Potsdam Institut für Klimafolgenforschung

Detection of Long-Range Dependence

Applications in Climatology and Hydrology

Dissertation zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)
in der Wissenschaftsdisziplin Theoretische Physik

Eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Potsdam

von
Henning Rust

Potsdam, im Januar 2007

Contents

Chapter 1

Introduction

The mystery of nature's variability has been perceived as a potential threat since ancient times (Genesis 41:29-30): "Seven years of great abundance are coming throughout the land of Egypt, but seven years of famine will follow them. Then all the abundance in Egypt will be forgotten, and the famine will ravage the land." Only this prophecy made by Joseph in the Old Testament led to the storage of excess food produced in the years of abundance and saved the ancient Egyptians from starving in the years of famine. Likewise in modern times the British hydrologist Harold Edwin Hurst studied the Nile River minimal flows to "prophesy" the inter-annual water level variability (Hurst, 1951). The aim of his work was to determine the size of a reservoir large enough to allow a constant outflow given the highly variable inflows and thereby finally ensure a constant food production under varying natural water resources. During his analysis, Hurst observed that water levels seem to be strongly influenced by those observed in the distant past. The slow algebraic decay of this interdependence (or autocorrelation) was different from standard models. This is later referred to as Hurst phenomenon or long-range dependence contrary to short-range dependence specifying an exponential decay.

Droughts and the resulting scarcity of aliments are still challenges to overcome (e.g., Rebetez et al., 2006). These challenges are enhanced by floods which result in a menacing excess of water, and create the need for hydraulic structures potent enough to withstand extreme (i.e. unlikely high) water discharges. It is desirable that the structure itself is conserved, in case of dams and dykes, as well as the properties behind it. In recent years floods have become more abundant and more destructive (Kundzewicz and Schellnhuber, 2004). Extreme floods have been recorded in central Europe in the last decades, e.g., at the Donau 1997 or at the Elbe 2002 (Becker and Grünewald, 2003). Therefore flood risk assessment and the associated uncertainty have become a highly relevant topic on the scientific as well as on the political agenda (e.g., Apel et al., 2004; Merz, 2006; WMO/UNEP, 2002). Flood risk assessment involves an estimate of the probability of exceeding a certain water level or discharge, i.e. an extreme value analysis. In a similar way as Hurst's storage problem, this analysis is affected by the autocorrelation of the observed water levels (Koutsoyiannis, 2003; Coles and Tawn, 1999). It reduces the information obtained from a new observation because the interdependence determines to a certain extent the subsequent observation. In this way the autocorrelation structure of the underlying process influences the uncertainty of quantities estimated from its samples. Especially processes with long-range dependence, as observed by Hurst, show this effect. If the autocorrelation of the underlying process is not known, appropriate assumptions are needed for a meaningful uncertainty analysis.

Assumptions on such a "natural variability" are also necessary in other contexts. The

detection of particular departures from a natural behaviour requires the specification of what is still consistent with random deviations inherent to the system. The most prominent example of this kind of problem is the detection of climate change and the associated attribution to anthropogenic influence (IPCC, 2001). An outstanding ansatz in this respect is the so-called fingerprint method (e.g., Hasselmann, 1997). Spatial patterns of climate variables under different forcing conditions are obtained from general circulation models and are sought in the observations. Instead of these multivariate patterns one can study global or hemispheric temperature observations with respect to the detection of an anthropogenic trend. For these one-dimensional records one might hypothesise polynomial trends and, instead of physically motivated processes, various types of correlated noise terms to model the "natural variability". Such approaches have been used by, e.g., Wigley and Jones (1981); Woodward and Gray (1995); Bloomfield (1992) and Cohn and Lins (2005). The selection of an adequate noise model plays a decisive role in these analyses. In their work on the northern hemisphere temperature series, Cohn and Lins (2005) demonstrated the susceptibility of trend tests to changes in the assumption made for the noise model. In particular, the assumption of a long-range dependent process alters the test result in a way that a linear trend is not found to be significant. This analysis is an instructive example of the interplay of long-range memory and deterministic trends.

A reliable detection of long-range dependence has, thus, relevant consequences in various respects. Long-range dependence influences a trend assessment in geophysical records, examples range from temperature (e.g., Bloomfield, 1992; Craigmile et al., 2004; Cohn and Lins, 2005) over stratospheric ozone (e.g., Vyushin et al., 2007) to water discharge records (e.g., Kallache, Rust, and Kropp, 2005). Furthermore, uncertainty analyses might be affected, e.g. in the case of flood return level estimates (e.g., Koutsoyiannis, 2003; Rust et al., 2006). These problems have also been addressed within the framework of the project "Scaling Analysis of Hydro-Meteorological Time Series", funded by the German Federal Ministry of Education and Research (BMBF). Aims of this project were, among others, the development of methods for the detection and quantification of long-range dependence, as well as to provide reliable strategies for trend detection and uncertainty analysis in extreme value statistics under consideration of dependence effects. This thesis sets out to tackle the fundamental problem of detection and quantification of long-range dependence.

Long-range dependence is an asymptotic property of a stochastic process, characterising the decay of the correlation between observations separated by increasingly large time lags. Therefore, it is in principle not possible to unambiguously infer an underlying long-range dependent process from a finite sample. The strategy suggested in this thesis, is to escape this dilemma by rephrasing the detection of long-range dependence as a model selection problem: does a short-range or a long-range dependent model more adequately describe the observations? To answer this question it is necessary to have a suitable set of long-range and short-range dependent models along with an adequate parameter estimation strategy. Furthermore, it is inevitable to impose a principle which allows to compare such models, particularly the most suitable long-range and the most suitable short-range dependent model.

Models to describe the Hurst phenomenon, or long-range dependence, were not at hand when Hurst first observed the slowly decaying dependence for the Nile river minimum flows, or for other geophysical observations. A seminal contribution towards the development of such models was made in the 1960s by the French mathematician Benoît Mandelbrot. Mandelbrot introduced a stochastic process into the statistics com-

munity which was able to reproduce the Hurst phenomenon (Mandelbrot and van Ness, 1968). This so-called fractional Gaussian noise model is related to the concept of self-similarity, or fractals. Fractional Gaussian noise as well as other self-similarity concepts became increasingly popular during the 1980s and are still frequently used. They can be found, for example, in the theory of turbulence (e.g., Frisch, 1995), seismology (e.g., Okubo and Aki, 1987), atmospheric science (e.g., Cahalana et al., 1994) , and many more (e.g., Bunde and Havlin, 1994). Critical votes against the ubiquitousness of fractals were raised in the late 1990s (e.g., Avnir et al., 1998). Despite being an attractive concept, simple scaling relations do not always provide an adequate description of the observed data, as examples from time series analysis (e.g., Metzler, 2003; Maraun, Rust, and Timmer, 2004) or turbulence (e.g., Renner et al., 2002) show.

Regarding the description of the Hurst phenomenon, flexible alternatives to fractional Gaussian noise emerged in the framework of stochastic processes in the years following Mandelbrot's seminal contribution. A milestone was set with fractional differencing, a concept independently introduced by Granger and Joyeux (1980) and Hosking (1981). They defined a new class of stochastic processes, called fractional difference processes, having the same asymptotic properties (i.e. for large time scales) as fractional Gaussian noise. Thus, they also provide a description of the Hurst phenomenon, or long-range dependence. Formulated in the framework of linear stochastic models, fractional difference processes can be combined with the widely used autoregressive moving average processes (Box and Jenkins, 1976) to create a flexible class of models capable of describing the Hurst phenomenon. This class forms the basis for a strategy to detect and quantify long-range dependence, which is developed in this thesis.

The above mentioned problems from hydrology and climatology as well as the demands on the methods needed to resolve them can be summarised in the following questions which will be addressed in the following:

- What is an adequate strategy for choosing between different models which may be used to describe the autocorrelation structure? How can long-range dependence be detected in a reliable way?
- What effect does autocorrelation have on trend detection?
- What are the consequences of considering or ignoring autocorrelations for flood risk assessment? How can the uncertainty of flood return level estimates in autocorrelated records be quantified?

The last two questions address effects of long-range dependence and are discussed along examples from hydrology and climatology in Chapters 5, 6 and 7. Prior to this discussion, the methodological apparatus required to tackle the first question is developed in Chapters 3 and 4.

This thesis is structured as follows. Basic concepts of time series analysis and the notion of long-range dependence are introduced in the second chapter. This provides the basis for the subsequent discussion of the stochastic processes used in the remainder of this thesis, with respect to parameter estimation and simulation. A central point in the stochastic modelling approach pursued here is the selection of an adequate process to represent the autocorrelation function of the observations. The third chapter thus presents a goodness-of-fit test and standard approaches for choosing between nested models[1].

[1]The model g is nested in model f if it constrains one or more parameters of f, typically to zero. This notion is different from nested in the context of climate models.

These methods are enhanced by a simulation-based selection approach for non-nested models.

The framework constructed in the first chapters is used in Chapter 4 to formulate a strategy for the detection of long-range dependence based on a parametric modelling approach. The major difference to other methods is that, beyond investigating the observed large scale behaviour, this method investigates whether this behaviour is more likely to be consistent with a short-range or a long-range dependent model.

In Chapter 5, we investigate an outstandingly long record of surface air temperatures for long-range dependence, and use the resulting model for a trend test. A prominent example is the record of northern hemisphere mean temperature anomalies. This series has been investigated for trends under consideration of various models for the residual fluctuations by, e.g., Bloomfield (1992); Woodward and Gray (1995) and Cohn and Lins (2005). Using the parametric modelling approach, we re-investigate for long-range dependence, and discuss the assumption of a linear trend as suggested by Cohn and Lins (2005).

In Chapter 6, we analyse two run-off records with daily time resolution from southern Germany and one record of monthly mean values from Poland, with respect to long-range dependence. For one of these records, we estimate a hundred-year return level, i.e. the water level (or discharge) exceeded on average once every hundred years, and use the stochastic models found previously for an uncertainty analysis (Chapter 7).

The various sections in the appendix complement this thesis by providing an extensive review on detrended fluctuation analysis (Appendix A) and additional material regarding long-range dependence (Appendix B). Furthermore, an approach for estimating confidence intervals for records from autocorrelated series is proposed in Appendix C. Finally, the data sources and data preprocessing is discussed (Appendix D).

Chapter 2

Time Series Analysis and Stochastic Modelling

A time series is a sequence of data points measured at successive time intervals. Understanding the mechanism that generated this time series or making predictions are the essence of time series analysis. Applications range from physiology (e.g., Timmer, 1998) and systems biology (e.g., The et al., 2005), over economy (e.g., Lo, 1991) and geophysical problems as daily weather forecasts (e.g., Orrel et al., 2001) or detection and prediction of climate change (e.g., Held and Kleinen, 2004) to astronomy (e.g., Timmer et al., 2000).

This chapter introduces some concepts of linear time series analysis and stochastic modelling. Starting with random variables, we briefly introduce spectral analysis and discuss some special stochastic processes. An emphasis is made on the difference between short-range and long-range dependence, a feature especially relevant for trend detection and uncertainty analysis. Equipped with a canon of stochastic processes, we present and discuss ways of estimating optimal process parameters from empirical data.

2.1 Basic Concepts of Time Series Analysis

2.1.1 Random Variables

A random variable X is a mapping $X : \Omega \to \mathbb{R}$ from a sample space Ω onto the real axis. Given a random variable one can define probabilities of an event. As a simple example we consider a die with a sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$. The probability P for the event $d \in \Omega$, $d =$"the number on a die is smaller than 4" is denoted as $P(X < 4)$. Such probabilities can be expressed using the cumulative probability distribution function

$$F_X(x) = P(X \leq x). \tag{2.1}$$

For a continuous random variable X, $F_X(x)$ is continuous. A discrete random variable X leads to $F_X(x)$ being a step function.

With miniscules x we denote a realisation $x \in \mathbb{R}$, a possible outcome of a random variable X. A sample is a set of N realisations $\{x_i\}_{i=1,\ldots,N}$.

Dependence

Two random variables X and Y are *independent* if their joint probability distribution $P(X < x, Y < y)$ can be written as a product of the individual distributions: $P(X < x, Y < y) =$

$P(X < x)P(Y < y)$, or, using the conditional distribution, $P(X < x \mid Y < y) = P(X < x)$. They are called *dependent* otherwise. A particular measure of dependence is the covariance which specifies the linear part of the dependence:

$$\mathrm{cov}(X, Y) = \mathrm{E}\left[(X - \mathrm{E}\left[X\right])(Y - \mathrm{E}\left[Y\right])\right], \tag{2.2}$$

where $\mathrm{E}\left[X\right] = \int_{\mathbb{R}} x f_X(x) \mathrm{d}x$ denotes the expectation value with $f_X(x) = \mathrm{d}F_X(x)/\mathrm{d}x$ being the probability density function (Feller, 1968).

A normalised measure quantifying the strength and direction[1] of the linear relationship of two random variables X and Y is the correlation

$$\mathrm{cor}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}}, \tag{2.3}$$

with $\mathrm{var}(X) = \mathrm{cov}(X, X)$ being the variance of X. It is a normalised measure taking values in the interval $-1 \leq \mathrm{cor}(X, Y) \leq 1$. An absolute correlation of unity, i.e. $|\mathrm{cor}(X, Y)| = 1$, implies Y being a linear function of X.

The variables X and Y are said to be uncorrelated if their covariance function, and thus their correlation, vanishes:

$$\mathrm{cov}(X, Y) = 0. \tag{2.4}$$

Correlated random variables are also dependent. The opposite statement is not necessarily true, because the correlation captures only the linear part of the dependence. Independent random variables X_i with identical distribution function are referred to as independent and identically-distributed random variables (IID).

For a set of random variables $X_1, X_2, \ldots, X_M$ the covariance matrix $\mathbf{\Sigma}$ with elements Σ_{ij} is defined as

$$\Sigma_{ij} = \mathrm{cov}(X_i, X_j). \tag{2.5}$$

2.1.2 Stochastic Processes

The word *stochastic* originates from the Greek *stochazesthai* ($\sigma\tau\omega\chi\acute{\alpha}\zeta\epsilon\sigma\theta\alpha\iota$) meaning "to aim at" or "to guess at" (Merriam-Webster, 2005). It is used in the sense of random in contrast to deterministic. While in a deterministic model the outcome is completely determined by the equations and the input (initial conditions), in a stochastic model no exact values are determined but probability distributions. In that sense, a stochastic model can be understood as a means to guess at something.

The choice between a deterministic and a stochastic model is basically one of what information is to be included in the equations describing the system. On the one hand information can be limited simply by the lack of knowledge. On the other hand it might not be benefitting the modelling objective to include certain information.

A stochastic process $X(t)$ or X_t is an indexed collection of random variables with the indices specifying a time ordering. The index set can either be discrete ($t \in \mathbb{N}$) or continuous ($t \in \mathbb{R}$). In the latter case the collection consists of an uncountable infinite number of random variables.

With $\mathbf{x}_{t=1,\ldots,N} = (x_1, x_2, \ldots, x_N)$ we denote a realisation of the stochastic process X_t for $1 \leq t \leq N$. Where unambiguous, we omit the index. Depending on the context, $\mathbf{x}$ also denotes an empirical data record which is considered as a realisation of a possibly unknown process.

[1] Direction specifies the sign of the correlation. The direction is positive if X increases when Y increases and negative otherwise. It is not meant in the sense of variable X influencing variable Y or vice versa.

Stationarity

If the joint probability distribution $P(X_{t_1} < x_1, X_{t_2} < x_2, \ldots, X_{t_n} < x_n)$, with $t_i \in \mathbb{N}$, is identical with a displaced one $P(X_{t_1+k} < x_1, X_{t_2+k} < x_2, \ldots, X_{t_n+k} < x_n)$ for any admissible $t_1, t_2, \ldots, t_n$ and any k, the process is called *completely stationary*.

An alleviated concept is *stationarity up to an order m* with the first m moments of the joint probability distribution existing and being equal to those of the displaced one. For practical applications, frequently stationarity up to order $m = 2$ or *weak stationarity* is required. Testing for higher orders or for complete stationarity is usually not feasible.

