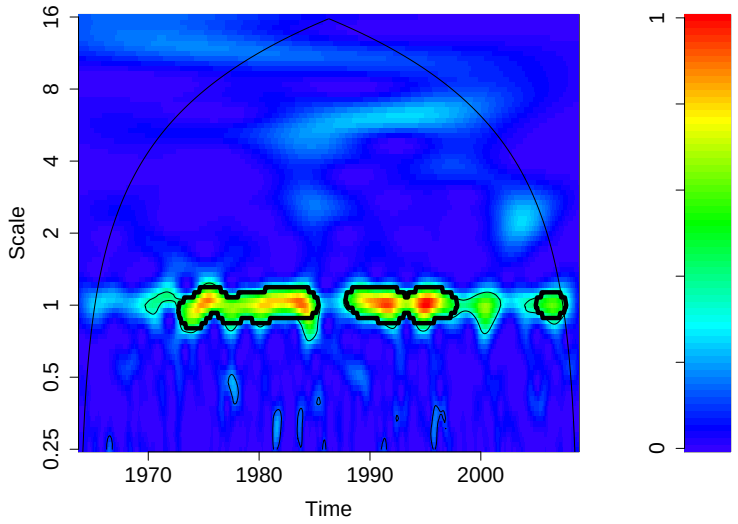

FIGURE B.6: Wavelet power spectrum of monthly maxima series - River Clyde (gauging station 84013). Both time and scale axis are in years. The thick black line highlight areas of significant variability at the 90% level

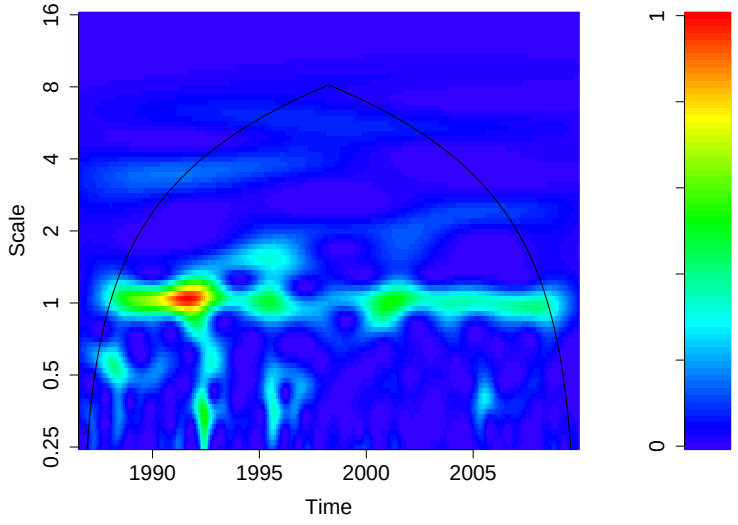

FIGURE B.7: Wavelet power spectrum of monthly maxima series - Water of Minnoch (gauging station 81006). Both time and scale axis are in years. The thick black line highlight areas of significant variability at the 90% level

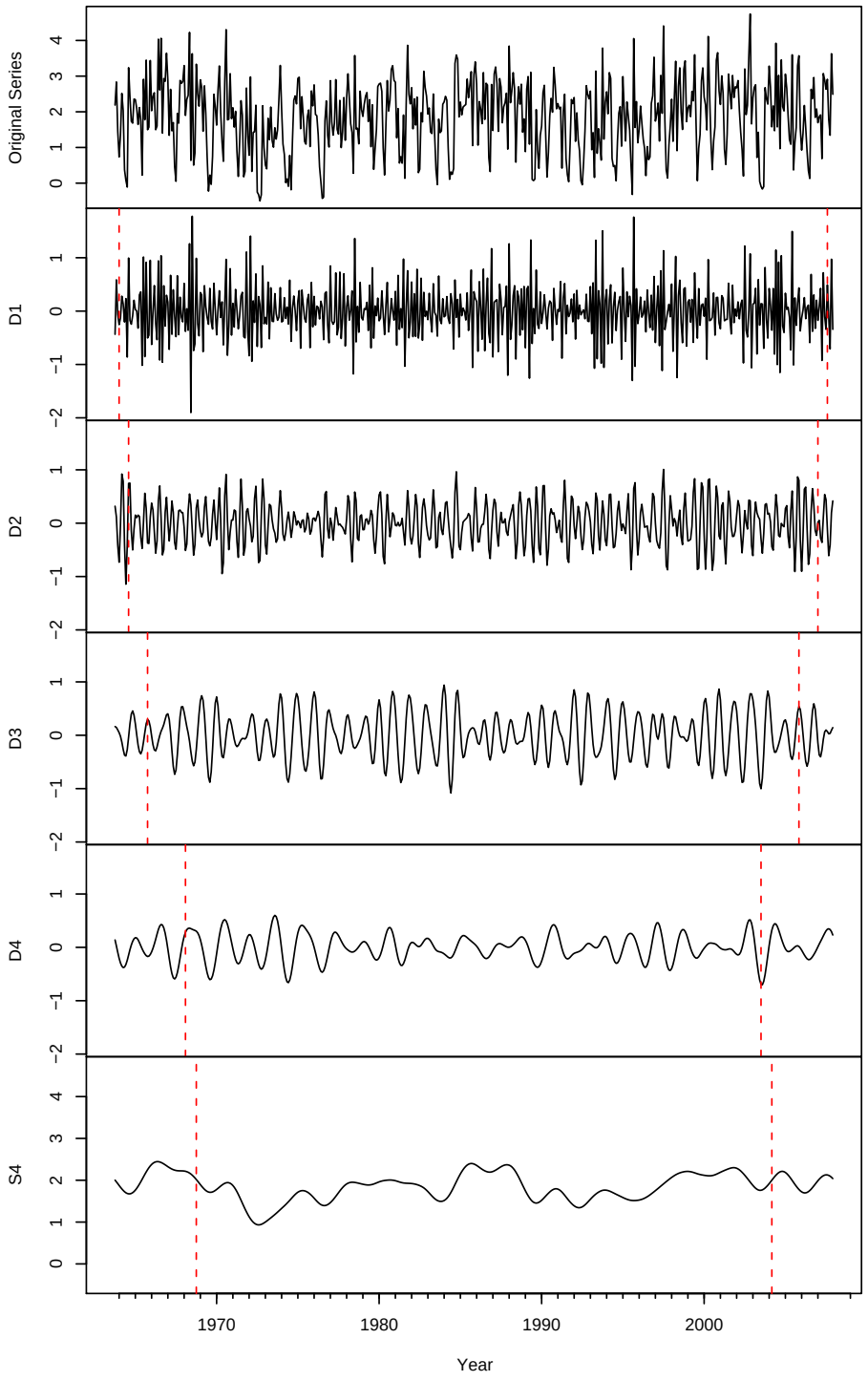

FIGURE B.8: Multiresolution analysis of monthly series - River Lossie (gauging station 7003)