Autocovariance

The covariance of two instances X_{t_1} and X_{t_2} at two different times t_1 and t_2 of a stochastic process X_t is called *autocovariance*

$$\operatorname{cov}(X_{t_i}, X_{t_j}) = \mathrm{E}\left[\left(X_{t_i} - \mathrm{E}\left[X_{t_i}\right]\right)\left(X_{t_j} - \mathrm{E}\left[X_{t_j}\right]\right)\right]. \tag{2.6}$$

For processes which are stationary at least up to order $m = 2$ the autocovariance depends only on the differences $\tau = t_i - t_j$ and can be written as

$$\operatorname{cov}(\tau) = \mathrm{E}\left[\left(X_t - \mathrm{E}\left[X_t\right]\right)\left(X_{t+\tau} - \mathrm{E}\left[X_t\right]\right)\right]. \tag{2.7}$$

Analogously, the *autocorrelation* can be defined as

$$\rho(X_{t_i}, X_{t_j}) = \frac{\operatorname{cov}(X_{t_i}, X_{t_j})}{\sqrt{\operatorname{var}(X_{t_i})}\sqrt{\operatorname{var}(X_{t_j})}}, \tag{2.8}$$

and in case of at least weak stationarity it can be written as

$$\rho(\tau) = \rho(X_t, X_{t+\tau}). \tag{2.9}$$

In the following, we refer to Equation (2.9) as the *autocorrelation function* or ACF.

Given a sample of an unknown process, we can estimate process characteristics as the ACF. An estimator for a characteristic T is a function of the sample and will be denoted by $\widehat{T}$. A nonparametric estimate for the autocovariance of a zero mean record $\mathbf{x}_{t=1,\ldots,N}$ is the sample autocovariance function or autocovariance sequence (Brockwell and Davis, 1991)

$$\widehat{\operatorname{cov}}(\tau) = \frac{1}{N}\sum_{t=1}^{N-\tau} x_t x_{t+\tau}, \quad 0 \leq \tau < N. \tag{2.10}$$

Using the divisor N instead of $N - \tau$ ensures that the estimate for the autocovariance matrix $\mathbf{\Sigma}$ is non-negative definite. Consequently an estimate for the autocorrelation function can be formulated as

$$\widehat{\rho}(\tau) = \frac{\widehat{\operatorname{cov}}(\tau)}{\widehat{\sigma}_x^2}, \tag{2.11}$$

with $\widehat{\sigma}_x^2 = \widehat{\operatorname{cov}}(0)$ being an estimate of the record's variance.

For a more detailed description of these basic concepts refer to, e.g., Priestley (1992); Billingsley (1995), or Brockwell and Davis (1991).

2.1.3 Spectral Analysis

While the time domain approach to time series analysis originates from mathematical statistics, the spectral or frequency domain approach has its root in communication engineering. Whenever a signal fluctuates around a certain stable state we might use periodic functions to describe its behaviour. Spectral analysis aims at splitting the total variability of a stationary stochastic process into contributions related to oscillations with a certain frequency.

For the definition of the spectral density of a continuous stationary process $X(t)$ consider initially the process

$$X_T(t) = \begin{cases} X(t), & \text{if} \quad -T \leq t \leq T \\ 0, & \text{otherwise} \end{cases} . \tag{2.12}$$

This process is absolutely integrable due to the compact support $[-T, T]$. We now can define a Fourier integral as

$$G_T(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-T}^{T} X_T(t) \mathrm{e}^{-i\omega t} \mathrm{d}t. \tag{2.13}$$

and a power spectral density

$$\widetilde{S}(\omega) = \lim_{T \to \infty} \frac{|G_T(\omega)|^2}{2T}. \tag{2.14}$$

We use the notion of power as energy per time to obtain a finite spectral density. The aim is to represent the stochastic process and not only a single realisation. We thus have to average over multiple realisations. This leads us to a definition of a *power spectral density* for stochastic processes

$$S(\omega) = \lim_{T \to \infty} \mathrm{E}\left[\frac{|G_T(\omega)|^2}{2T}\right]. \tag{2.15}$$

In the following, we refer to the power spectral density simply as the spectral density. A detailed derivation of this concept is given by Priestley (1992).

Relation to the Autocovariance

The relation between the spectral density $S(\omega)$ and the autocovariance function $\mathrm{cov}(\tau)$ of a zero mean process is surprising at first sight. Using some basic properties of the Fourier transform, we can establish that the spectral density can be expressed as the Fourier transform of the autocovariance function:

$$S(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{cov}(\tau) \mathrm{e}^{-i\omega\tau} \mathrm{d}\tau. \tag{2.16}$$

In a more general setting, when the spectral density function does not exist but the integrated spectrum $F(\omega)$ does, this result manifests in the Wiener-Khinchin theorem using the more general Fourier-Stieltjes form of the integral (Priestley, 1992)

$$\rho(\tau) = \int_{-\infty}^{\infty} \mathrm{e}^{i\omega\tau} \mathrm{d}F(\omega), \tag{2.17}$$

with $\rho(\tau)$ being the ACF. Thus, the spectral density and the ACF are equivalent descriptions of the linear dynamic properties of a process.

Estimation of the Spectral Density

Given a zero mean series **x** sampled at discrete time points $t = 1, \ldots, N$, an estimator for the spectral density function (2.15) can be formulated using the periodogram $I(\omega_j)$

$$\widehat{S}(\omega_j) = I(\omega_j) = \frac{1}{2\pi N} |\sum_{t=1}^{N} x_t \mathrm{e}^{-it\omega_j}|^2, \tag{2.18}$$

with the Fourier frequencies $\omega_j = 2\pi j/N$ and $j = 1, \ldots, [(N-1)/2]$, where $[.]$ denotes the integer part. The periodogram is not a consistent[2] estimator for the spectral density because its variance does not decrease with increasing length N of the sample. To obtain a consistent estimator we can locally smooth the periodogram with, e.g., a suitable rectangular or bell-shaped window (e.g., Priestley, 1992).

Instead of the angular frequency ω, it is convenient to use the frequency $f = \omega/2\pi$ in practical applications. In this way, we can specify frequencies as reciprocal periods, which are expressed in units of the sampling interval, e.g., 1/days for data with a daily resolution.

2.1.4 Long-Range Dependence

For some stochastic processes, such as the popular autoregressive and moving average type models that will be introduced in Section 2.2, the ACF, denoted as $\rho_{\mathrm{SRD}}(\tau)$, decays exponentially and is thus summable:

$$\sum_{\tau=-\infty}^{\infty} \rho_{\mathrm{SRD}}(\tau) = \mathrm{const} < \infty. \tag{2.19}$$

These processes are called *short-range dependent* (SRD) or short-range correlated.

This characteristic contradicted other findings from, e.g., a spatial analysis of agricultural data by Smith (1938) or the famous Nile River flow minima studied by Hurst (1951). The behaviour they observed for the ACF for large lags – often referred to as Hurst phenomenon – was not consistent with hitherto existing models describing dependence. It was well represented assuming an algebraic decay of the ACF: $\rho_{\mathrm{LRD}}(\tau) \propto \tau^{-\gamma}$. This type of decay leads to a diverging sum

$$\sum_{\tau=-\infty}^{\infty} \rho_{\mathrm{LRD}}(\tau) = \infty. \tag{2.20}$$

Processes with an ACF following (2.20) are called *long-range dependent* (LRD), long-range correlated or long-memory processes.

Alternative Definitions of Long-Range Dependence

0 Today, it is common to use equation (2.20) as definition for a long-range dependent process (Robinson, 2003). The following alternative formulations in the time and spectral domain, respectively, are consistent with (2.20).

[2] An estimator $\widehat{T}_N$ for the parameter T based on N observations is consistent iff for all $\epsilon > 0$ $\lim_{N\to\infty} P((\widehat{T}_N - T) < \epsilon) = 1$.

Time Domain Let X_t be a stationary process. If there exists a real number $\gamma \in (0,1)$ and a constant $c_\rho > 0$ such that

$$\lim_{\tau \to \infty} \frac{\rho(\tau)}{\tau^{-\gamma}} = c_\rho \tag{2.21}$$

holds, then X_t is called a process with long-range dependence or long memory.

Spectral Domain If there exists a real number $\beta \in (0,1)$ and a constant $c_f > 0$ such that

$$\lim_{\omega \to 0} \frac{S(\omega)}{|\omega|^{-\beta}} = c_S \tag{2.22}$$

holds then X_t is called a process with long-range dependence.

The concept of long-memory refers to non-periodic stochastic processes. Therefore the recurrence due to periodicities, such as the Milankovitch cycles[3] in the climate system, is not to be considered as long-range dependence, even if their (deterministic) behaviour causes dependence for infinite time lags.

2.2 Some Stationary Stochastic Processes

This section introduces some examples of stationary stochastic processs, SRD, as well as LRD processes. Throughout this thesis, they are used to describe temperaure anomalies (Chapter 5) or run-off records (Chapter 6). Furthermore, they are the fundament of the bootstrap based approach for the quantification of uncertainty of flood return level estimates for dependent processes (Chapter 7).

Gaussian Processes

A stochastic process X_t with joint probability distribution $P(X_{t_1} < x_1, X_{t_2} < x_2, \ldots, X_{t_k} < x_k)$ for all $t_k \in \mathbb{N}$ being multivariate normal is called a Gaussian process. It is completely defined by its first two moments, the expectation value $\mathrm{E}[X_t]$ and the autocovariance function $\mathrm{cov}(X_{t_i}, X_{t_j})$ or autocovariance matrix $\mathbf{\Sigma}$. This implies that a weakly stationary Gaussian process is also completely stationary (cf. Section 2.1.2). Furthermore, a Gaussian process is thus a linear process, because the autocovariance matrix specifies only linear relationships between the different instances X_{t_k}. Simple examples for discrete time Gaussian stochastic processes are the Gaussian white noise and autoregressive processes described in the following.

2.2.1 Processes with Short-Range Dependence

Gaussian White Noise Process

Let $\{X_t\}_{t=0,\pm1,\pm2,\ldots}$ be a sequence of uncorrelated[4] Gaussian random variables, then X_t is called a Gaussian white noise process or a purely random process. It possesses "no memory" in the sense that the value at time t is not correlated with a value at any other time s: $\mathrm{cov}(X_t, X_s) = 0, \forall t \neq s$. Hence, the spectrum is flat. A plot of a realisation and the corresponding periodogram of a zero mean Gaussian white noise process is given in Figure 2.1. In the following, we frequently refer to a Gaussian white noise process simply

[3] Milankovitch cycles are quasi-periodic changes in the earth's orbital and rotational properties resulting from being exposed to the gravitational force of multiple planets. The change in the orbital parameters effect the earth's climate.

[4] or, more generally, independent

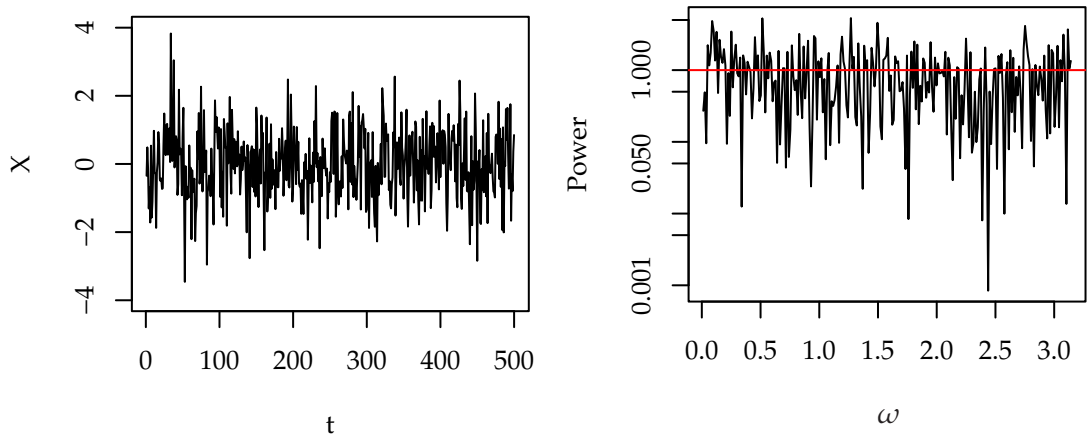

Figure 2.1: Realisation of a white noise process with 500 points (left) and the corresponding periodogram with the spectral density as red solid line (right).

as white noise. To state that X_t is such a process with mean μ and variance σ^2, we use the notation $X_t \sim \mathcal{WN}(0, 1)$.

First Order Autoregressive Process (AR$[1]$)

An instance X_t at time t of an autoregressive process depends on its predecessors in a linear way. A first order autoregressive process (AR$[1]$), depends thus only on its last predecessor and it includes a *stochastic noise* or *innovations* term η_t in the following way:

$$X_t = aX_{t-1} + \eta_t. \tag{2.23}$$

If not explicitely stated otherwise, we assume throughout this thesis the driving noise term η_t to be a zero mean Gaussian white noise process with variance σ_η^2: $\eta_t = \mathcal{WN}(0, \sigma_\eta^2)$. Figure 2.2 shows a realisation and the corresponding periodogram of an AR$[1]$ process.

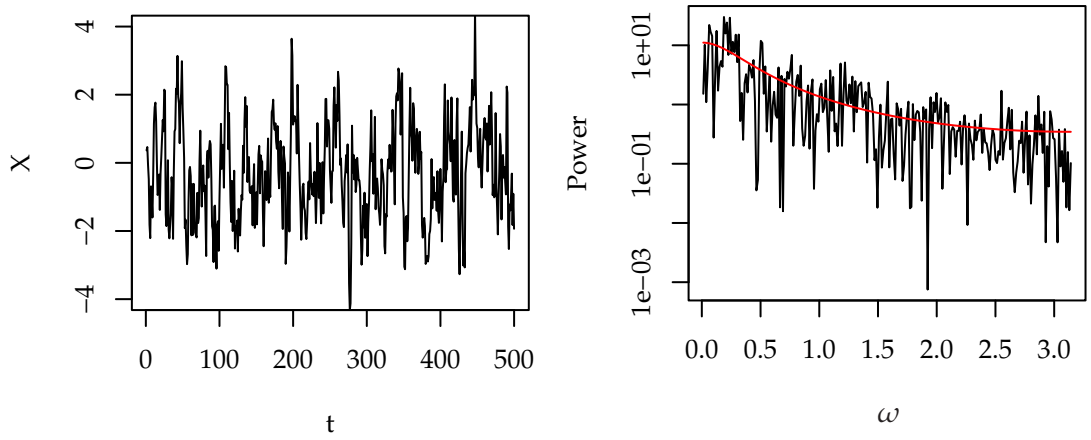

Figure 2.2: Realisation of an AR[1] process with 500 points (left) and the corresponding periodogram with the spectral density as red solid line (right). The autoregressive parameter is set to $a = 0.7$.

Mean and Autocovariance function Assuming $X_0 = 0$ the derivation of the mean and autocovariance function for this process is straightforward. Iteratively using (2.23) yields

$$\begin{aligned} X_t &= aX_{t-1} + \eta_t \\ &= a(aX_{t-2} + \eta_{t-1}) + \eta_t \\ &\vdots \\ &= a^{t-1}\eta_1 + a^{t-2}\eta_2 + \cdots + a^2\eta_{t-2} + a\eta_{t-1} + \eta_t. \end{aligned} \tag{2.24}$$

Taking expectation values on both sides and using $\mathrm{E}[\eta_t] = 0$ yields $\mathrm{E}[X_t] = 0$. Using this result, the autocovariance function for $\tau \geq 0$ is

$$\begin{aligned} \mathrm{cov}(X_t, X_{t+\tau}) &= \mathrm{E}\left[(a^{t-1}\eta_1 + a^{t-2}\eta_2 + \cdots + \eta_t)(a^{t+\tau-1}\eta_1 + a^{t+\tau-2}\eta_2 + \cdots + \eta_{t+\tau})\right] \\ &= \sigma_\eta^2(a^{2(t-1)+\tau} + a^{2(t-2)+\tau} + \cdots + a^\tau) \\ &= \begin{cases} \sigma_\eta^2 t, & \text{if} |a| = 1 \\ \sigma_\eta^2 a^\tau \left(\frac{1-a^{2t}}{1-a^2}\right), & \text{else.} \end{cases} \end{aligned} \tag{2.25}$$

Stationarity Setting $|a| < 1$ leads to $\mathrm{cov}(X_t, X_{t+\tau})$ being asymptotically (i.e. for large t) independent of t and thus the process is called asymptotically stationary. For $a = 1$ we obtain a non-stationary process, i.e. with $X_0 = 0$

$$X_t = X_{t-1} + \eta_t = X_{t-2} + \eta_{t-1} + \eta_t = \cdots = \sum_{i=0}^{t} \eta_i \,. \tag{2.26}$$

This process is known as the *random walk* (Feller, 1968). In the limit of small step sizes, the random walk approximates *Brownian motion*.

Higher Order Autoregressive Processes (AR$[p]$)

The autoregression in (2.23) can be straightforwardly extended to include more regressors with larger time lags. A general definition of an AR$[p]$ process with p regressors is

$$X_t = \sum_{i=1}^{p} a_i X_{t-i} + \eta_t, \tag{2.27}$$

with $\eta_t \sim \mathcal{WN}(0, \sigma_\eta)$ being white noise. Rearranging the terms and using the back-shift operator $BX_t = X_{t-1}$, (2.27) reads

$$(1 - a_1 B - a_2 B^2 - \cdots - a_p B^p) X_t = \eta_t. \tag{2.28}$$

It is convenient to define the autoregressive polynomial of order p

$$\Phi(z) = (1 - a_1 z - a_2 z^2 - \cdots - a_p z^p). \tag{2.29}$$

Now, (2.27) reads simply

$$\Phi(B) X_t = \eta_t. \tag{2.30}$$

Deriving the condition for asymptotic stationarity of an AR$[p]$ process is somewhat more intricate. It leads to investigating whether the roots of $\Phi(z)$, with $z \in \mathbb{C}$, lying outside the unit circle of the complex plane (cf. Priestley, 1992; Brockwell and Davis, 1991).

The general form of the autocorrelation function is given by

$$\rho(\tau) = A_1\phi_1^{|\tau|} + A_2\phi_2^{|\tau|} + \cdots + A_p\phi_p^{|\tau|}, \tag{2.31}$$

with ϕ_i being the reciprocals of the roots of $\Phi(z)$:$\{\phi_i \in \mathbb{C} \mid \Phi(1/\phi_i) = 0\}$. For real values $\phi_i \in \mathbb{R}$ the terms in (2.31) decay exponentially and the asymptotic behaviour for large τ is thus also an exponential decay. Complex values $\phi_i \in \mathbb{C}$ with non-zero imaginary part lead to damped oscillatory terms indicating a "pseudo-periodic" behaviour of X_t. A realisation of an AR[2] process with $a_1 = 0.4$ and $a_2 = -0.8$ is depicted in Figure 2.3 together with the corresponding periodogram. The two reciprocal roots of the AR poly-

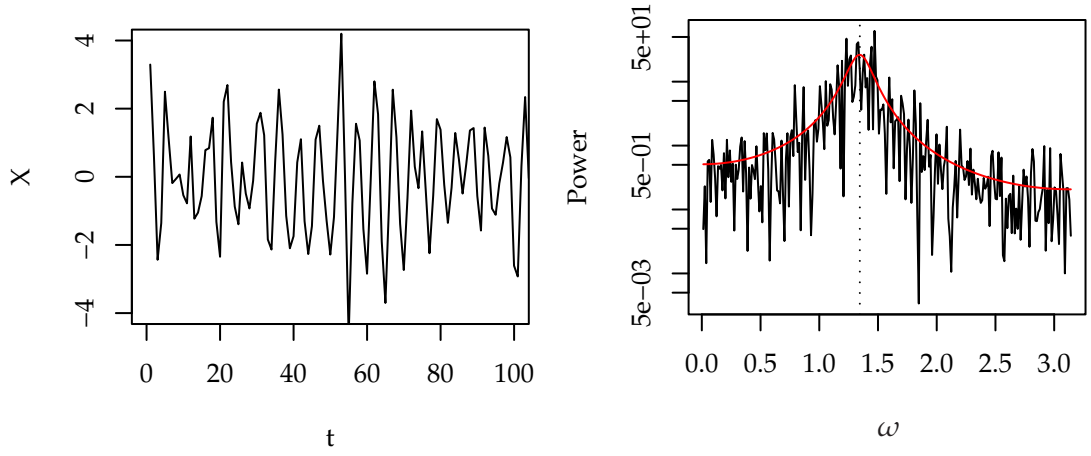

Figure 2.3: Realisation (left) and spectral density and periodogram (right) of an AR[2] process with a pseudo-periodic behaviour. The dotted vertical line at $\omega = 1.35$ marks the predominant frequency of the oscillating behaviour .

nomial are $\phi_1 = 0.2 - 0.87i$ and $\phi_2 = 0.2 + 0.87i$. The argument of the complex number $\arg(\phi_1) = 1.35$ gives the dominant frequency of the oscillating behaviour. The absolute value $|\phi_1| = 0.89$ is related to the damping of the oscillation. A characteristic time scale is given by $T_1 = -1/\ln|\phi_1| = 8.96$.

Moving Average Processes (MA$[q]$)

A different class of linear processes is based on the idea of averaging random shocks η_t which occur independently in time, e.g., $\eta_t \sim \mathcal{WN}(0, \sigma_\eta)$. Such processes can be written as

$$X_t = \eta_t + b_1\eta_{t-1} + \cdots + b_q\eta_{t-q} \tag{2.32}$$

and are called moving average processes of order q or MA$[q]$ processes. Analogously to AR$[p]$ process we can use the back-shift operator B for a convenient notation:

$$X_t = \Psi(B)\eta_t, \tag{2.33}$$

involving the moving average polynomial

$$\Psi(z) = (1 + b_1z + b_2z^2 + \cdots + b_qz^q). \tag{2.34}$$

The MA process (2.32) is thus a special case of the general linear process

$$X_t = \sum_{i=-\infty}^{\infty} b_iB^i\eta_t, \tag{2.35}$$

which allows also for future shocks to be included. In contrast, the MA$[q]$ process defined in (2.32) includes only those terms involving past and present shocks η_{t-i} with $i \geq 0$, leading to a *causal process*.

Autocovariance Different from AR processes, where X_t is a linear combination of its predecessors and a random shock, a variable X_t of a MA process depends only on a finite amount of past and present random shocks. This leads to a cut-off in the autocorrelation function for MA$[q]$ processes for lags larger than q, whereas the ACF of AR$[p]$ processes gradually fades to zero (2.31).

Stationarity and Invertibility Being a linear combination of independent identically distributed random variables η_t, the MA process is trivially stationary for any choice of the parameters b_i. A more interesting characteristic of these processes is the invertibility. In case (2.33) can be written as

$$\eta_t = \Psi^{-1}(B)X_t, \tag{2.36}$$

the process is called *invertible*. Invertibility requires the complex zeros of $\Psi(z)$ lying outside the unit circle (cf. Brockwell and Davis, 1991).

ARMA and ARIMA Processes

The combination of AR$[p]$ and MA$[q]$ processes leads to the concept of ARMA$[p,q]$ processes

$$X_t = \sum_{i=1}^{p} a_i X_{t-i} + \sum_{j=0}^{q} b_j \eta_{t-j}. \tag{2.37}$$

A convenient notation using the back-shift operator is

$$\Phi(B)X_t = \Psi(B)\eta_t \tag{2.38}$$

with $\Phi(z)$ and $\Psi(z)$ being the AR and MA polynomials, respectively. If $\Phi(z)$ and $\Psi(z)$ have no common zeros, we can write (2.38) as

$$X_t = \sum_{i=0}^{\infty} \gamma_i \eta_{t-i}, \tag{2.39}$$

with

$$\Gamma(z) = \sum_{i=1}^{\infty} \gamma_i z^i = \Psi(z)/\Phi(z) \tag{2.40}$$

being the quotient of the MA and AR polynomial. Such ARMA$[p,q]$ processes are *invertible* if the complex roots of $\Psi(z)$ lie outside the unit circle. Analogously to AR processes, they are stationary if we find the roots of $\Phi(z)$ outside the unit circle. As a linear stationary stochastic process, an ARMA$[p,q]$ process is completely defined by its autocovariance function $\mathrm{cov}(\tau)$ or, equivalently, in terms of the spectral density.

Despite the general nature of this formulation, ARMA$[p,q]$ models cannot describe all stationary linear processes. Their generality can be compared to that of a rational function. This analogy arises from their spectral density function being essentially the quotient of $\Psi(\mathrm{e}^{i\omega})$ and $\Phi(\mathrm{e}^{i\omega})$:

$$S(\omega) = \frac{\sigma_\eta^2}{2\pi} \frac{|\Psi(\mathrm{e}^{i\omega})|^2}{|\Phi(\mathrm{e}^{i\omega})|^2}, \tag{2.41}$$

and thus a rational function in the spectral domain.

Integrated Processes Box and Jenkins (1976) included a class of non-stationary processes Y_t in the framework, namely those being partial sums of stationary ARMA$[p,q]$ processes X_t: $Y_t = \sum_{i=0}^{t} X_i$. Using the difference operator[5] $(1-B)^d$, we can also obtain X_t as the *increment process* of Y_t writing

$$X_t = (1-B)^d Y_t, \tag{2.42}$$

with d being at first an integer value, later we allow $d \in \mathbb{R}$. Now, equation (2.38) can be generalised to

$$\Phi(B)(1-B)^d Y_t = \Psi(B)\eta_t \tag{2.43}$$

including now also integrated processes. Since Y_t is the outcome of the sum or integration of X_t such processes are called integrated ARMA$[p,q]$ or ARIMA$[p,d,q]$ processes, with the integer d specifying the degree of integration.

2.2.2 Processes with Long-Range Dependence

In the following, we introduce two representatives of long-range dependent processes. The first example, the fractional Gaussian noise (fGn), has its roots in the fractal community and has been made popular in the 1960s by the French mathematician Benoît Mandelbrot. It was the first mathematical model to describe long-range dependence. About two decades later Granger and Joyeux (1980) and Hosking (1981) proposed a different way to describe this phenomenon. They introduced the concept of fractional differencing leading to fractional differenced processes (FD) and fractional ARIMA processes. Further examples of models for long-range dependence can be found in, e.g., Beran (1994) or Doukhan et al. (2003).

Increments of Self-Similar Processes

In the 1960s Mandelbrot introduced self-similar stochastic processes, actually dating back to Kolmogorov (1941), in the statistics community (Mandelbrot and Wallis, 1968a). A process Y_t is called self-similar if $Y_t \overset{d}{=} c^{-H} Y_{ct}$, where $a \overset{d}{=} b$ means a and b are equal in distribution. H is called the self-similarity parameter or Hurst exponent and c is a positive constant. To model data with a stationary appearance one considers the stationary increment process $X_t = Y_t - Y_{t-1}$ of the self-similar process Y_t. Following Beran (1994), its autocorrelation function reads

$$\rho(\tau) = \frac{\sigma}{2}\left[(\tau+1)^{2H} - 2\tau^{2H} + (\tau-1)^{2H}\right], \tag{2.44}$$

with asymptotic behaviour ($\tau \to \infty$) resulting from a Taylor expansion

$$\rho(\tau) \to H(2H-1)\tau^{2H-2}, \quad \text{for } \tau \to \infty. \tag{2.45}$$

Thus, for $0.5 < H < 1$ the variance decays algebraically and the increment process X_t is long-range dependent. For $0 < H < 0.5$ the ACF is summable; it even sums up to zero, resulting in SRD. This "pathologic" situation is rarely encountered in practice; it is mostly the result of over-differencing (Beran, 1994).

The increments of self-similar processes are thus capable of reproducing LRD and offered an early description of the Hurst phenomenon. The well-known *fractional Brownian motion* is a Gaussian representative of a self-similar process. Its increment series is a long-range dependent stationary Gaussian process with asymptotic properties as given in (2.45) and is referred to as *fractional Gaussian noise* (fGn).

[5] For $d = 1$, $(1-B)X_t = X_t - X_{t-1}$. This can be regarded as the discrete counterpart of the derivative.

Fractional Difference Processes (FD)

After Mandelbrot promoted self-similar processes, it were Granger and Joyeux (1980) and Hosking (1981) who formulated a different model to describe long-range dependence. This model fits well into the framework of the linear processes discussed in Section 2.2.1. They allowed the difference parameter d in (2.42) to take real values. This results in a fractional difference filter

$$(1-B)^d X_t = \eta_t, \tag{2.46}$$

with $d \in \{d \in \mathbb{R} \mid -1/2 < d < 1/2\}$ and η_t being white noise. The fractional difference operator is defined using an infinite power series

$$(1-B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-1)^k B^k, \tag{2.47}$$

with the binomial coefficient

$$\binom{d}{k} = \frac{\Gamma(d+1)}{\Gamma(k+1)\Gamma(d-k+1)}, \tag{2.48}$$

and $\Gamma(x)$ denoting the gamma function. The process X_t has the asymptotic properties of a long-range dependent process as given in (2.21) and is referred to as *fractional difference process* (FD). For $d > 1/2$ the second moment is not finite anymore. In such cases the increment process $X = (1-B)Y$ can be described as a stationary FD process.

Fractional ARIMA Processes (FARIMA$[p,d,q]$)

The FD process has been formulated in the framework of linear models. Although the power series expansion of $(1-B)^d$ is infinite, it can be included into the concept of ARIMA models leading to fractional ARIMA or FARIMA[6] processes. Using the notation of (2.43), we get

$$\Phi(B)(1-B)^d X_t = \Psi(B)\eta_t, \tag{2.49}$$

with $\Phi(z)$ and $\Psi(z)$ being again the autoregressive and moving average polynomials defined in (2.29) and (2.34). Now, we allow d being a real number with $d \in \{d \in \mathbb{R} \mid -1/2 < d < 1/2\}$. The spectral density of FARIMA processes can be obtained from the corresponding result for ARMA processes (2.41), multiplied by $|1-\mathrm{e}^{i\omega}|^{-2d}$:

$$S(\omega) = \frac{\sigma_\eta^2}{2\pi} \frac{|\Psi(\mathrm{e}^{i\omega})|^2}{|\Phi(\mathrm{e}^{i\omega})|^2} |1-\mathrm{e}^{i\omega}|^{-2d}. \tag{2.50}$$

For small frequencies $\omega \to 0$ the limiting behaviour for the spectral density is given by

$$S(\omega) \approx \frac{\sigma_\eta^2}{2\pi} \frac{|\Psi(1)|^2}{|\Phi(1)|^2} |\omega|^{-2d} = c_f |\omega|^{-2d}. \tag{2.51}$$

With $2d = \beta$ (2.51) recovers the asymptotic behaviour required for a long-range dependent process given by (2.22). The relation to the Hurst exponent is $2d = \beta = 2H - 1$.

[6] Sometimes referred to as ARFIMA processes

2.2.3 Motivation for Autoregressive Moving Average Models

The ARMA$[p,q]$ model class presented in the preceeding sections has a flexibility or generality which can be compared to that of a rational function as can be seen from (2.41). Including fractional differencing (2.50), we obtain the FARIMA$[p,d,q]$ class and this rational function is supplemented by a term converging for small frequencies to a power-law, cf. (2.51). This allows for flexible modelling of the spectral density (or ACF) including LRD. Because here the main goal is a suitable description of the ACF and especially the detection of LRD, the latter is a motivation for using FARIMA$[p,d,q]$ processes.

Besides flexibility and feasibility, there are physically based arguments for autoregressive moving average models. Quite generally autoregressive processes can be regarded as discretised linear ordinary differential equations including a stochastic noise term. This allows to describe relaxations and oscillations and offers thus a framework for modelling many physical systems. In Chapter 6 we motivate the use of ARMA$[p,q]$ models for river run-off with a conceptional water balance model.

2.3 Parameter Estimation for Stochastic Processes

The model building process consists of two parts: the inference of a proper model structure and the estimation of the unknown model parameters. A thorough discussion of the model selection is presented in the subsequent chapter. Here we assume that a suitable model is known and present parameter estimation strategies for the stochastic processes (Section 2.2.2), particularly a maximum likelihood approach to FARIMA$[p,d,q]$ processes.

Besides the likelihood approach there are also other ways to estimate the fractional difference parameter. We discuss the detrended fluctuation analysis (DFA) as a frequently used heuristic method[7] in the following, while the description of the rescaled range analysis and the the semi-parametric[8] log-periodogram regression is deferred to Appendices B.3.1 and B.3.2, respectively. These semi-parametric and heuristic approaches are formulated on the basis of the asymptotic behaviour and focus on estimating the Hurst exponent H or the fractional difference parameter d to quantify long-range dependence. It is not intended to include the high frequency behaviour in the description, it is rather attempted to reduce its influence on the estimation of H or d.

A section on generating realisations of a FARIMA$[p,d,q]$ process concludes this chapter.

2.3.1 Maximum Likelihood for FARIMA$[p,d,q]$ Processes

Consider a zero mean stationary FARIMA$[p,d,q]$ process. As a Gaussian process, it is fully specified by the autocovariance matrix $\mathbf{\Sigma}(\boldsymbol{\theta})$, which in turn depends on the process parameters $\boldsymbol{\theta} = (d, a_1, \ldots, a_p, b_1, \ldots, b_p, \sigma_\eta)$. We can thus formulate the probability density for obtaining a realisation $\mathbf{x}$ as

$$p(\mathbf{x}|\boldsymbol{\theta}) = (2\pi)^{-\frac{N}{2}} |\mathbf{\Sigma}(\boldsymbol{\theta})|^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}\mathbf{x}^\dagger \mathbf{\Sigma}^{-1}(\boldsymbol{\theta})\mathbf{x}}, \tag{2.52}$$

[7] As heuristic we denote an estimator without an established limiting distribution. The specification of confidence intervals and thus statistical inference is not possible.