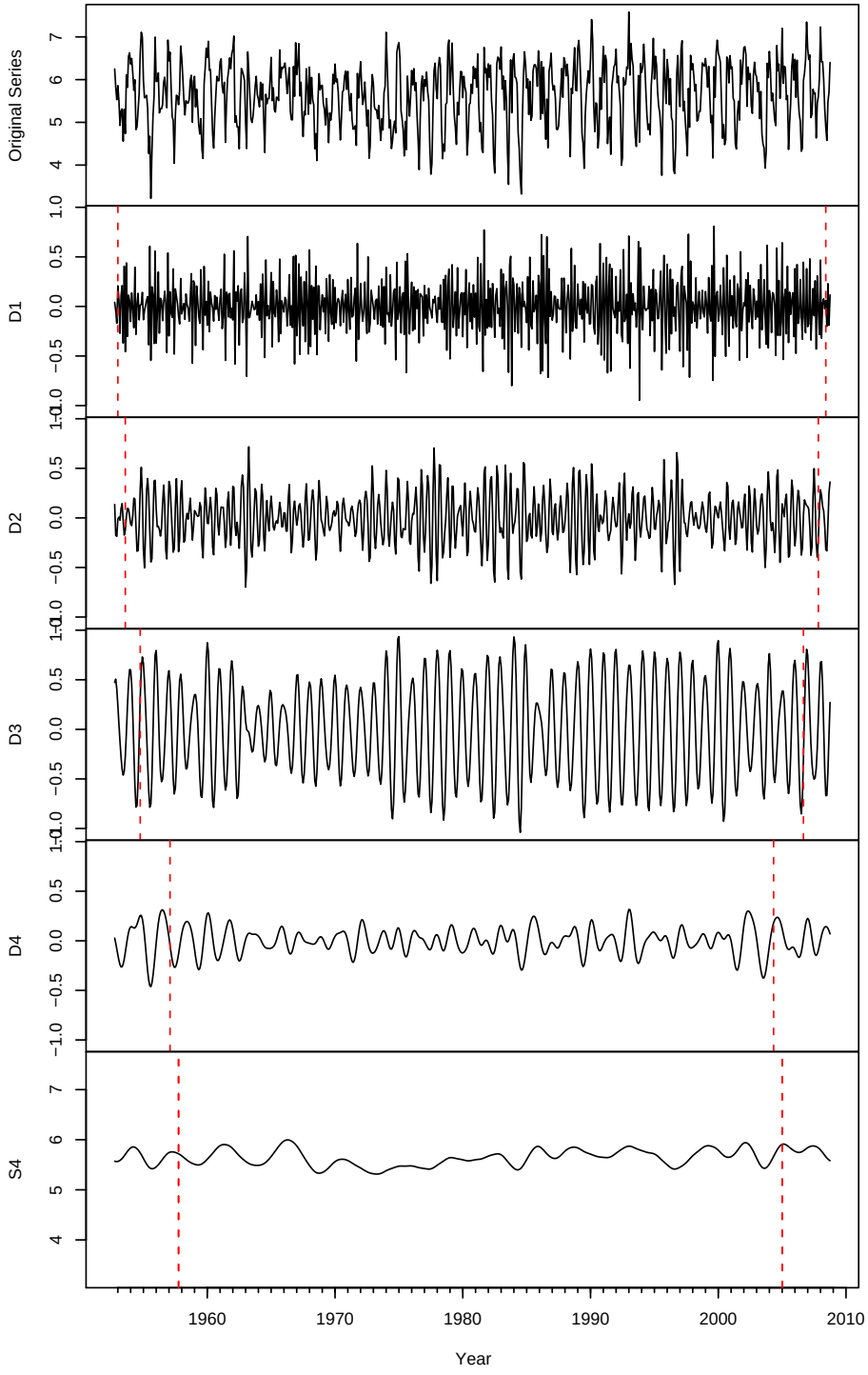

FIGURE B.9: Multiresolution analysis of monthly series - River Tay (gauging station 15006)