[8] We call an approach semi-parametric if only the behaviour for large scales (small frequencies) is described. Contrary to a heuristic approach, the limiting distribution for the estimator is available.

with $\mathbf{x}^\dagger$ denoting the transposed of $\mathbf{x}$. For a given realisation $\mathbf{x}'$, the same expression can be interpreted as the likelihood for the parameter vector $\boldsymbol{\theta}$

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{x}') = p(\mathbf{x}'|\boldsymbol{\theta}). \tag{2.53}$$

This suggests the formulation of an estimator for $\boldsymbol{\theta}$ based on a realisation $\mathbf{x}'$: the estimate $\widehat{\boldsymbol{\theta}}$ is chosen such that (2.53) is maximised. It is convenient to use the logarithm of the likelihood, $L(\boldsymbol{\theta}|\mathbf{x}') = \log \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}')$. Due to its monotonicity, the estimator can then be formulated as the argument of the maximum log-likelihood

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathbf{x}'). \tag{2.54}$$

Asymptotic Properties of the Maximum Likelihood Estimator

For causal and invertible Gaussian processes and an increasing sample size $N \to \infty$ the maximum-likelihood estimator (MLE) is unbiased and converges to the true parameter vector $\boldsymbol{\theta}^0$ almost surely (Yajima, 1985; Dahlhaus, 1989). Furthermore, the estimator converges in distribution

$$N^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \xrightarrow{d} \xi \tag{2.55}$$

to a Gaussian random variable $\xi \sim \mathcal{N}(0, \mathbf{V}_{\boldsymbol{\theta}^0})$, with $\mathbf{V}_{\boldsymbol{\theta}^0} = 2\mathbf{D}^{-1}(\boldsymbol{\theta}^0)$ being the covariance matrix for the parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_M)^\dagger$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^0$. The $M \times M$ matrix $\mathbf{D} = D_{ij}(\boldsymbol{\theta}^0)$ is defined by

$$D_{ij}(\boldsymbol{\theta}^0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\partial}{\partial \theta_i} \log S(\omega; \boldsymbol{\theta}) \frac{\partial}{\partial \theta_j} \log S(\omega; \boldsymbol{\theta}) \mathrm{d}\omega|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0}. \tag{2.56}$$

The rate of convergence of $\widehat{\boldsymbol{\theta}}$ is $N^{-1/2}$ as expected also for the short-range dependent case. This is particularly interesting since some estimates exhibit a slower rate of $N^{-\alpha}$ in the presence of long-range dependence as, e.g., regression parameters or the mean value as shown in (B.6).

Dahlhaus (1989) established asymptotic efficiency of the MLE for long-range dependent processes [9]. This implies minimal variance (as specified by the Cramér-Rao bound), or, equivalently, the Fisher information matrix

$$\boldsymbol{\Gamma}_n(\boldsymbol{\theta}^0) = \mathrm{E}\left[[\mathbf{L}'(\widehat{\boldsymbol{\theta}}|\mathbf{x})][\mathbf{L}'(\widehat{\boldsymbol{\theta}}|\mathbf{x})]^\dagger\right] \tag{2.57}$$

converges to the inverse of the estimators covariance matrix

$$\lim_{n\to\infty} \boldsymbol{\Gamma}_n(\boldsymbol{\theta}^0) = \mathbf{V}_{\boldsymbol{\theta}^0}^{-1}. \tag{2.58}$$

The maximum likelihood estimation (2.54) is formulated as a nonlinear optimisation problem which can be solved numerically. This requires inverting the covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ for every optimisation step in the parameter space. This is a costly and potentially unstable procedure and thus not feasible for long records. A convenient approximation to the MLE has been proposed by Whittle (1953) and is described in the following.

[9] See also the correction notes at
`http://math.uni-heidelberg.de/stat/people/dahlhaus/ExtendedCorrNote.pdf`

2.3.2 Whittle Estimator

The main idea of Whittle (1953) was to give suitable approximations for the terms in the log-likelihood

$$L(\boldsymbol{\theta}|\mathbf{x}') = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \frac{1}{2}\mathbf{x}^{\dagger}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\mathbf{x} \tag{2.59}$$

which are dependent on $\boldsymbol{\theta}$: the determinant $\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})|$ and $\mathbf{x}^{\dagger}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\mathbf{x}$. For both terms the approximation involves an integral over the spectral density $S(\omega;\boldsymbol{\theta})$ of the process which is in a successive step approximated by a Riemann sum. Finally, an appropriate rescaling of the spectral density $S(\omega;\boldsymbol{\theta}) = \theta_1 S(\omega;\boldsymbol{\theta}^*)$ with $\boldsymbol{\theta}^* = (1,\theta_2,\ldots,\theta_M)^{\dagger}$ and $\theta_1 = 2\pi\sigma_\eta^{-2}$ yields a discrete version of the Whittle estimator:

1. Minimise
$$Q(\boldsymbol{\theta}^*) = \sum_{j=1}^{[(N-1)/2]} \frac{I(\omega_j)}{S(\omega_j;\boldsymbol{\theta}^*)} \tag{2.60}$$
with respect to $\boldsymbol{\theta}^*$.

2. Set
$$\hat{\sigma}_\eta^2 = 2\pi\hat{\theta}_1 = \frac{4\pi}{n}Q(\boldsymbol{\theta}^*). \tag{2.61}$$

$I(\omega_j)$ denotes the periodogram of the realisation $\mathbf{x}'$ at the Fourier frequencies $\omega_j = \frac{2\pi j}{N}$ with $j = 1,\ldots,\left[\frac{N-1}{2}\right]$ where $[.]$ denotes the integer part. A detailed derivation of the discrete Whittle estimator for LRD processes has been given by Beran (1994).

Asymptotic Properties of the Whittle Estimator

It can be shown, that the Whittle aproximation has the same asymptotic distribution as the exact ML estimator (Beran, 1994). Therefore, it is asymptotically efficient for Gaussian processes. The assumption of a zero mean process has been made for simplification of the representation, the asymptotic result does not change if the mean is consistently estimated and substracted (Beran, 1995). Besides simplifying the optimisation, the choice of the scale factor brings about a further convenience: the estimate of the innovations variance $\sigma_\eta^2 = 2\pi\theta_1$ is asymptotically independent (i.e. for $N \to \infty$) of the other parameter estimates.

Confidence Intervals

The asymptotic distribution (2.55) can be used to obtain approximate $\alpha 100\%$ confidence regions for the parameter estimate $\widehat{\boldsymbol{\theta}}$ as

$$\mathrm{CI}_\alpha = \{\boldsymbol{\theta} \in \mathbb{R}^{(p+q+1)} \mid (\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})^{\dagger}\mathbf{V}^{-1}(\widehat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}) \leq N^{-1}\chi^2_{(p+q+1),\alpha}\}, \tag{2.62}$$

with $\chi^2_{m,\alpha}$ specifying the α-quantile of the χ^2 distribution with m degrees of freedom, where m is the number of parameters (Brockwell and Davis, 1991). For a single parameter estimate $\widehat{\theta}_j$ we can also obtain approximate confidence intervals using the standard normal distribution. Writing the jth diagonal element of the estimator's covariance matrix $\mathbf{V}(\widehat{\theta})$ as v_{jj} we obtain

$$\mathrm{CI}_\alpha = \{\theta \in \mathbb{R} \mid |\theta - \widehat{\theta}_j| \leq N^{-1/2}\Phi_{((1+\alpha)/2)}v_{jj}^{1/2}\}, \tag{2.63}$$

with $\Phi_{((1+\alpha)/2)}$ being the $(1+\alpha)/2$-quantile of the standard normal distribution (Beran, 1994).

An alternative variant is to use a bootstrap approach. This approach consists of generating an ensemble of time series of original length using the model obtained for the empirical series and a subsequent parameter estimation from all ensemble members. Confidence intervals can then be obtained from the empirical frequency distribution of the estimates following Davison and Hinkley (1997). A variant based on resampling the residuals can be found in Hipel and McLeod (1994). Bootstrap approaches are especially useful if the asymptotic confidence intervals are not reliable, e.g., if the residual distribution is not Gaussian.

Non-stationary Long-Range Dependence

So far, we have assumed a fractional difference parameter in the stationary range $-1/2 < d < 1/2$. Velasco and Robinson (2000) showed that the Whittle estimation presented here can be extended to non-stationary processes with $1/2 \leq d < 1$. It is consistent for $d < 1$ and preserves its asymptotic normality for $d < 3/4$. For larger d the asymptotic normality can be recovered using a cosine bell taper for the calculation of the periodogram. For such a non-stationary process we cannot easily express the likelihood as in (2.53). We can, however, calculate an approximate log-likelihood using the ML-estimate of the residuals variance $\widehat{\sigma}_\epsilon^2$ as

$$L(\widehat{\boldsymbol{\theta}}|\mathbf{x}) \approx -\frac{N}{2}\log 2\pi - \frac{N}{2}\log \widehat{\sigma}_\epsilon^2. \tag{2.64}$$

Performance

Smith et al. (1997) performed an extensive simulation study on bias and variance of the fractional integration parameter and the autoregressive and moving average parameters in FARIMA$[p,d,q]$ models using ML estimation. Given the correct model the ML estimation outperforms the log-periodogram regression. Mis-specifying the model orders results in biased estimates of d, as well as of the autoregressive parameters. Similar results have been obtained by Taqqu and Teverovsky (1997) investigating three types of Whittle estimators. The estimator proposed here is the most accurate of those under investigation provided the correct model is chosen. This stresses the need of reliable model selection (Chapter 3).

Numerical Aspects

In contrast to moment based estimators, such as the Yule-Walker equations or the innovations algorithms (e.g., Brockwell and Davis, 1991) the optimisation problem in (2.54) is generally carried out numerically. It is thus burdened with problems inherent to non-linear optimisation, for instance, getting trapped in local minima. This problem can be partially surmounted by initialising the algorithm with preliminary estimates. For the AR or ARMA components, these estimates can be obtained using the Yule-Walker equations, the innovations algorithm, or an estimate obtained by minimising the conditional sum of squares (Brockwell and Davis, 1991; Schumway and Stoffer, 2006).

The optimisation routines used in the following are those implemented in `R`'s[10] `optim` routine, namely the simplex method (Nelder and Mead, 1965) or, alternatively, a quasi-

[10] `R` is a software package for statistical computing and freely available from `http://www.r-project.org/` (R Development Core Team, 2004)

Newton algorithm (Press et al., 1992). While the latter is faster, the simplex method is more robust. For box-constrained numerical optimisation, the algorithm by Byrd et al. (1995) is available.

Implementation

The numerical algorithm given in `S-Plus` by Beran (1994) and translated to `R` by Martin Mächler was used in a modified and supplemented form available as the `R`-package `farisma`[11]. Extensions are made with respect to the initial guesses of the parameter values and tools for visualisation, simulation and model selection were supplemented.

2.3.3 Detrended Fluctuation Analysis – A Heuristic Approach

The first attempts to describe the long-range dependent phenomenon were made before suitable stochastic models had been developed. Starting with the rescaled range statistic (cf. Appendix B.3.1) proposed by Hurst in 1951, several heuristic methods have emerged to estimate the Hurst coefficient H. The approaches are based, for example, on a direct calculation of the autocovariance sequence or on studying the behaviour of the variance of the mean of subsamples with increasing length in a log-log plot. A relatively recent approach investigates the variance of residuals of a regression in subsamples and is thus to a certain extend robust to instationarities. This approach has become widely used in the physics community and is known as detrended fluctuation analysis (DFA) or residuals of regression. Other heuristic methods are discussed by, e.g., Beran (1994) or Taqqu et al. (1995). These heuristic methods are not suitable for statistical inference about the long-range dependence parameter, but are rather diagnostic tools to start with. Confidence intervals cannot be obtained straightforwardly which renders the interpretation of the result difficult. For statistical inference one can either consider the log-periodogram regression (Appendix B.3.2) or the full parametric modelling approach presented above (Beran, 1994).

Residuals of regression or detrended fluctuation analysis (DFA) was developed by Peng et al. (1993, 1994) while studying DNA nucleotides and later heartbeat intervals (e.g., Peng et al., 1995). Their aim was to investigate for long-range dependent processes underlying these records excluding the influence of a possible trend. Direct estimation of the autocorrelation function from empirical data is limited to rather small time lags s and is affected by observational noise and instationarities like trends. DFA received a lot of attention in recent years and became a frequently used tool for time series analysis with respect to LRD (e.g., Kantelhardt et al., 2001). Many works investigating for LRD in various fields of research are based on this method. Some examples are empirical temperature records (e.g., Koscielny-Bunde et al., 1996, 1998; Monetti et al., 2001; Eichner et al., 2003; Blender and Fraederich, 2003), or temperature series from climate models (e.g., Bunde et al., 2001; Govindan et al., 2002), run-off (e.g., Koscielny-Bunde et al., 2006) and wind speed records (e.g., Govindan and Kantz, 2004). Also with respect to extreme events DFA has been used (e.g., Bunde et al., 2005). This popularity is probably due to the simplicity of the method. There are, however, difficulties and pitfalls which are easily overlooked when applying DFA. The latter are described in more detail in Appendix A.2.

For a description of the method, consider again the time series $\{X_i\}_{i=1,\ldots,N}$. Similar to Hurst's rescaled range analysis (cf. Appendix B.3.1), the detrended fluctuation analysis

[11] `http://www.pik-potsdam.de:~hrust/tools/`

is based on the aggregated time series $Y(t) = \sum_{i=1}^{t} X_i$. First, divide $Y(t)$ into M non-overlapping segments of length s. Then, for DFA of order n (DFAn), in each segment $m = 1, \ldots, M$ fit a polynomial of order n. This polynomial trend $p_{s,m}^{(n)}$ is subtracted from the aggregated series:

$$Y_{s,m}(t) = Y(t) - p_{s,m}^{(n)}(t). \tag{2.65}$$

For every segment m the squared fluctuation is calculated as

$$F_m^2(s) = \frac{1}{s} \sum_{t=(m-1)s+1}^{ms} Y_{s,m}(t)^2. \tag{2.66}$$

Averaging over all segments $m = 1, \ldots, M$ yields the squared fluctuation function for the time scale s:

$$F^2(s) = \frac{1}{M} \sum_{m=1}^{M} F_m^2(s). \tag{2.67}$$

This procedure is repeated for several scales s. The time scales s are limited at the lower bound by the order n of DFA and at the upper bound by the length of the record. Due to the increasing variability of $F^2(s)$ with s, a reasonable choice for the maximum scale is $s_{\max} \approx N/10$. The fluctuation function $F(s)$ is then investigated in a double logarithmic plot.

Interpretation of the Slope

Using the asymptotic behaviour of fractional Gaussian noise (fGn) and FARIMA processes, Taqqu et al. (1995) showed that the resulting fluctuation function obtained with DFA is asymptotically proportional to s^H, with H being the Hurst exponent. Thus the asymptotic slope in the log-log plot yields an estimate $\widehat{H}_{\mathrm{DFA}n}$ (or, equivalently, $\widehat{d}_{\mathrm{DFA}n}$ (2.49)) of the Hurst exponent (or fractional difference parameter)(Taqqu et al., 1995). For increments of self-similar processes, as fGn, the asymptotic behaviour is already reached for very moderate sample size. For those processes, one can choose to fit a straight line in the range $1 < \log s < \log N/10$. Values for $\log s > \log N/10$ are frequently excluded due to a large variability.

Asymptotically, we can relate the Hurst exponent H to the exponent γ quantifying the algebraic decay of the ACF (2.21) by

$$H = 1 - \gamma/2\,, \quad \text{with} \quad 0.5 < H < 1. \tag{2.68}$$

It can be as well related to the fractional difference parameter (2.49) by $d = H - 0.5$ (cf. Section 2.2.2; Taqqu et al., 1995). For an uncorrelated process we get a squared fluctuation function $F^2(s) \propto s$ (i.e. $F(s) \propto s^{0.5}$) which reflects the linear increase of the variance of the aggregated series $Y(t)$ with t (Beran, 1994).

More details on the basics of DFA can be found in (Kantelhardt et al., 2001). Hu et al. (2001) and Chen et al. (2002) investigate extensively the effects of various types of instationarities on DFA using simulation studies with realisations of the self-similar increment process fGn.

Limiting Behaviour, Uncertainty Analysis

Studies regarding the variability of the estimate $\widehat{H}_{\mathrm{DFA}n}$ are rare. Analytical approaches towards a derivation of the limiting distribution could not be found in the literature. While

a first impression of bias and variance can be obtained from the simulations in Taqqu et al. (1995), a more detailed study is given in Appendix A.1. Furthermore, Weron (2002) performed a systematic Monte Carlo study to quantify the variability of $\widehat{H}_{\mathrm{DFA}n}$ based on realisations of Gaussian white noise. He derived empirical confidence intervals which can be used for testing the null hypothesis of a Gaussian white noise process. Weron estimated the slope of the logarithmic fluctuation function for scales $s > 10$ and $s > 50$. For many practical applications the asymptotic behaviour has not been reached for such small scales. This implies that suitable conditions to transfer these confidence intervals are rare. Furthermore, neither the upper bound for the straight line fit has been specified, nor the number of sampling points s for which the fluctuation function has been calculated. Although Weron's confidence bands might therefore not be suitable for a direct transfer to many practical applications, the idea he presented leads to a promising approach of obtaining confidence intervals using a simulation approach. If it is possible to specify a parametric model for the observed record, such as a FARIMA$[p, d, q]$, confidence intervals can be easily estimated using a parametric bootstrap (Davison and Hinkley, 1997). If, however, such a model has been identified, an estimate including asymptotic confidence intervals for the Hurst coefficient can be derived directly (Section 2.3.1).

2.4 Simulations from Long-Range Dependent Processes

There are several ways of obtaining a realisation of a LRD model. An extensive survey of generators has been conducted by Bardet et al. (2003). Here, we focus on the description of the direct spectral method which was implemented and used in this work. It makes use of some essentials of spectral analysis: a) the relation between the spectrum $S(\omega)$ and the ACF (2.16), b) the periodogram $I(\omega)$ (2.18) as an estimate for the spectral density and c) the sampling properties of $I(\omega)$. The latter are such that the Fourier coefficients a_j and b_j of a Gaussian process $\{X_t\}_{t=1,\ldots,N}$, defined by

$$a_j = N^{-1/2} \sum_{t=1}^{N} X_t \cos(\omega_j t), b_j = N^{-1/2} \sum_{t=1}^{N} X_t \sin(\omega_j t), \tag{2.69}$$

are Gaussian random variables. This implies that the periodogram

$$I(\omega_j) = \frac{1}{2\pi} \begin{cases} a_j^2 + b_j^2, & j = 1, \ldots, (N/2) - 1 \\ a_j^2, & j = 0, N/2 \end{cases} \tag{2.70}$$

is basically the sum of two squared normal variables and thus follows a scaled χ-squared distribution with two degrees of freedom and an expectation value $S(\omega_j; \boldsymbol{\theta})$. The periodogram can be written as a scaled χ-squared distributed random variable with two degrees of freedom (Priestley, 1992)

$$I(\omega_j) = \frac{1}{2} S(\omega_j; \boldsymbol{\theta}) \zeta \text{ , with } \quad \zeta \sim \chi_2^2. \tag{2.71}$$

Exploiting these properties, we can specify Gaussian random numbers a_j and b_j such that (2.71) holds, i.e.

$$a_j, b_j = \frac{1}{2} S(\omega_j; \boldsymbol{\theta}) \xi \text{ , } \qquad \text{for } \quad j = 1, \ldots, (N/2) - 1, \tag{2.72}$$

and

$$a_j = S(\omega_j; \boldsymbol{\theta}) \xi \text{ , } \qquad \text{for } \quad j = 0, N/2, \tag{2.73}$$

where $\xi \sim \mathcal{N}(0,1)$ is a Gaussian random variable. Using further a property of Fourier series $f(\omega_j)$ from real valued records $\{x_t\}_{t=1,\ldots,N}$: $f(-\omega_j) = f^*(\omega_j)$, we obtain a realisation of the process with spectral density $S(\omega_j;\boldsymbol{\theta})$ by the inverse Fourier series of a realisation of $Z_j = a_j + ib_j$ for $j = -N,\ldots,N$. Using the Fast Fourier Transform (FFT), this algorithm is fast at least for N being dyadic (order $\mathcal{O}(N \log N)$) and thus it is particularly interesting for simulating long records.

Besides the effect of aliasing (Priestley, 1992), the periodicity of the realisation is a problem: the end of the record is highly correlated with the beginning. It is thus necessary to generate much longer records and extract a series of needed length from it.

Chapter 3

Model Selection

Entia non sunt multiplicanda praeter necessitatem.
William of Ockham (1295-1349)

The most crucial and delicate step in model building is the choice of an adequate model which satisfactorily describes the data. The notion of a satisfactory description is subjective and depends on the research questions. If physical understanding or explanation is the goal of the modelling efforts, *lex parsimoniae* – the principle of parameter parsimony – may be a guiding principle. It states that among models with the same explanatory power the one with fewer parameters is to be preferred. The principle dates back to the English Franciscan friar William of Ockham and is also known as Ockhams razor. If, on the other hand, prediction is the focus of the modelling effort, useful models in the sense of Box and Jenkins (1976) are those with a high predictive power regardless of a physical interpretation; model complexity – in terms of number of parameters – is secondary in this respect.

The final aim of our modelling effort is the derivation of statistical quantities from time series such as trend (Section 5.1.3) and quantile (Chapter 7) estimates. For a reliable inference, we need the information about the dependence structure. Therefore, it is not the prediction of future observations which is in the focus, it is rather the reliable identification of the ACF. In this respect, we use lex parsimoniae as a guiding principle.

For model choice or model selection, we can identify two different concepts: goodness-of-fit tests and model comparison strategies. With a goodness-of-fit test, the hypothesis is tested that the observed record is a realisation of the model proposed. The second concept compares the fitness of two or more models in order to choose one or a few models out of a set of suitable models. Commonly, model selection involves goodness-of-fit tests and model comparison strategies. A plausible strategy is to base the model comparison on a set of models which pass the goodness-of-fit test. Within the family of the FARIMA$[p, d, q]$ processes, model selection reduces to choosing appropriate model orders p and q, and to deciding whether a fractional difference parameter d is needed or not.

This chapter presents a goodness-of-fit test particularly suitable within the framework of FARIMA$[p, d, q]$ models and Whittle maximum likelihood estimation. Further, we discuss standard model comparison strategies. These are approaches within a test-theoretical setting, such as the likelihood-ratio test, and criteria from information theory, such as the Akaike information criterion. Being confronted with non-nested models[1],

[1]The model g is nested in model f if it constrains one or more parameters of f, typically to zero. This notion is different from nested in the context of climate models.

non-Gaussian processes or simply situations where asymptotic results are not applicable, other than the standard approaches have to be considered. In this context, we take up an approach proposed by Cox (1961) and Hinde (1992) and develop a model selection strategy for non-nested FARIMA$[p, d, q]$ models.

3.1 Goodness-of-Fit Tests

It seems plausible to expect from a "good" model a description of the data such that the difference between data and model output (the residuals) does not exhibit any structure of interest with respect to the modelling task. If the aim is to describe the correlation structure, it is plausible to demand for independent (or at least uncorrelated) residuals. This implies that there is no evidence for more (linear) structure in the underlying dynamics. Consequently many goodness-of-fit tests are based on testing the residuals for compatibility with a white noise process. A fundamental test in this respect is the Portmanteau test. It is based on the sum over the squared autocovariance series of the residuals. Several modifications have been proposed to improve the finite sample properties of the Portmanteau test. A comprehensive discussion can be found in Li (2004). Here, we restrict ourselves to a brief description of the basic Portmanteau statistic and a spectral variant which fits well in the framework of Whittle-based parameter estimation. A short note on hypothesis testing precedes the discussion.

3.1.1 Hypothesis Testing

Hypothesis testing is an algorithm to decide for or against an uncertain hypothesis minimising a certain risk. An example for a hypothesis is the "goodness-of-fit", i.e. H_0 "the observed data is a plausible realisation of the model". The information relevant for this hypothesis is summarised in a test statistic, e.g., the sum of the squared residual autocovariance series. Knowing the distribution of the test statistic under this *null hypothesis* H_0, we choose a *critical region* including those values of the test statistic which we consider as extreme and as evidence against the hypothesis. The probability of finding the test statistic under the assumption of H_0 in this critical region is called the *α-value* or *size* of the test. Common α-values are 0.01 and 0.05 corresponding to a 1% and 5%-level of significance.

The probability of the test statistic falling into the critical region under an alternative hypothesis H_A is called the *power* (pow) of the test; $1 - \text{pow}$ is referred to as the *β-value*. If the observed value falls inside the critical region, we reject the null hypothesis. If it is found outside we conclude that there is not enough evidence to reject H_0. Note, that this lack of evidence against H_0 does not imply evidence for the null hypothesis (Cox and Hinkley, 1994).

A test with low power cannot discriminate H_0 and H_A and is said to be not sensitive to the alternative hypothesis. For large α-values the test is referred to as not being specific; H_0 is frequently rejected even if it is true. The optimal test would be both, sensitive and specific.

If not stated otherwise, we use a 5%-level of significance in the following.

3.1.2 Portmanteau Test

Let $\{x_t\}_{t=1,\ldots,N}$ be the record under consideration and $f_t(\boldsymbol{\theta})$ the model output at index t. The parameter vector $\boldsymbol{\theta}$ has elements $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^\dagger$. The residual series is denoted as

$r_t = x_t - f_t(\boldsymbol{\theta})$. The Portmanteau or Box-Pierce statistic is

$$Q_l = N \sum_{\tau=1}^{l} \widehat{\rho}^2_{\text{res}}(\tau), \tag{3.1}$$

with $\widehat{\rho}_{\text{res}}(\tau)$ being the autocorrelation sequence of the residuals r_t. Under the null hypothesis H_0 "the residuals are uncorrelated", Q_N is asymptotically χ-squared distributed with $l - m$ degrees of freedom (Box and Jenkins, 1976; Li, 2004).

3.1.3 Spectral Variant of the Portmanteau Test

Beran (1992) suggested a goodness-of-fit test for long-memory processes which is equivalent to the Portmanteau test but formulated in the spectral domain. The test statistic involves quantities which are necessarily calculated during the Whittle parameter estimation, thus it integrates smoothly into the framework. The test is a generalisation of a result from Milhoj (1981) for short memory processes. Define

$$A_N = \frac{4\pi}{N} \sum_j \left(\frac{I(\omega_j)}{\mathcal{S}(\omega_j;\boldsymbol{\theta})} \right)^2 \quad \text{and} \quad B_N = \frac{4\pi}{N} \sum_j \frac{I(\omega_j)}{\mathcal{S}(\omega_j;\boldsymbol{\theta})}, \tag{3.2}$$

with $S(\omega_j;\boldsymbol{\theta})$ being the spectral density of the model under consideration. $I(\omega_j)$ denotes the periodogram and the sums extend over all Fourier frequencies ω_j. The test statistic is now defined as

$$T_N(\widehat{\boldsymbol{\theta}}) = \frac{A_N(\widehat{\boldsymbol{\theta}})}{B^2_N(\widehat{\boldsymbol{\theta}})}. \tag{3.3}$$

It can be shown that $N^{1/2}(A_N(\widehat{\boldsymbol{\theta}}), B_N(\widehat{\boldsymbol{\theta}}))$ converges to a bivariate Gaussian random variable. We can simplify the test considering that $T_N(\widehat{\boldsymbol{\theta}})$ itself is asymptotically normal with mean π^{-1} and variance $2\pi^{-2}N^{-1}$:

$$P(T_N \leq c) \approx \Phi(\pi\sqrt{N/2}(c - (1/\pi))), \tag{3.4}$$

with Φ denoting the standard normal distribution function. Milhoj (1981) showed numerically that this approximation is acceptable already for moderate sample size of $N \approx 128$.

3.2 Model Comparison

A central idea in model selection is the comparison of the model residual variances, residual sums of squares or likelihoods. Intuitively, it is desirable to obtain a small residual variance or a large likelihood. However, increasing the number of parameters trivially reduces the variance. Consequently, one has to test whether a reduction in residual variance due to an additional parameter is random or a significant improvement. An improvement is called random, if the effect could have been induced by an arbitrary additional parameter. An effect is called significant if a special parameter improves the quality of the fit in a way which is not compatible with a random effect.

In the following, we present standard procedures from the test theoretical framework, such as the likelihood-ratio test, and criteria from information theory, such as the criteria of Akaike, Hannan and Quinn or Schwarz.

3.2.1 Likelihood-Ratio Test

The likelihood-ratio test (LRT) is based on the ratio of the two likelihood functions for two nested models f and g evaluated at the maximum. We assume that g is nested in f and constrains k-parameters. According to lex parsimoniae, we ask if the models perform equally well and take then the simpler one. The null hypothesis is now H_0 "g is an admissible simplification of f". A convenient asymptotic distribution for the test statistic is obtained using the log-likelihood functions l_f and l_g. The log-likelihood-ratio is then expressed as the difference

$$\mathrm{lr} = -2(l_g - l_f). \tag{3.5}$$

Under H_0 this test statistic is asymptotically χ-squared distributed with k degrees of freedom. Increasing the α-values with the sample size N leads to a consistent model selection strategy, i.e. the probability of identifying the right model approaches one with increasing sample size.

3.2.2 Information Criteria

A different approach was carried forward with a paper by Akaike (1973). It had tremendous influence on the way model comparison was carried out in the 1970s. Hitherto existing strategies relied on eye-balling or heuristic criteria. Due to increasing computational power it became possible to estimate parameters in models involving considerably more than two or three parameters. Therfore, it was now necessary to find a reliable procedure to choose among them.

Akaike Information Criterion

Akaike suggested a simple and easy to apply criterion. It is basically an estimate of the relative Kullback-Leibler distance to the "true" model. The Kullback-Leibler distance, or negentropy, is a measure of distance between a density function $p(\mathbf{x}|\boldsymbol{\theta})$ and the true probability density. It is thus different from the test theoretical approach described above. This measure was initially called "An Information Criterion" which later changed into *Akaike Information Criterion* (AIC). It is defined as

$$\mathrm{AIC} = -2l(\boldsymbol{\theta}|\mathbf{x}) + 2m, \tag{3.6}$$

with $l(\boldsymbol{\theta}|\mathbf{x})$ being the log-likelihood and m the number of parameters of the model under consideration. The term $2m$, frequently referred to as "penalty term", is indeed a bias correction. The model with the smallest AIC, i.e. the smallest distance to the true model, is typically chosen as the best. Within the framework of a multi-model approach, one might also retain all models within 2 of the minimum (cf. Ripley, 2004).

In the decades after it has been proposed AIC enjoyed great popularity in many branches of science. One major reason why it was much more used than, e.g., Mallows' C_p criterion (Mallows, 1973) developed at the same time might be its simplicity. Simulation studies, and later Hannan (1980), showed that the AIC systematically selects too complex models, which led to modified approaches.

Hannan-Quinn Information Criterion

A consistent criterion was later proposed by Hannan and Quinn (1979). The difference to the Akaike criterion is a modified "penalty term" including the number of data points N

$$\mathrm{HIC} = -2l(\boldsymbol{\theta}|\mathbf{x}) + 2mc \log\log N, \tag{3.7}$$

with $c > 1$. This criterion is known as the *Hannan-Quinn information criterion* (HIC).

Bayesian-Schwarz Information Criterion

Another consistent criterion was formulated within the framework of Bayesian modelling by Schwarz (1978):

$$\text{BIC} = -2l(\boldsymbol{\theta}|\mathbf{x}) + m \log N, \tag{3.8}$$

termed Schwarz information criterion or *Bayesian information criterion* (BIC).

3.2.3 Information Criteria and FARIMA$[p, d, q]$ Models

Beran et al. (1998) investigated these three information criteria regarding order selection in FARIMA$[p, d, 0]$ processes and proved consistency for HIC and BIC regarding the selection of the autoregressive order p. An extensive simulation study was conducted by Bisaglia (2002) also in the context of FARIMA$[p, d, q]$ but with nontrivial moving average order q. As a result Bisaglia obtained that HIC and BIC can be quite successful when choosing among the class of FARIMA$[p, d, q]$ processes with $0 \leq p, q \leq 2$ for realisations coming from a FARIMA$[1, d, 1]$ process with $d = 0.3$. The rate of success, i.e. the fraction of correctly identified models, depends, however, heavily on the AR and MA components of the true model. It varies between 0.5% ($N = 500, a_1 = -0.5$ and $b_1 = -0.3$) and 99% ($N = 500, a_1 = 0.9$ and $b_1 = 0$). Although in many cases the rate of success is larger than 70%, in specific cases it can be unacceptable small.