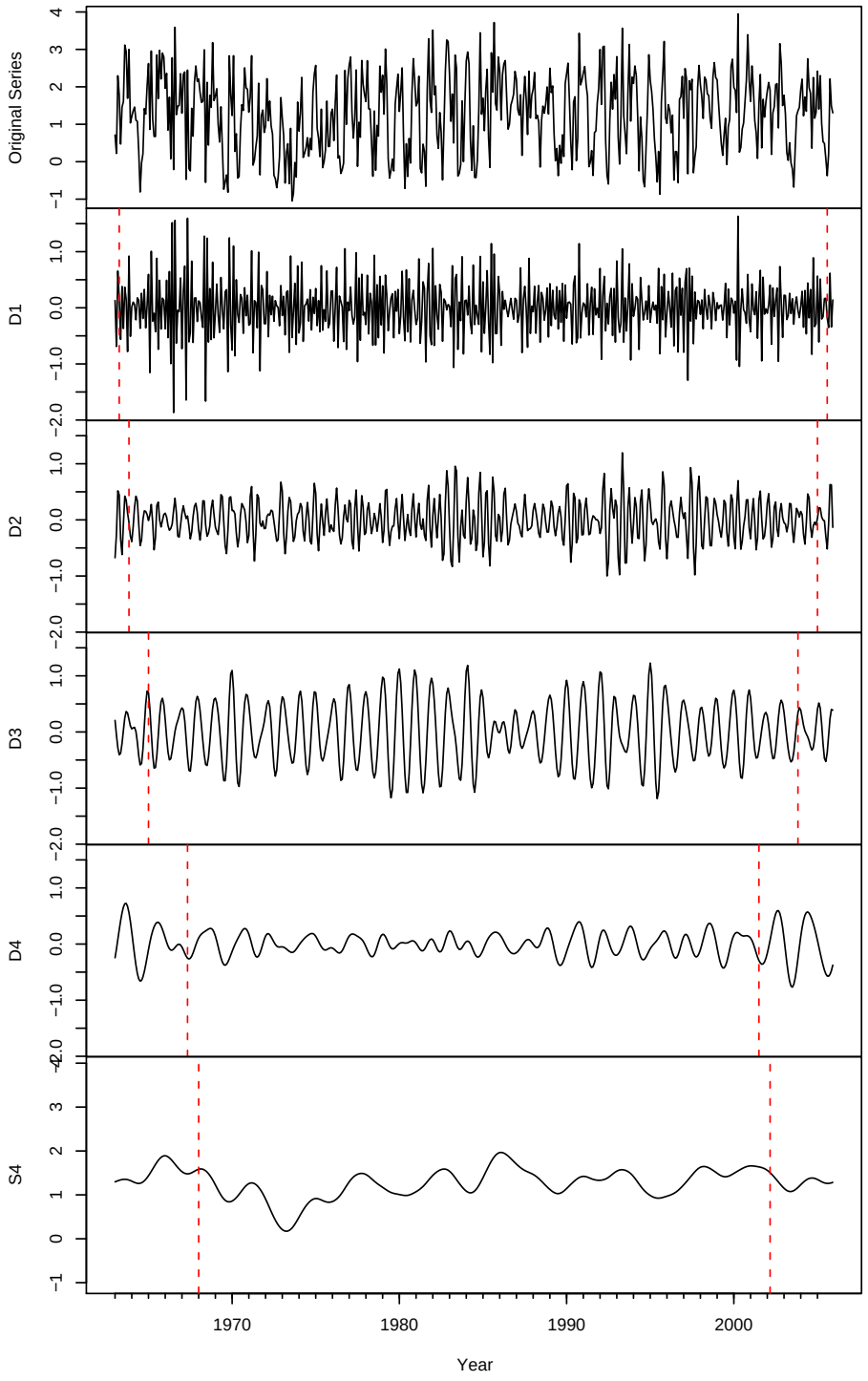

FIGURE B.10: Multiresolution analysis of monthly series - Water of Leith (gauging station 19006)

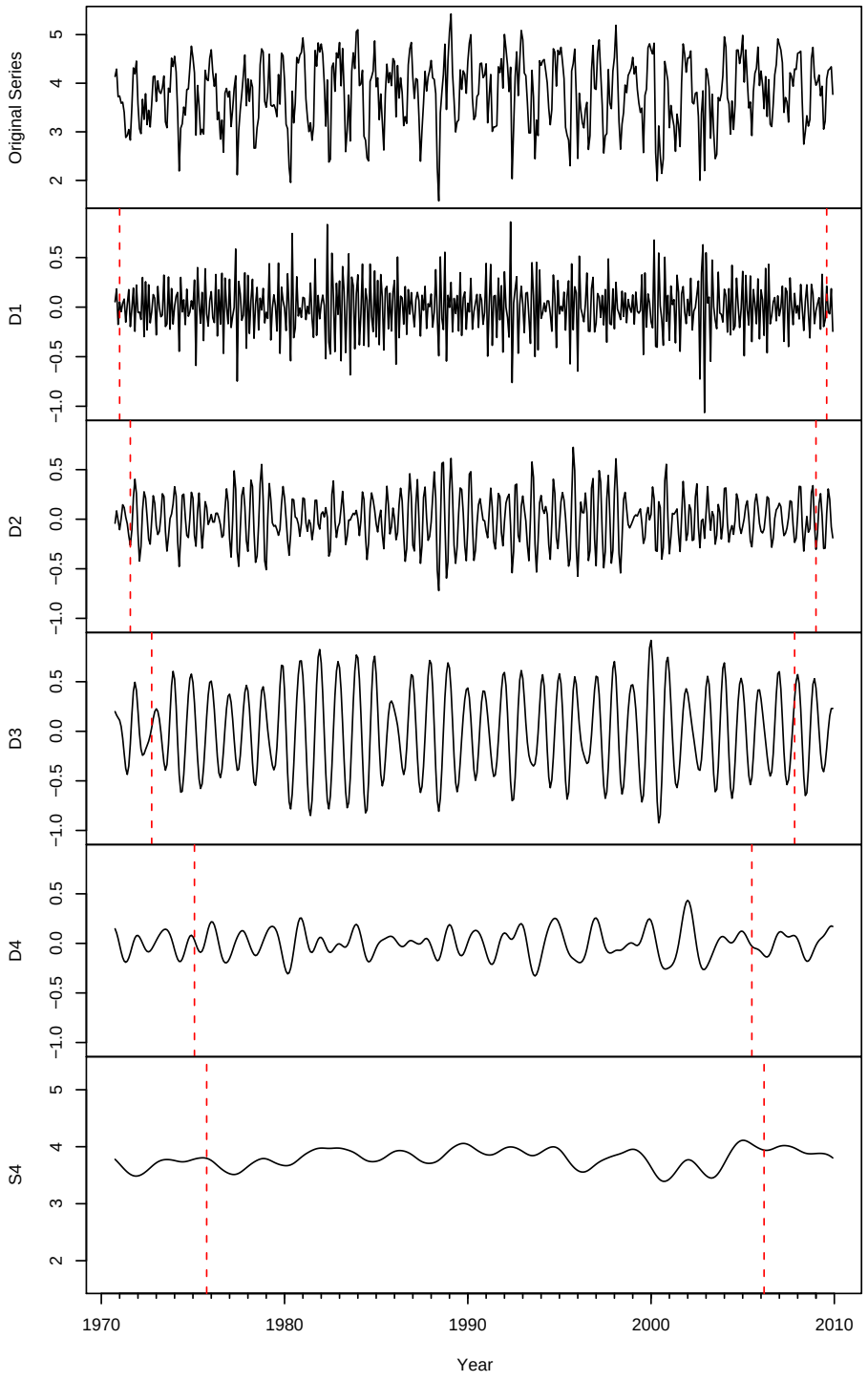

FIGURE B.11: Multiresolution analysis of monthly series - River Ewe (gauging station 94001)