Unfortunately, the information criteria do not provide any measure of performance by themselves. It is, however, desirable to identify situations with a success rate as low as 0.5%. In these situations one might either consult a different selection strategy or, at least, one wishes to be informed about the uncertainty in the selection. In situations which can be related to Bisaglia's simulation studies one might refer to those for a measure of performance. In other situations one has to conduct such studies for the specific situation at hand.

A further problem is that Akaike derived his information criterion for nested models (Akaike, 1973; Ripley, 2004) which makes it inappropriate when asking for a choice between, e.g., a FARIMA$[1, d, 0]$ and an ARMA$[3, 2]$. Reverting to a common model where both variants are nested in (here FARIMA$[3, d, 2]$) and testing the two original models for being admissible simplifications might result in a loss of power (Cox, 1961).

Confronted with these difficulties, we develop a simulation-based model selection approach for FARIMA$[p, d, q]$ processes which is particularly suitable for non-nested models.

3.3 Simulation-Based Model Selection

In the previous section, we have discussed some standard techniques to discriminate between nested models. In many practical data analysis problems one is often confronted with the task of choosing between non-nested models. Simple examples are linear regression problems, where we might have to decide between two different sets of explanatory variables (e.g., Cox, 1961). More complex examples include discriminating different mechanisms in cellular signal transduction pathways (e.g., Timmer et al., 2004).

Cox (1961) was the first to consider the problem of testing two non-nested hypotheses within the framework of the likelihood approach and to derive an asymptotic result. Later a simulation approach was suggested by Williams (1970). It is based on the idea

to obtain a distribution of the test statistic on the basis of the two fitted models. We take up this idea and develop and test a framework for discriminating between non-nested FARIMA$[p, d, q]$ processes.

3.3.1 Non-Nested Model Selection

Cox (1961) suggested a test statistic for discriminating non-nested models based on the log-likelihood. Consider $l_f(\widehat{\boldsymbol{\theta}})$ and $l_g(\widehat{\boldsymbol{\Xi}})$, the log-likelihood of two models f and g with parameter estimates $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\Xi}}$, respectively. The test statistic is defined as

$$T_f = \left(l_f(\widehat{\boldsymbol{\theta}}|\mathbf{x}) - l_g(\widehat{\boldsymbol{\Xi}}|\mathbf{x}) \right) - \mathrm{E}_{\widehat{\boldsymbol{\theta}}} \left[l_f - l_g \right], \tag{3.9}$$

where $\mathrm{E}_{\widehat{\boldsymbol{\theta}}} \left[l_f - l_g \right]$ denotes the expected likelihood ratio under the hypothesis H_f "model f is true". A large positive value of T_f can be taken as evidence against H_g "model g is true", and a large negative value as evidence against H_f. Note that the expression in (3.9) is not symmetric in f and g due to the expectation value in the last term. It can be shown that T_f is asymptotically normal under the null hypothesis H_f. The variance of the distribution of T_f is not known in general, which renders testing impossible. For specific situations, however, it is possible to calculate the variance; for some examples, see also Cox (1961).

In the following section, we develop a simulation-based approach for discriminating models coming from the FARIMA$[p, d, q]$ class based on this test statistic.

3.3.2 Simulation-Based Approach for FARIMA$[p, d, q]$

Instead of (3.9) we consider only the difference $\mathrm{lr}_{\mathrm{obs}} = l_f(\widehat{\boldsymbol{\theta}}|\mathbf{x}) - l_g(\widehat{\boldsymbol{\Xi}}|\mathbf{x})$ and compare it to the distributions of lr_f and lr_g (Hinde, 1992). These two distributions are obtained similarly to the value for $\mathrm{lr}_{\mathrm{obs}}$ but from records $\mathbf{x}_f$ and $\mathbf{x}_g$ simulated with the respective model. If the two distributions of lr_f and lr_g are well separated, we can discriminate the two models. Depending on the observed record, we favour one or the other model or possibly reject both. A large overlap of the distributions indicates that the two models are hard to distinguish.

Schematically this procedure can be described as follows:

1. Estimate parameters $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\Xi}}$ for models f and g from the observed record, calculate the log-likelihood-ratio $\mathrm{lr}_{\mathrm{obs}} = l_f(\widehat{\boldsymbol{\theta}}|\mathbf{x}) - l_g(\widehat{\boldsymbol{\Xi}}|\mathbf{x})$.

2. Simulate R datasets $\mathbf{x}_{f,r}, r = 1, \ldots, R$ with model f and parameters $\widehat{\boldsymbol{\theta}}$. Estimate parameters $\widehat{\boldsymbol{\theta}}_{f,r}$ for model f and $\widehat{\boldsymbol{\Xi}}_{f,r}$ for model g for each ensemble member $\mathbf{x}_{f,r}$. Calculate $\mathrm{lr}_{f,r} = l_f(\widehat{\boldsymbol{\theta}}_{f,r}|\mathbf{x}_{f,r}) - l_g(\widehat{\boldsymbol{\Xi}}_{f,r}|\mathbf{x}_{f,r})$. This yields the distribution of the test statistic under the hypothesis H_f.

3. Simulate R datasets $\mathbf{x}_{g,r}, r = 1, \ldots, R$ with model g and parameters $\widehat{\boldsymbol{\Xi}}$. Estimate parameters $\widehat{\boldsymbol{\theta}}_{g,r}$ for model f and $\widehat{\boldsymbol{\Xi}}_{g,r}$ for model g for each ensemble member $\mathbf{x}_{g,r}$. Calculate $\mathrm{lr}_{g,r} = l_f(\widehat{\boldsymbol{\theta}}_{g,r}|\mathbf{x}_{g,r}) - l_g(\widehat{\boldsymbol{\Xi}}_{g,r}|\mathbf{x}_{g,r})$. This yields the distribution of the test statistic under the hypothesis H_g.

4. Compare the observed ratio $\mathrm{lr}_{\mathrm{obs}}$ to the distribution of $\mathrm{lr}_{f,r}$ and $\mathrm{lr}_{g,r}$. In case the distributions are well separated (Figure 3.1), this might yield support for the one or the other model (Figure 3.1, (a), solid line), or evidence against both of them (Figure 3.1,

(b), dashed line). If the two distributions show a large overlap (Figure 3.2) models f and g cannot be discriminated. Depending on the observed value, we find situations where we can still reject one model (Figure 3.2, b) and situations where no model can be rejected (Figure 3.2, a).

The representation of the distributions as density estimates or histograms (Figures 3.1 and 3.2, left) appear more intuitive because the separation or overlap of the distributions is directly visible. Density estimates or histograms require, however, additional parameters, such as the smoothing band width or the bin size. Therefore we prefer in the following the representation using the empirical cumulative distribution function (ECDF) (Figures 3.1 and 3.2, right).

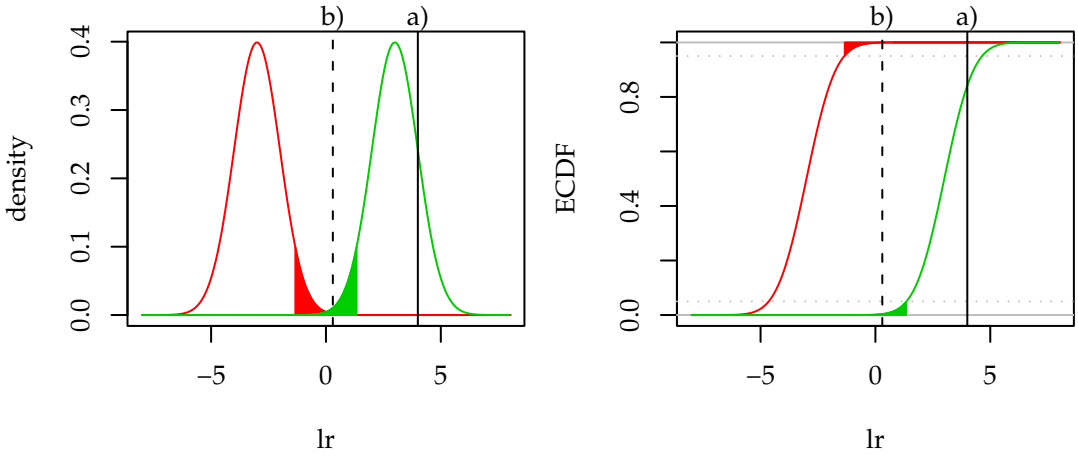

Figure 3.1: Possible outcomes of the simulation-based model selection. The two distributions are well separated. A one-sided 5% critical region is marked by the filled areas. The solid and dashed vertical line exemplify two possible values for an observed log-likelihood-ratio. The left plot shows the density function and the right plot the empirical cumulative distribution function.

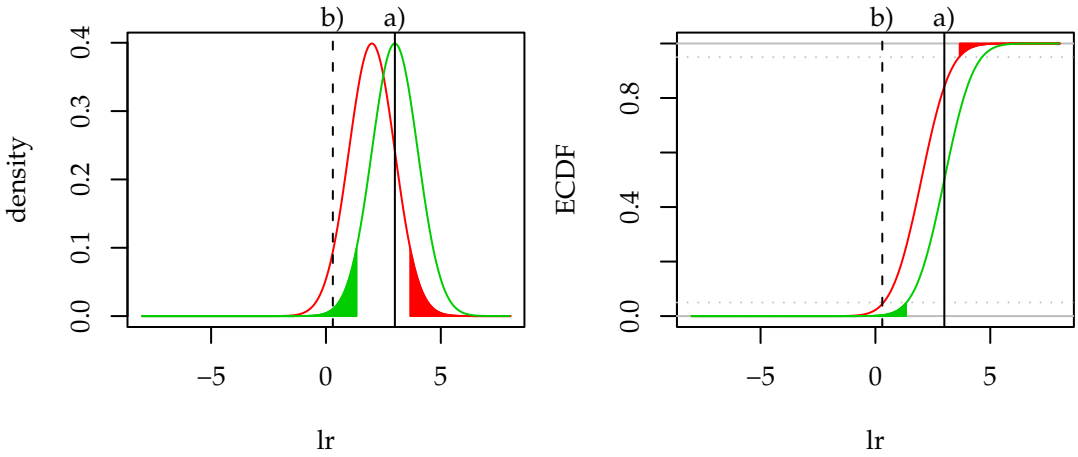

Figure 3.2: Possible outcomes of the simulation-based model selection. The two distributions are not well separated. The colour and line-coding is the same as in Figure 3.1.

The diagram in Figure 3.3 depicts the procedure in a flow-chart-like way. Besides a

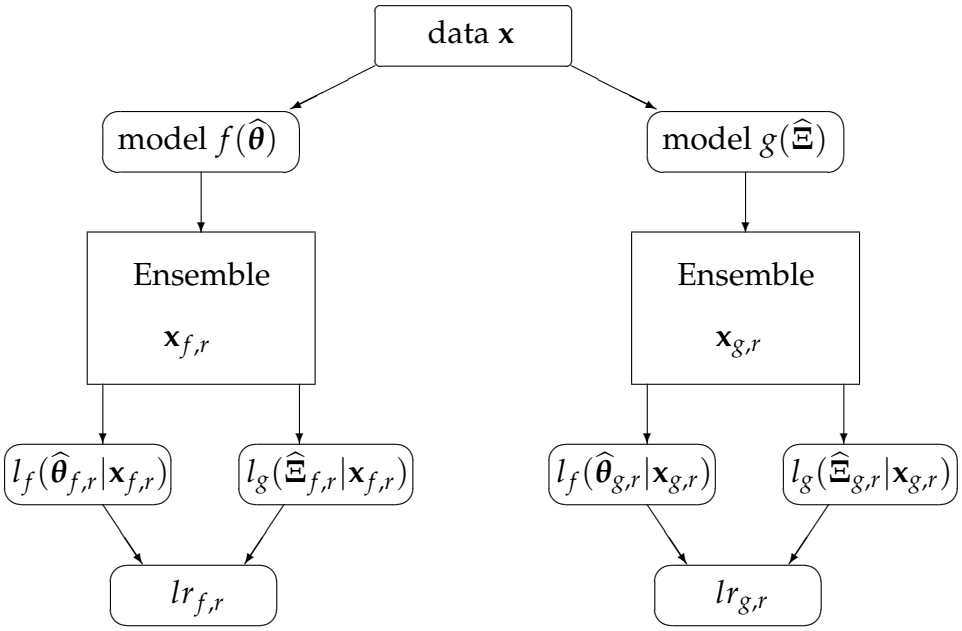

Figure 3.3: Schematic representation of the simulation-based model selection.

visual inspection of the result in a plot of the densities (histograms) or the empirical cumulative distribution functions, we can estimate critical regions, p-values and the power of the tests. Examples of this simulation-based model selection strategy in different settings are given by, e.g., Hinde (1992) or Timmer et al. (2004).

Critical Regions, p-values and Power

Critical Regions If we follow the specification (3.9) of the test statistic for testing the hypothesis H_f against H_g, we expect a deviation towards negative values if H_f is not true. To estimate the limit of a critical region, i.e. a critical value, at a nominal level α for R runs, we use

$$\widehat{\mathrm{lr}}^{\mathrm{crit}}_{f,\alpha} = \mathrm{lr}_{f,([\alpha R])}\,. \tag{3.10}$$

$\mathrm{lr}_{f,(r)}$ denotes the r-th largest value and $[\alpha R]$ the integer part of αR.

Reversing the hypothesis and testing H_g against H_f, we expect a deviation towards positive values if H_g is not true and use $\mathrm{lr}_{g,([(1-\alpha)R])}$ as an estimate for the critical value.

p-values Furthermore, we might estimate p-values in a similar way. A one-sided p-value for testing hypothesis H_f against H_g can be estimated as

$$\widehat{p}_f(\mathrm{lr}_{\mathrm{obs}}) = \frac{\#(\mathrm{lr}_{f,r} < \mathrm{lr}_{\mathrm{obs}})}{R}, \tag{3.11}$$

where $\#(\mathrm{lr}_{f,r} < \mathrm{lr}_{\mathrm{obs}})$ denotes the number of likelihood ratios $\mathrm{lr}_{f,r}$ smaller than the observed value $\mathrm{lr}_{\mathrm{obs}}$.

Power The power $\mathrm{pow}_f(\alpha, g)$ of testing H_f against H_g associated with a specified level α is also straightforwardly estimated. The power is defined as the probability of finding the

statistic under the alternative hypothesis in the critical region (Cox and Hinkley, 1994). An estimate for the power is thus given by

$$\widehat{\mathrm{pow}}_f(\alpha, g) = \frac{\#(\mathrm{lr}_{g,r} < \widehat{\mathrm{lr}}^{\mathrm{crit}}_{f,\alpha})}{R}. \tag{3.12}$$

3.3.3 An Illustrating Example

As an example consider the problem of discriminating between an AR[1] and a FD process on the basis of a time series of length $N = 400$. We start with a realisation of an AR[1] process and consider this series as an observed one. This time series is depicted in Figure 3.4 in the time domain as well as in the spectral domain. We suppose that the un-

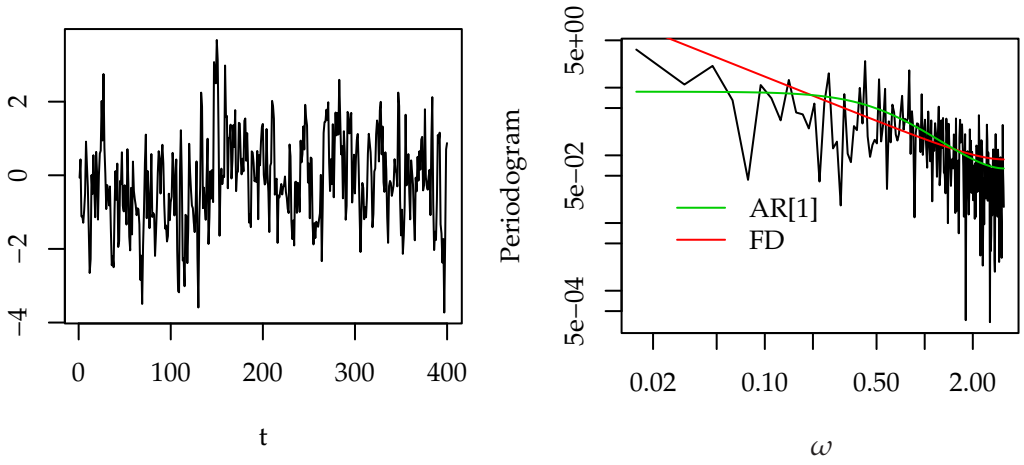

Figure 3.4: Realisation of an AR[1] process with $a_1 = 0.6, \sigma_\eta^2 = 1$ and length $N = 400$ in the time domain (left) and spectral domain (right) together with the spectral densities of the AR[1] (green) and FD (red) processes.

derlying process is either AR[1] or FD and we use the simulation-based model selection to discriminate the two models.

Fitting the AR[1] and the FD Model

Following the procedure given above, we use the Whittle estimator (Section 2.3.2) to obtain the following parameter estimates $\widehat{\boldsymbol{\theta}} = (\widehat{a}_1, \widehat{\sigma}_{\eta,f})^\intercal$ and $\widehat{\boldsymbol{\Xi}} = (\widehat{d}, \widehat{\sigma}_{\eta,g})^\intercal$ for model f and g, respectively. This results in $\widehat{a}_1 = 0.57(0.04)$, with the value in parentheses denoting the standard deviation, and $\widehat{\sigma}^2_{\eta,f} = 0.995$ for the AR[1] process (model f); $\widehat{d} = 0.47(0.04)$ and $\widehat{\sigma}^2_{\eta,g} = 1.049$ for the FD process (model g).

In both cases the goodness-of-fit test (3.4) does not reject the model on the 5%-level of significance: the p-values obtained are $\widehat{p}_f = 0.551$ and $\widehat{p}_g = 0.063$ for the AR[1] and FD process, respectively, and thus both are larger than 0.05. This implies that we are not able to discriminate the processes solely on the basis of this test. The periodogram of the original series is compared to the spectral density of the fits in Figure 3.4 (right).

The Simulation-Based Model Selection

Following the scheme depicted in Section 3.3.2, we are left with two sets of log-likelihood-ratios: $\mathrm{lr}_{f,r}$ obtained from the ensemble $\mathbf{x}_{f,r}$ with the underlying process being AR[1] and $\mathrm{lr}_{g,r}$ obtained from the ensemble $\mathbf{x}_{g,r}$ with the underlying process being FD. The two hypotheses are shown in Figure 3.5 as histograms (left) and cumulative distribution functions (right). The likelihood ratio calculated for the original series $\mathrm{lr}_{\mathrm{obs}}$ (black vertical line)

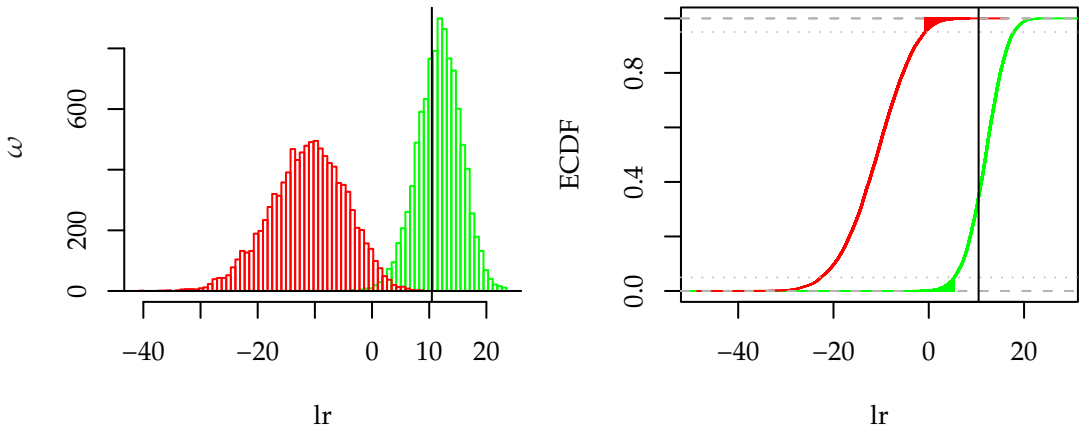

Figure 3.5: Distributions of the log-likelihood-ratio for the AR[1] (lr_f, green) and the FD (lr_g, red) model as histogram (left) and cumulative distributions function (right). The critical regions are indicated as filled regions and the log-likelihood-ratio for the original series is shown as black vertical line. The distributions are obtained from $R = 10\,000$ runs.

is located close to the centre of H_f and in the 5% critical region of H_g. This thus supports the AR[1] in favour of the FD process.

The visualisation of the hypotheses in Figure 3.5 provides us with an idea to what extent the two models are distinguishable and which hypothesis to reject. In this example the distributions are well separated and the observed value is close to the centre of the AR-hypothesis but at the edge of the distribution resulting from the FD model. Additional to this visual impression, we estimate also critical regions, the power of the tests and p-values in the way described above.

Estimating Critical Regions Consider the AR[1] model as the hypothesis under test and the FD as the alternative. For the given example using (3.10) we find a critical value at $\alpha = 0.05$ of $\widehat{\mathrm{lr}}^{\mathrm{crit}}_{f,0.05} = \mathrm{lr}_{f,([500])} = 5.393$. The critical region extends towards smaller values and thus the observed value $\mathrm{lr}_{\mathrm{obs}} = 10.463$ lies outside. Thus the hypothesis H_f "the realisation stems from an AR[1] model", cannot be rejected

For the inverse situation of testing FD against the alternative AR[1], we find a lower limit of a critical region at $\widehat{\mathrm{lr}}^{\mathrm{crit}}_{g,0.05} = \mathrm{lr}_{g,([500])} = 1.115$. The observed value is thus inside and we can reject the hypothesis that the realisation stems from an FD model.

Estimating the Power of the Test Considering again the AR[1] model as the hypothesis under test, with (3.12) we obtain an estimate $\widehat{\mathrm{pow}}_f(\alpha = 0.05, g) = 0.9978$. Simultaneously, we have obtained an estimate $\beta = 1 - \widehat{\mathrm{pow}}_f(\alpha = 0.05, g) = 0.0022$. This estimate of the power being close to one is a result of the distributions being well separated. This high

power implies that H_f (AR[1]) is very likely rejected for realisations from an FD process (H_g).

The corresponding power estimate for the inverse situation amounts to $\widehat{\text{pow}}_g(\alpha = 0.05, f) = 0.999$.

Estimating p-Values For the AR[1] model we estimate a p-value $\widehat{p}_f = 0.3533$ using (3.11). We thus have no evidence to reject this hypothesis. Since in our example, no $\text{lr}_{g,r}$ exceeds the observed likelihood ratio, we find $\widehat{p}_g = 0$ and thus strong evidence against the FD hypothesis.

On the basis of the above procedure we can thus very confidently reject the hypothesis of the data being generated from an FD model in favour of the autoregressive alternative. In the given example, we know the underlying process and find that it was correctly identified.

In order to investigate whether the simulation-based model selection strategy can be referred to as a statistical test, we have to evaluate whether the rate of rejection corresponds to the nominal level α.

3.3.4 Testing the Test

In the setting presented above, we straightforwardly estimated critical regions, p-values and the power, for testing simple hypotheses, i.e. we assume that the null as well as the alternative hypotheses are models with known parameters, namely the ML estimates for the specified models. This is different from the situation where the parameters are not known or, in other words, from testing composite hypotheses, i.e. the record is either compatible with an AR[1] or with an FD process. In the latter situation the parameters are not known but have to be estimated. This is the question we actually have to address. Thus we test the size of such a test in a simulation study. In the current setting, this implies to investigate whether the rejection rate according to the estimated p-values correspond to the nominal rate α. We furthermore investigate to what extend the estimate of the power is a useful approximation.

Testing the Size

We use again the AR[1] example process and perform an extensive simulation study using different values for the parameter $a_1 \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and various sample sizes $N \in \{100, 200, 400, 800, 2\,000, 4\,000\}$[2]. For a_1 and N fixed, we generate $M = 1\,000$ realisations. For each such realisation we use the simulation-based model selection (again with 1 000 runs) to estimate p-values for a test against a FD model. The fractional difference parameter d of this alternative hypothesis is estimated from the realisations. On the basis of this p-value for a given realisation, we can either reject or accept the null hypothesis on a prescribed nominal level of significance α. The fraction of the number of rejections of the total number of realisations M is called the rejection rate r_α. For a statistical test this rate must agree with the nominal level α for large M, or, equivalently, $r_\alpha/\alpha \to 1$ for $M \to \infty$.

[2]Obtaining results for larger N, e.g., $N = 8\,000$ could not be accomplished within one week of computation time for a given a_1 and $M = 1\,000$ runs on one node of the IBM p655 Cluster equipped with IBM Power4 CPUs with 1.1 GHz.

Figure 3.6 shows the relative rate of rejection r_α/α for three nominal levels of α. In the four upper panels (a-d) the dependence on the sample size N is given for four values of a_1. The two bottom panels (e,f) depict the dependence on the parameter a_1 for a small and a large sample size. The largest deviations from unity in all panels can be observed for a nominal level of $\alpha = 0.01$. For this level, we expect only ten of the 1 000 realisations to be rejected. Thus a deviation of 50% implies five instead of ten rejected realisation. In this range the statistic is very limited.

The plot for the N-dependence with a small autoregressive parameter (a_1=0.1, panel a), suggests that an increasing sample size reduces the difference between nominal level and calculated rejection rate. For larger a_1, the sample size does not have a clear influence anymore (panels b-d). Except for $a_1 = 0.9$, the rejection rates are predominantly smaller than the corresponding nominal level. The a_1-dependence for small sample size ($N = 100$) (panel e) shows that the rejection rate increases with the autoregressive parameter a_1. For $a_1 = 0.1$, we find the rejection rate being about 50% smaller than the nominal level. Only for $a_1 = 0.9$ more realisations of the AR[1] process are rejected than expected from the value of α. For a larger sample size ($N = 4\,000$), the mismatch between nominal level and rejection rate reduces, especially for small a_1 (panel f). In this case, we observe a relative rejection rate $r_\alpha/\alpha < 1$ for $a_1 = 0.9$ and $\alpha \in \{0.05, 0.1\}$.

In situation where we find a rejection rate smaller than the nominal level, we speak of a conservative test; the rejection under the null hypothesis is not as likely as expected. In the opposite case of a rejection rate larger than the nominal value, one is more likely to reject a true null hypothesis than expected. For the present example, this implies that the AR hypothesis is rejected in favour of an FD process. The consequence is falsely detecting LRD. Thus in the case of discriminating between an AR[1] and an FD process, special attention has to be paid to processes with large autoregressive parameters.

For the situation of the example under consideration ($a_1 = 0.6$ and $N = 400$), the relative rejection rates are close to one for $N = 400$ and $a_1 = 0.5$ ($r_{0.05}/0.05 = 1.14$) or $a_1 = 0.7$ ($r_{0.05}/0.05 = 0.96$). We can thus expect that the selection procedure is close to a statistical test in this case.

Estimating the Power

With further simulations we study the estimates of the power for the simulation-based model selection. The power is the rate of rejection for records from the alternative hypothesis. Thus we need to study realisations of a suitable process serving as such. In this setting this is an FD process with a difference parameter d chosen adequately to the corresponding null hypothesis. This parameter changes with the length N and the parameter a_1 of the AR[1] process. Thus, we estimate d from 20 realisations of the null hypothesis AR[1] for given a_1 and N and take the average value for the alternative hypothesis. With this alternative hypothesis we generate 1 000 realisations and obtain an estimate for the power according to (3.12) for each run. This sample of power estimates is then compared to the rate of rejection obtained from the ensemble by means of a box-plot. The results for various a_1 and N is shown in Figure 3.7. The blue crosses mark the difference parameter used for the alternative hypothesis. The power increases with N but in a different manner for different a_1. The most rapid increase can be observed for $a_1 = 0.5$ and $a_1 = 0.7$. In general the power estimates are very variable in regions of low power and especially for parameters $a_1 = 0.1$ and $a_1 = 0.9$ where the AR[1] and the FD process have quite similar spectral densities. For the other two parameter values ($a_1 = 0.5$ and $a_1 = 0.7$) the estimate of the power is on average good. Particularly, for the setting of the example

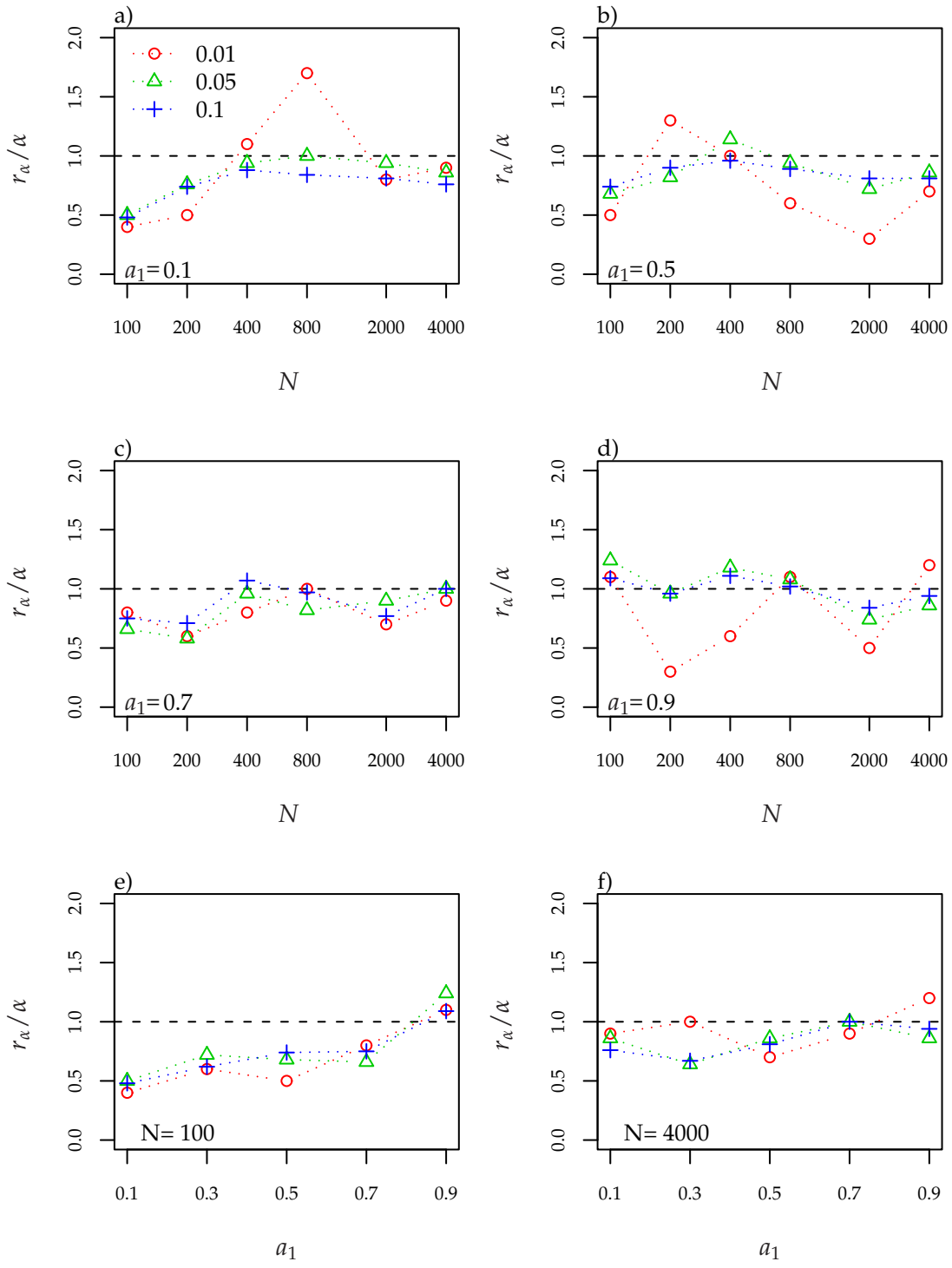

Figure 3.6: Relative rate of rejection at the 1% (circles), 5% (triangles) and 10% (crosses) nominal levels of significance. The four top panels show the dependence on the length N for $a_1 = 0.1$ (a), $a_1 = 0.5$ (b), $a_1 = 0.7$ (c) and $a_1 = 0.9$ (d). The two bottom panels show the dependence on the AR-parameter a_1 for $N = 100$ (e) and $N = 4\,000$ (f).