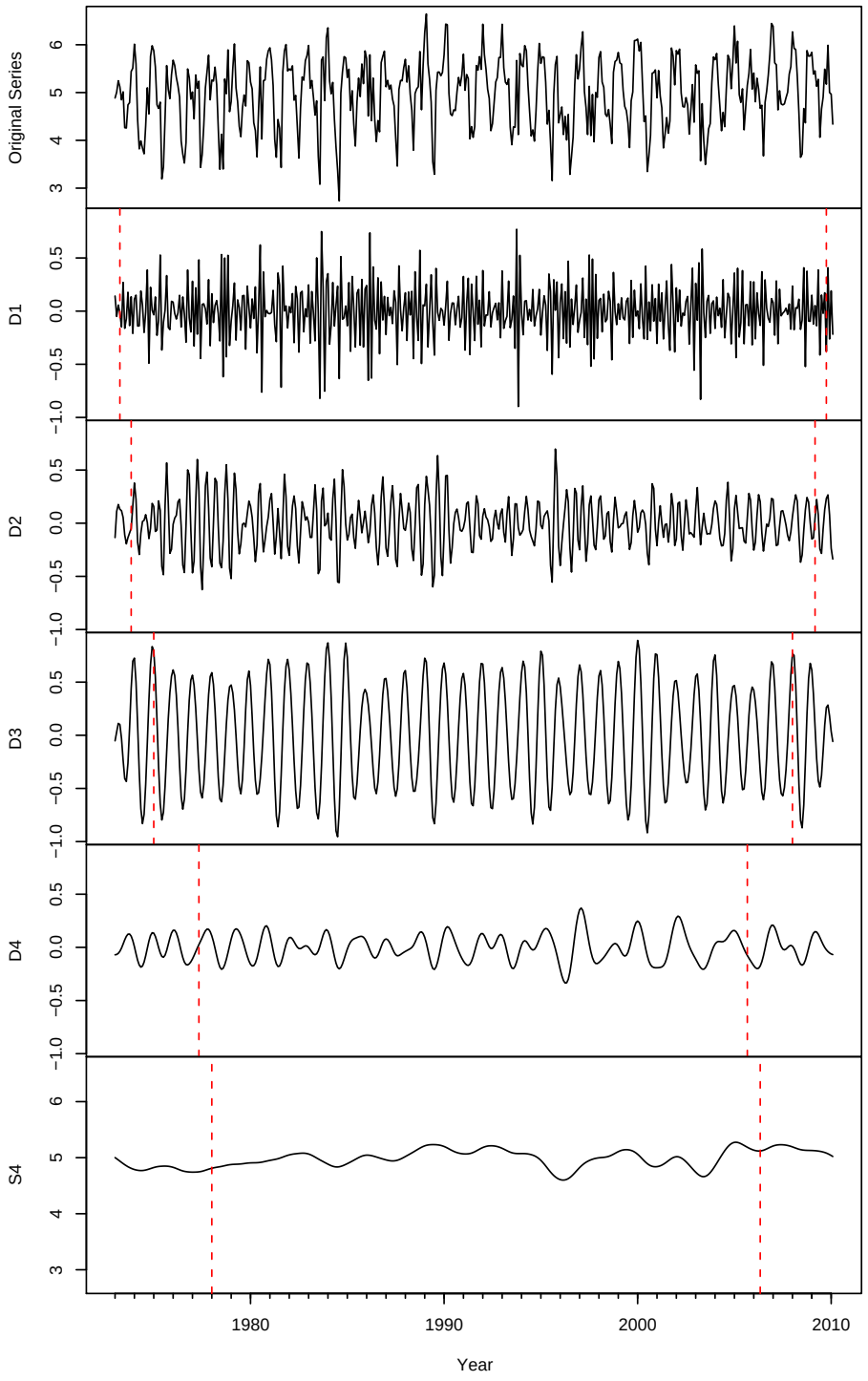

FIGURE B.12: Multiresolution analysis of monthly series - River Ness (gauging station 6007)

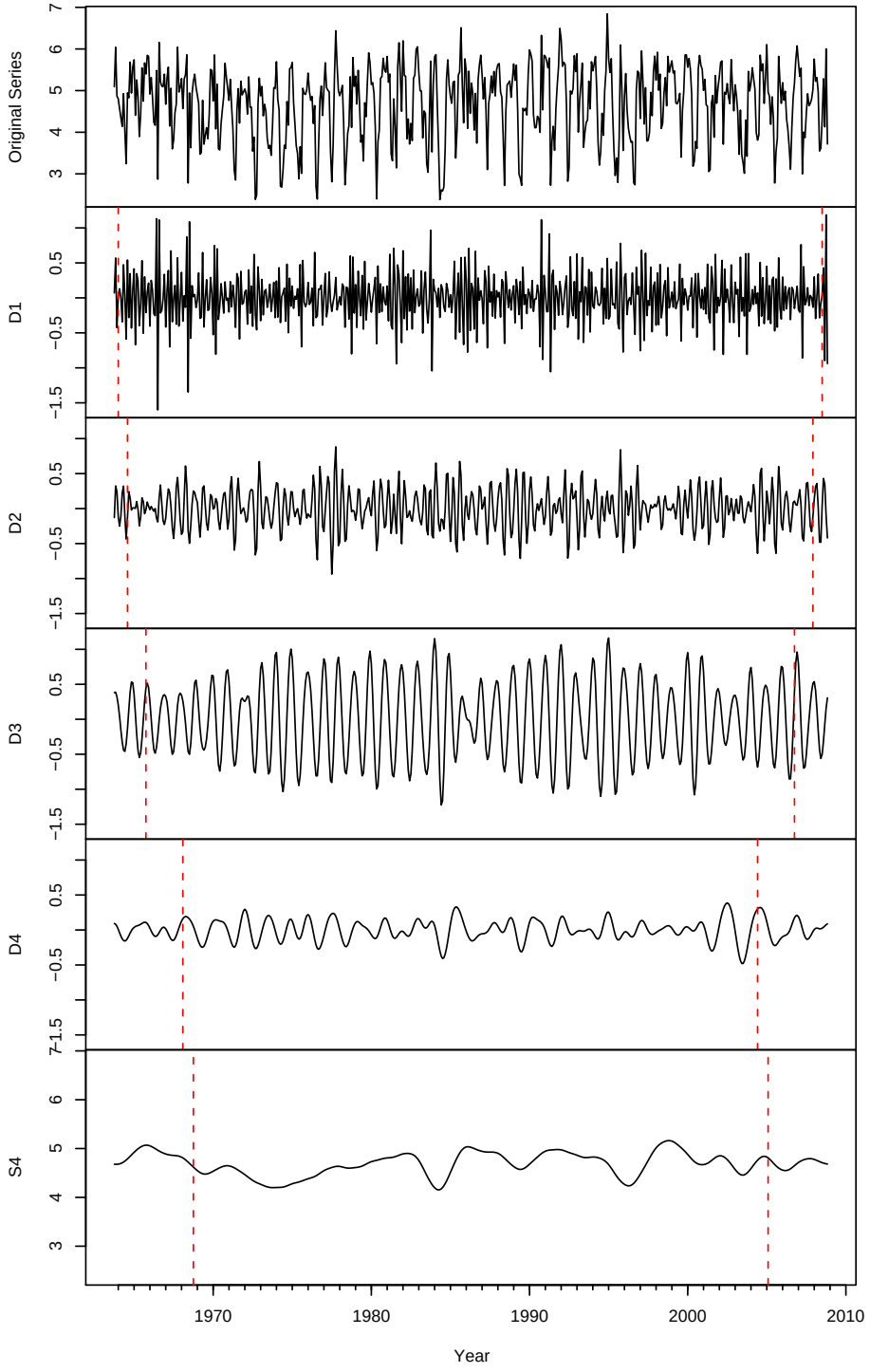

FIGURE B.13: Multiresolution analysis of monthly series - River Clyde (gauging station 84013)

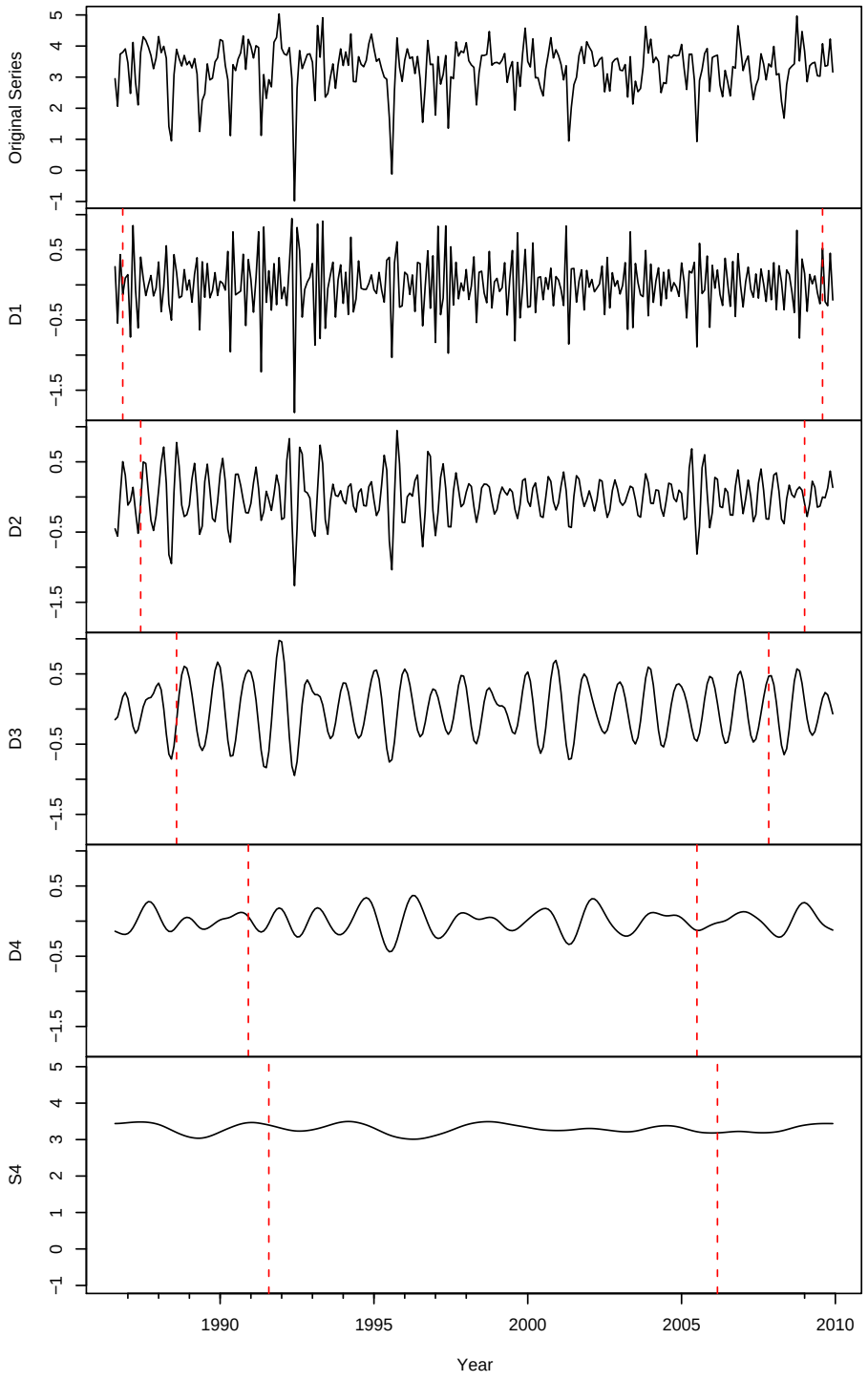

FIGURE B.14: Multiresolution analysis of monthly series - Water of Minnoch (gauging station 81006)

Bibliography

Acreman, M. and C. Sinclair (1986). Classification of drainage basins according to their physical characteristics; an application for flood frequency analysis in Scotland. *Journal of Hydrology 84*, 365–380.

Alexander, L., N. Tapper, X. Zhang, J. Fowler, C. Tebaldi, and A. Lynch (2009). Climate extremes: progress and future directions. *International Journal of Climatology 29*, 317–319.

Arnell, N. (1996). *Global warming, river flows and water resources*. Wiley.

Barnett, C., J. Hossell, M. Perry, and G. Hughes (2006). A handbook of climate trends across Scotland. Technical report, SNIFFER project CC03, Scotland and Northern Ireland Forum for Environmental Research: Edinburgh.

Bell, B. (1970). The oldest records of the Nile floods. *The Geographical Journal 136*(4), 569–573.

Beran, J. (1994). *Statistics for Long-Memory Processes*, Volume 61 of *Monographs on Statistics and Applied Probability*. Champman and Hall.

Black, A. (1996). Major flooding and increased flood frequency in Scotland since 1988. *Physics and Chemistry of the Earth 20*(5-6), 463–468.

Black, A. and A. Bennett (1994). Regional flooding in Strathclyde December 1994. *Hydrological Data - Institute of Hydrology - Clyde River Purification Board*, 29–34.

Black, A. and C. Burns (2002). Re-assessing the flood risk in Scotland. *The Science of the Total Environment 294*, 169–184.

Black, A. and A. Werritty (1997). Seasonality of flooding: a case study of North Britain. *Journal of Hydrology 195*, 1–25.

Black, A. R. and A. Werritty (1993). Seasonality and the generation of flood peaks in small Scottish upland catchments. In *Proceedings of the Fourth National Hydrology Symposium, Cardiff*.

Bouwer, L., J. Vermaat, and J. Aerts (2008). Regional sensitivities of mean and peak river discharge to climate variability in Europe. *Journal of Geophysical Research 113*, D19103.

Bowman, A., M. Giannitrapani, and E. Scott (2009). Spatiotemporal smoothing and sulphur dioxide trends over europe. *Journal of the Royal Statistical Society, series C 58*(5), 737–752.

Box, G. and G. Jenkins (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day.

Brown, B. and S. Resnick (1977). Extreme values of independent stochastic processes. *Journal of Applied Probability 14*, 732–739.

Buchinsky, M. (1998). Recent advances in quantile regression models: a practical guideline for empirical research. *The Journal of Human Resources 33*(1), 88–126.

Cade, B. and B. Noon (2003). A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment 1*(8), 412–420.

Cai, Z. (2002). Regression quantiles for time series. *Econometric Theory 18*, 169–192.

Cameron, D. (2006). An application of the UKCIP02 climate change scenarios to flood estimation by continuous simulation for a gauged catchment in the Northeast of Scotland, UK (with uncertainty). *Journal of Hydrology 328*, 212–226.

Cannas, B., A. Fanni, L. See, and G. Sias (2006). Data preprocessing for river flow forecasting using neural networks: Wavelet transforms and data partitioning. *Physics and Chemistry of the Earth 31*.

Clegg, R. (2005). A practical guide to measuring the Hurst parameter. Technical report, 21st UK Performance Engineering Workshop, School of Computing Science Technical Report Series, CSTR-916, University of Newcastle,pp 4355.

Cleveland, R., W. Celveland, J. McRae, and I. Terpenning (1990). Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics 6*(1), 3–73.

Coles, E. (2004). *An Introduction to Statistical Modelling of Extreme Values*. Springer Series in Statistics.

Cooley, D., D. Nychka, and P. Naveau (2007). Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association 102*(479), 824–840.

Craigmile, P. and D. Percival (2002). *Encyclopedia of Environmetrics*, Volume 4, Chapter Wavelet-based trend detection and estimation, pp. 2334–2338. John Wiley, Hoboken, N. J.

Cressie, N. (1993). *Statistics for Spatial Data (revised edition)*. Wiley Series in Probability and Mathematical Statistics.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Philadelphia: SIAM.

Davison, A. and M. Gholamrezaee (2012). Geostatistics of extremes. *Proceedings of the Royal Society A 468*, 581–608.

Davison, A. and D. Hinkley (1997). *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics.

Davison, A., S. Padoan, and M. Ribatet (2012). Statistical modelling of spatial extremes. *Statistical Science*.

Delworth, T. and M. Mann (2000). Observed and simulated multidecadal variability in the Northern Hemisphere. *Climate Dynamics 16*, 661–676.

Dettinger, M. and H. Diaz (2000). Global characteristics of stream flow seasonality and variability. *Journal of Hydrometeorology 1*, 289–310.

Diggle, P. and P. Ribeiro Jr. (2007). *Model-based Geostatistics*. Springer Series in Statistics.

Doukhan, P., G. Oppenheim, and M. Taqqu (2003). *Theory and Applications of Long-Range Dependence*. Birkhäuser.

Draghicescu, D., S. Guillas, and W. Wu (2009). Quantile curve estimation and visualization for nonstationary time series. *Journal of Computational and Graphical Statistics 18*(1), 1–20.

Draghicescu, D., S. Guillas, and W. B. Wu (2003). Direct quantile estimation for locally stationary time series with applications to ozone data. Technical report, Center for Integrating Statistical and Environmental Science, The University of Chicago.

Eastoe, E. and J. Tawn (2009). Modelling non-stationary extremes with application to surface level ozone. *Journal of the Royal Statistical Society - Series C 58*(1), 25–45.

Eilers, P. and B. Marx (1996). Flexible smoothing with B-splines and penalties. *Statistical Science 11*(2), 89–121.

Eilers, P. and B. Marx (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems 66*, 159–174.

Eilers, P. and B. Marx (2004). Splines, knots, and penalties.

Eilers, P. and B. Marx (2009). The craft of smoothing.

Fowler, H. and M. Ekström (2009). Multi-model ensemble estimates of climate change impacts on UK seasonal precipitation extremes. *International Journal of Climatology 29*, 385–416.

Fowler, H. and C. Kilsby (2003). Implications of changes in seasonal and annual extreme rainfall. *Geophysical Research Letters 30*(13), 1720. DOI:10.1029/2003GL017327.

Fowler, H. and R. Wilby (2010). Detecting changes in seasonal precipitation extremes using regional climate model projections: Implications for managing fluvial flood risk. *Water Resources Research 46*, W03525.

Fox, I. and R. Johnson (1997). The hydrology of the River Tweed. *The Science of the Total Environment 194-195*, 163–172.

Franco-Villoria, M., M. Scott, T. Hoey, and D. Fischbacher-Smith (2012). Temporal investigation of flow variability in Scottish rivers using wavelet analysis. *Journal of Environmental Statistics 3*(6).

Fuentes, M., J. Henry, and B. Reich (2012). Nonparametric spatial models for extremes: application to extreme temperature data. *Extremes*, 1–27. 10.1007/s10687-012-0154-1.

Galvao, A. F. J., G. Montes-Rojas, and S. Park (2009). Quantile autoregressive distributed lag model with an application to house price returns. City University Economics Discussion Papers 09/04, Department of Economics, City University, London.

Ghosh, S., J. Beran, and J. Innes (1997). Nonparametric conditional quantile estimation in the presence of long memory. *Student 2*(2), 109–117.

Giannitrapani, M. (2006). *Nonparametric methodologies for regression models with correlated data*. Ph. D. thesis, Department of Statistics, University of Glasgow.

Granger, C. and R. Joyeux (1980). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis 1*(1), 15–29.

Grew, H. and A. Werritty (1995). Changes in flood frequency and magnitude in Scotland 1964-1992. *Proceedings of the BHS,Fifth National Hydrological Symposium 1995, Edinburgh*, 3.1–3.9.

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

Hall, P., R. Wolff, and Q. Yao (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association 94*(445), 154–163.

Hallin, M., Z. Lu, and K. Yu (2009). Local linear spatial quantile regression. *Bernoulli 15*, 658–686.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. Chapman and Hall.

He, X., P. Ng, and S. Portnoy (1998). Bivariate quantile smoothing splines. *Journal of the Royal Statistical Society B 60*(3), 537–550.

Heffernan, J. and J. Tawn (2004). A conditional approach for multivariate extreme values. *Journal of the Royal Statistical Society - Series B 66*(3), 497–546.

Hendricks, W. and R. Koenker (1991). Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American Statistical Association 87*(417), 58–68.

Horowitz, J. and S. Lee (2005). Nonparametric estimation of an additive quantile regression. *Journal of the American Statistical Association 100*(472), 1238–1249.

Hosking, J. (1981). Fractional differencing. *Biometrika 68*(1), 165–176.

Hunt, J. (2002). Floods in a changing climate: a review. *Philosophical Transactions of the Royal Society London A 360*, 1531–1543.

Hurrell, J. (1995). Decadal trends in the North Atlantic Oscillation: Regional temperatures and precipitation. *Science 269*, 676–679.

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers 116*, 770–808.

Institute of Hydrology (1999). *Flood Estimation Handbook*. Center for Ecology & Hydrology.

Jenkins, G. J., J. M. Murphy, D. M. H. Sexton, J. A. Lowe, P. Jones, and C. G. Kilsby (2009). UK climate projections: Briefing report. Technical report, Met Office Hadley Centre, Exeter, UK.

Katz, R. and B. Brown (1992). Extreme events in a changing climate: variability is more important than averages. *Climate Change 21*, 289–302.

Keef, C., C. Svensson, and J. Tawn (2009). Spatial dependence in extreme river flows and precipitation for Great Britain. *Journal of Hydrology 378*, 240–252.

Keef, C., J. Tawn, and C. Svensson (2009). Spatial risk assessment for extreme river flows. *Applied Statistics 58*(5), 601–618.

Kerr, R. (2000). A North Atlantic climate pacemaker for the centuries. *Science 288*(5473), 1984–1985.

Kidson, R. and K. Richards (2005). Flood frequency analysis: assumptions and alternatives. *Progress in Physical Geography 29*, 392–410.

Kingston, D., D. Hannah, D. Lawler, and G. McGregor (2009). Climate-river flow relationships across montane and lowland environments in northern Europe. *Hydrological Processes 23*, 985–996.

Kisi, O. (2010). Wavelet regression model for short-term streamflow forecasting. *Journal of Hydrology 389*, 344–353.

Knight, J., C. Folland, and A. Scaife (2006). Climate impacts of the Atlantic Multidecadal Oscillation. *Geophysical Research Letters 33*, L17706.

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs.

Koenker, R. (2006). *Quantile Regression in R: a vignette*.

Koenker, R. (2011). Additive models for quantile regression: model selection and confidence bandaids. *Brazilian Journal of Probability and Statistics 25*(3), 239–262.

Koenker, R. and G. J. Bassett (1978). Regression quantiles. *Econometrica 46*(1), 33–50.

Koenker, R. and K. Hallock (2000). Quantile regression: an introduction. *Journal of Economic Perspectives*.

Koenker, R. and K. Hallock (2001). Quantile regression. *The Journal of Economic Perspectives 15*(4), 43–156.

Koenker, R. and J. Machado (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association 94*(448), 1296–1310.

Koenker, R., P. NG, and S. Protnoy (1994). Quantile smoothing splines. *Biometrika 81*(4), 673–680.

Koenker, R. and Z. Xiao (2006). Quantile autoregression. *Journal of the American Statistical Association 101*(475), DOI 10.1198/016214506000000672.

Koutsoyiannis, D. (2002). The Hurst phenomenon and fractional Gaussian noise made easy. *Hydrological Sciences 47*(4), 573–595.

Kyselý, J., J. Pcek, and R. Beranová (2010). Estimating extremes in climate change simulations using the peaks-over-threshold method with a non-stationary threshold. *Global and Planetary Change 72*, 55–68.

Labat, D. (2005). Recent advances in wavelet analyses: Part 1. A review of concepts. *Journal of Hydrology 314*, 275–288.

Labat, D. (2010). Cross wavelet analyses of annual continental freshwater discharge and selected climate indices. *Journal of Hydrology 385*, 269–278.

Labat, D., J. Ronchail, and J. Guyot (2005). Recent advances in wavelet analyses: Part 2 - Amazon, Parana, Orinoco and Congo discharges time scale variability. *Journal of Hydrology 314*, 289–311.

Lamb, H. (1972). *British Isles weather types and a register of daily sequence of circulation patterns, 1861-1971.* Geophysical Memoir 116, HMSO, London.

Lee, D. and T. Neocleous (2010). Bayesian quantile regression for count data with application to environmental epidemiology. *Journal of the Royal Statistical Society. Series C 59*(5), 905–920.

Macdonald, N., A. Werritty, A. Black, and L. McEwen (2006). Historical and pooled flood frequency analysis for the River Tay at Perth, Scotland. *Area 38*(1), 34–46.

Macklin, M. and B. Rumsby (2007). Changing climate and extreme floods in the British uplands. *Transactions of the Institute of British Geographers 32*, 168–187.

Mansell, M. (1997). The effect of climate change on rainfall trends and flooding risk in the West of Scotland. *Nordic Hydrology 28*, 37–50.

Maraun, D. and J. Kurths (2004). Cross wavelet analysis: significance testing and pitfalls. *Nonlinear Processes in Geophysics 11*, 505–514.

Markovic, D. and M. Koch (2005). Wavelet and scaling analysis of monthly precipitation extremes in Germany in the 20_{th} century: Interannual to interdecadal oscillations and the North Atlantic Oscillation influence. *Water Resources Research 41*, W09420.

Marsh, T. (1995). The 1995 drought - a water resources review in the context of the recent hydrological instability. *Hydrological Data UK Series*, 25–33.

Marsh, T. and J. Hannaford (2008). UK hydrometric register. Hydrological data UK series. Technical report, Centre for Ecology and Hydrology.

Maxim, V., L. Çendur, J. Fadili, J. Suckling, R. Gould, R. Howard, and E. Bullmore (2005). Fractional Gaussian noise, functional MRI and Alzheimer's disease. *Neuroimage 25*, 141–158.

Mayes, J. (1991). Regional airflow patterns in the British Isles. *International Journal of Climatology 11*, 473–491.

Mayes, J. (1996). Spatial and temporal fluctuations of monthly rainfall in the British Isles and variations in the mid-latitude westerly circulation. *International Journal of Climatology 16*, 585–596.

MetOffice (2012). Electronic Resource [Accessed 20/09/2012].

Montanari, A., R. Rosso, and M. Taqqu (1997). Fractionally differenced ARIMA models applied to hydrologic time series: identification, estimation, and simulation. *Water Resources Research 33*(5), 1035–1044.

Natural Environment Research Council (1975). *Flood Studies Report*. Institute of Hydrology (Great Britain).

Ng, P. and M. Maechler (2007). A fast and efficient implementation of qualitatively constrained quantile smoothing splines. *Statistical Modelling 7*, 315–328.

Northrop, P. and P. Jonathan (2011). Threshold modelling of spatially dependent non-stationary extremes with application to hurricane-induced wave heights. *Environmetrics*.

NRFA (2008). Electronic Resource [Accessed 1/11/08].

Opsomer, J., Y. Wang, and Y. Yang (2001). Nonparametric regression with correlated errors. *Statistical Science 16*(2), 134–153.

Palma, W. (2007). *Long-Memory Time Series. Theory and Methods*. Wiley Series in Probability and Statistics.

Pebesma, E. J. and R. N. M. Duin (2005). *Geostatistics for Environmental Applications*, Chapter Spatio-temporal mapping of sea floor sediment pollution in the North Sea, pp. 367–378. SpringerLink.

Percival, D. (1995). On estimation of the wavelet variance. *Biometrika 82*, 619–631.

Percival, D. and H. Mofjeld (1997). Analysis of subtidal coastal sea level fluctuations using wavelets. *Journal of the American Statistical Association 92*(439), 868–880.

Percival, D. and A. Walden (2006). *Wavelet Methods for Time Series Analysis*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Pratesi, M., M. Ranalli, and N. Salvati (2006). Nonparametric m-quantile regression via penalized splines. In *ASA Proceedings on Survey Research Methods, Alexandria, VA*, pp. 3596–3603.

Prudhomme, C., D. Jakob, and C. Svensson (2003). Uncertainty and climate change impact on the flood regime of small UK catchments. *Journal of Hydrology 277*, 1–23.

R Development Core Team (2001). A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Reich, B. (2012). Spatiotemporal quantile regression for detecting distributional changes in environmental processes. *Journal of the Royal Statistical Society. Series C 61*(4), 535–553.

Reich, B., M. Fuentes, and D. Dunson (2011). Bayesian spatial quantile regression. *Journal of the American Statistical Association 106*(493), 6–20.

Reiss, P. and L. Huang (2012). Smoothness selection for penalized quantile regression splines. *The International Journal of Biostatistics*. DOI:10.1002/env.1106.

Robson, A. (2002). Evidence for trends in UK flooding. *Philosophical Transactions of the Royal Society London A 360*, 1327–1343.

Rossi, A., N. Massei, B. Laignel, D. Sebag, and Y. Copard (2009). The response of the Mississipi River to climate fluctuations and reservoir construction as indicated by wavelet analysis of streamflow and suspended-sediment load, 1950-1975. *Journal of Hydrology 377*, 237–244.

Samanta, M. (1989). Non-parametric estimation of conditional quantiles. *Statistics and Probability Letters 7*, 407–412.

Sankarasubramanian, A. and U. Lall (2003). Flood quantiles in a changing climate: seasonal forecasts and causal relations. *Water Resources Research 39*(5), 1134.

Schlather, M. (2002). Models for stationary max-stable random fields. *Extremes 5*(1), 33–44.

Schulze, N. (2004). *Applied Quantile Regression: Microeconometric, Financial and Environmental analyses*. Ph. D. thesis, Tubingen.

Sen, A. (2009). Spectral-temporal characterization of riverflow variability in England and Wales for the period 1865-2002. *Hydrological Processes 23*, 1147–1157.

SEPA (2005). Scotland river basin district. Technical report, SEPA.

Serinaldi, F. (2010). Use and misuse of some Hurst parameter estimators applied to stationary and non-stationary financial time series. *Physica A 389*, 2770–2781.

Shaw, E. (1994). *Hydrology in Practice* (Third ed.). Chapman and Hall.

Sheather, S. and J. Marron (1990). Kernel quantile estimators. *Journal of the American Statistical Association 85*(410), 410–416.

Shorthouse, C. and N. Arnell (1997). Spatial and temporal variability in European river flows and the North Atlantic Oscillation. *FRIEND'97 - Regional Hydrology: Concepts and Models for Sustainable Water Resource Management 246*, 77–85.

Shumway, R. and D. Stoffer (2006). *Time Series Analysis and Its Applications - With R Examples* (Second ed.). Springer.

Simonsen, I. and I. Hansen (1998). Determination of the Hurst exponent by use of wavelet transforms. *Physical Review E 58*(3), 2779–2787.

Smith, L., D. Turcotte, and B. Isacks (1998). Stream flow characterization and feature detection using a discrete wavelet transform. *Hydrological Processes 12*, 233–249.

Smith, R. (1990). Max-stable processes and spatial extremes.

Smithson, P. (1969). Regional variations in the synoptic origin of rainfall across Scotland. *Scottish Geographical Magazine 85*(3), 182–195.

Sousa, S., J. Pires, F. Martins, M. Pereira, and M. Alvim-Ferraz (2009). Potentialities of quantile regression to predict ozone concentrations. *Environmetrics 20*, 147–158.

Spiegel, M. R. (1975). *Schaum's Outline of Theory and Problems of Probability and Statistics*. McGraw-Hill.

Steel, M. (1999). *Historic rainfall, climatic variability and flood risk estimation for Scotland*. Ph. D. thesis, Department of Geography, University of Dundee.

Sutton, R. and D. Hodson (2005). Atlantic Ocean forcing of North American and Europe summer climate. *Science 309*, 115–118.

Taqqu, M., V. Teverovsky, and W. Willinger (1995). Estimators for long-range dependence: and empirical study. *Fractals 3*(4), 785–798.

The Scottish Government (2010). Electronic Resource [Accessed 13/09/2010].

Torrence, C. and G. Compo (1998). A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society 79*, 61–78.

Torrence, C. and P. Webster (1999). Interdecadal changes in the ENSO-Monsoon system. *Journal of Climate 12*, 2679–2690.

Velasco, V. and B. Mendoza (2008). Assessing the relationship between solar activity and some large scale climatic phenomena. *Advances in Space Research 42*, 866–878.

Wadsworth, J. and J. Tawn (2012). Dependence modelling for spatial extremes. *Biometrika 99*(2), 253–272.

Werritty, A. (2002). Living with uncertainty: climate change, river flows and water resource management in Scotland. *The Science of the Total Environment 294*, 29–40.

Werritty, A. and J. Chatterton (2004). Foresight Future Flooding (Scotland). Technical report, Foresight Programme.

Werritty, A. and T. Hoey (2004). Geomorphical changes and trends in Scotland: river channels and processes. Technical Report 053, Scottish Natural Heritage Commissioned Report. ROAME No. F00AC107B.

Werritty, A., D. Houston, T. Ball, A. Tavendale, and A. Black (2007). Exploring the social impacts of flood risk and flooding in Scotland. Technical report, Scottish Executive Social Research.

Whitcher, B., P. Guttorp, and D. Percival (2000). Wavelet analysis of covariance with application to atmospheric time series. *Journal of Geophysical Research 105*(D11), 14941–14962.

Wood, S. (2006). *Generalized Additive Models - An Introduction with R*. Champman and Hall/CRC.

Xiao, Z. and R. Koenker (2009). Conditional quantile estimation for GARCH models. *Journal of the American Statistical Association 104*, 1696–1712.

Yu, K., Z. Lu, and J. Stander (2003). Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society. Series D (The Statistician) 52*(3), 331–350.

Zhou, Z. and W. Wu (2009). Local linear quantile estimation for nonstationary time series. *The Annals of Statistics 37*(5B), 2696–2729.