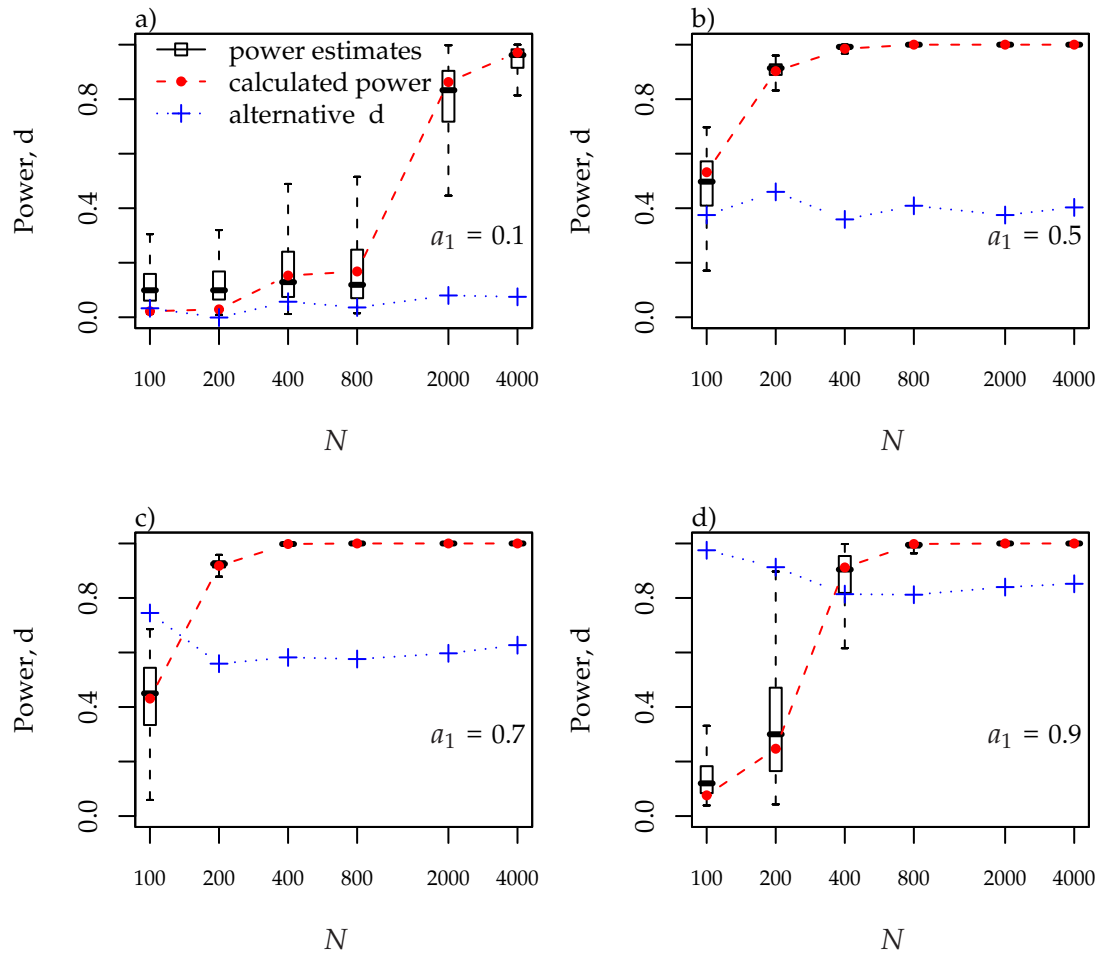

Figure 3.7: Box-plot of the power estimates for 1 000 runs dependent on the record length N. The power obtained from the simulation is marked by the red dots and the fractional difference parameter used for the alternative hypothesis as blue crosses. Different values of a_1 are shown: 0.1 (a), 0.5 (b), 0.7 (c) and 0.9 (d).

studied above ($a_1 = 0.6$, $N = 400$), the power estimate close to one should be reliable.

In some cases, we find a fractional difference parameter $d > 0.5$ which corresponds to the non-stationary domain. We are, however, still able to obtain estimates and an approximate likelihood as discussed in Section 2.3.2.

In case we cannot distinguish the two models, for example due to a low power, we can obtain a rough estimate of the time series length needed for a satisfactorily power on the basis of simulations, as described in the following.

3.3.5 Estimating the Required Sample Size

The power of the test changes with the length of the series under consideration. This is depicted in Figure 3.8. The model parameters used in the simulations are the ML estimates $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\Xi}}$ obtained for the original time series length.

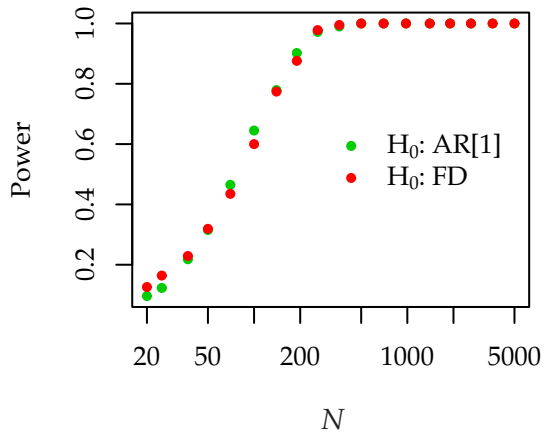

Figure 3.8: Power of testing the AR[1] hypothesis (green) and FD hypothesis (red) calculated for various lengths.

In cases where it is not possible to discriminate the models, i.e. the length of the observed record does not lead to satisfactory power of the test, we might repeat the procedure with increasing length of the simulated series. We then get an idea what record length is needed to distinguish the two models. Interpretation of this result has to be treated with care, because on the one hand, we use the estimates $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\Xi}}$ instead of the "true" parameter values. These estimates depend on the sample size N. On the other hand, every hypothesis is trivially falsified for a sufficient length of the time series. An increasing power does thus not mean that we are able to decide for one model; we might end up with rejecting both. However, for a moderate increase of sample length, we might still get an appropriate idea of the power.

In the context of geophysical records the lack of data basically implies to postpone the analysis until sufficient data has been collected. In other settings, with the possibility of influencing the size of the sample beforehand, this approach might be useful to design the experiment. This ensures that a sufficient size of the sample can be taken. Then the research question can be answered while an unnecessary surplus of data collection can be avoided, e.g., to reduce costs.

3.3.6 Bootstrapping the Residuals

Within the framework of FARIMA$[p, d, q]$ models, the ensemble members $\mathbf{x}_{f,r}$ or $\mathbf{x}_{g,r}$ can be generated from a white noise process η_t using a linear filter (cf. Section 2.2.1)

$$x_t = \sum_{i=0}^{M} \gamma_i \eta_{t-i}. \qquad [2.39]$$

In general the sum (2.39) extends over infinitively many elements M. In practical application, the filter is truncated at a suitable length. The γ_i can be derived from the model parameters (Brockwell and Davis, 1991; Beran, 1994).

For invertible models (2.39) can be inverted to obtain the residual series $\{r_t\}$ from the time series $\{x_t\}$ given the model. In situations where the distribution of the residual series is not close to a Gaussian, we might consider replacing the realisation $\{\eta_t\}$ of the white noise process with a realisation of a bootstrap resampling of the residuals $\{r_i\}$ (Hinde, 1992; Davison and Hinkley, 1997). In this way the distribution of the residuals is conserved.

3.4 Summary

We started the discussion on model selection with an example of a goodness-of-fit test: a spectral variant of the Portmanteau Test. This test is particularly suitable for the kind of models used and parameter estimation strategy we pursue here. For model comparison, we suggested the standard likelihood-ratio test and the Akaike-type information criteria (AIC, BIC, HIC). These methods can be used for choosing between nested models.

Starting from a general likelihood-ratio approach, we developed a simulation-based strategy for the discrimination of non-nested FARIMA$[p, d, q]$ models. The formulation of this approach is within the framework of statistical hypothesis testing, i.e. the concepts of critical regions, p-values, and power are used. We illustrated this method with a simulated example series from an AR$[1]$ process. On the basis of one realisation, the AR$[1]$ model could correctly be identified as the underlying process. The FD model as alternative hypothesis could be rejected, with a power close to one.

In order to study to what extend the estimates provided for the p-values and the power can be interpreted in the sense of a statistical test, we performed an extensive simulation study related to the above example. The results showed a dependence of the relative rejection rate on the parameter value a_1 and on the sample size N. The quality of the power estimates depend as well on the sample length and on the parameter values. We find useful estimates of the power in regions of high power and especially when the spectral densities of the AR$[1]$ and the FD process are not similar (here $a_1 = 0.5$ and $a_1 = 0.7$). In regions of low power and especially towards the boundaries of the interval $(0, 1)$ for a_1, estimates become highly variable and are to be interpreted with care.

If we cannot discriminate between the two models, a simulation study with an increased record length gives an idea how many data points are needed for a reliable decision. Because the simulations with increased length will be based on the parameters estimated for the observed length, the range of this analysis is limited.

Given the idea of generating ensembles from the two hypotheses under test, one can think of other statistics than the likelihood ratio. Especially in the context of detecting LRD, statistics focusing on the low frequency behaviour might be suitable. Two statistics, one based on the log-periodogram regression and one based on the DFA are discussed in

Appendix B.4. In the example presented here, the likelihood ratio as test statistic results, however, in the more powerful test.

In the following chapter, we develop a strategy to detect LRD on the basis of parametric modelling and the model selection strategies presented and developed here.

Chapter 4

Detection of Long-Range Dependence

"Nothing is what it seems!"
Morpheus to Neo in *The Matrix*, 1999

The phenomenon of long-range dependence (LRD, Section 2.1.4) has received attention since the 1950s, the time the British hydrologist H. E. Hurst studied the Nile River minimum flows, until today (Hurst, 1951; Hipel and McLeod, 1978; Montanari et al., 2000; Giratis et al., 2001; Doukhan et al., 2003; Robinson, 2003; Lohre et al., 2003). Its discussion is not limited to hydrology but is prevailing in other fields of the geosciences as well as in economics (e.g., Smith, 1993; Govindan et al., 2002; Govindan and Kantz, 2004; Lo, 1991; Henry and Zaffaroni, 2003). With this variety of applications different ways of detecting and describing the phenomenon arose. We can basically distinguish two strategies for the detection of LRD.

1. The historically older one is the search for a certain asymptotic behaviour in the autocorrelation function, in the spectrum, or similar statistics (Section 2.1.4). These approaches consist of a simple parametric description of the asymptotic behaviour of the statistic under consideration in regions where these asymptotics are assumed to hold. Such strategies are referred to as semi-parametric. If the asymptotic behaviour is sufficiently assured to hold and if the properties of the estimator are known, as it is the case for the log-periodogram regression (Appendix. B.3.2), inference about LRD can be made (Robinson, 2005). For estimators with unknown limiting distributions, as, e.g., DFA-based estimation (Section 2.3.3), inference in not possible straightforwardly. Such approaches are referred to as heuristic; they are useful for a preliminary analysis but not for statistical inference (Beran, 1994).

2. An alternative strategy is a full-parametric modelling approach including model selection as outlined in Chapter 3. The difference to semi-parametric or heuristic approaches is that non-asymptotic behaviour is not disregarded but explicitly modelled. Within the likelihood framework a limiting distribution for the estimator is available, and thus inference about the parameter values – in particular about the fractional difference parameter d – is possible (Section 2.3). More subtle is the model selection discussed in Chapter 3. The inference about model parameters, such as the fractional difference parameter, is only meaningful if the proper model is used. The problem of selecting a suitable model corresponds to selecting an adequate region for the log-periodogram regression (Beran, 1994): a small region as

well as a complex model lead to a smaller bias but brings along a larger variance for the estimate, and vice versa.

In this chapter, we develop a strategy for the detection of LRD based on the full-parametric modelling approach and the model selection strategy proposed in Chapter 3. In the first section, we generate two example series: one from a LRD and one from a SRD process. The latter is constructed such that its realisations with a given length can be easily mistaken for realisations of the LRD process. In the subsequent sections, these series are analysed with DFA and the log-periodogram regression. Finally, we develop the full-parametric strategy and illustrate its abilities to discriminate SRD and LRD with these example series.

4.1 Constructing the Example Process

Contrary to the example studied in Section 3.3.3, we consider two processes which are similar in their high frequency behaviour and differ only in the low frequency region, i.e. one is LRD, the other SRD. Realisations of them are then taken as "observed" records. Based on these "observations" we aim to decide whether the underlying process is SRD or the LRD.

In order to strongly challenge the various model selection strategies, we construct a SRD process such that it mimics a given LRD process as closely as possible. To this end, we generate a realisation of a LRD model and search for an SRD model representing this realisation reasonably well. Within the framework of the FARIMA$[p, d, q]$ family, a simple LRD process is the FARIMA$[1, d, 0]$ (or, for short, FAR$[1]$) process. A realisation of this process can be reasonably well modelled with an ARMA$[3, 2]$ process (for details, cf. Appendix B.5). We expect a realisation of this SRD model to be easily mistaken for stemming from the FAR$[1]$ process. It is thus an ideal candidate to illustrate and test detection strategies for LRD.

Figure 4.1 shows the realisation of the ARMA$[3, 2]$ (green) and the FAR$[1]$ process (red) in the time and frequency domain. Both realisations have $N = 2^{15} = 32\,768$ data points.

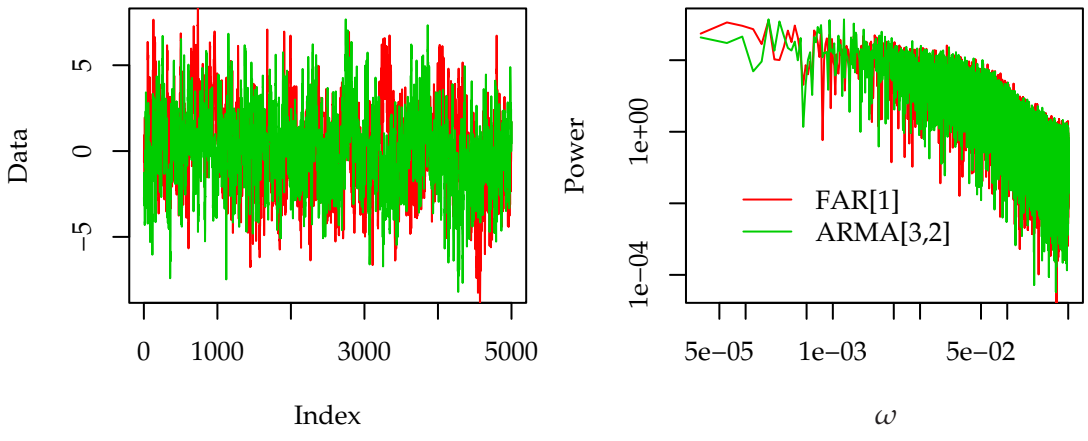

Figure 4.1: Time domain (left) and frequency domain (right) representation of the ARMA$[3, 2]$ realisation (green, $a_1 = 2.178, a_2 = -1.488, a_3 = 0.309, b_1 = -1.184, b_2 = 0.218$) compared to the FAR$[1]$ realisation (red, $d = 0.3, a_1 = 0.7$).

For the analyses in the following sections, we consider these two records as "observed" data and try to infer the underlying processes.

4.2 Detrended Fluctuation Analysis

For the detection of LRD it has become a widely used practice to compare the DFA fluctuation function (Section 2.3.3) in a double-logarithmic plot to an expected asymptotic slope for SRD or uncorrelated processes (e.g., Bunde and Havlin, 2002; Király and Jánosi, 2002; Bunde et al., 2004; Eichner et al., 2003). For the "observed" records, obtained from the ARMA$[3,2]$ and the FAR$[1]$ process, we depict these fluctuation function in Figure 4.2. From a visual inspection of the left panel one might be tempted to consider the slopes H

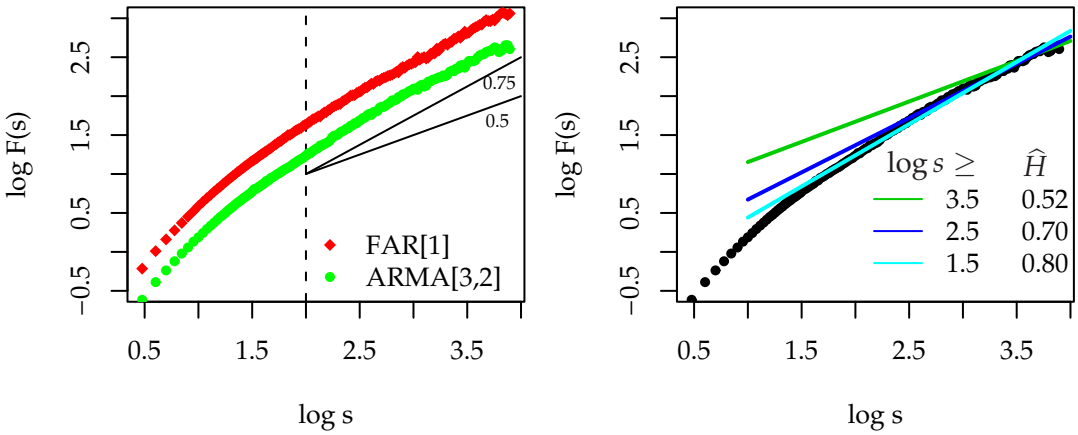

Figure 4.2: DFA1 fluctuation functions in double logarithmic representation for the ARMA$[3,2]$ (green) and the FAR$[1]$ (red) realisation (left). The two solid black lines mark the asymptotic behaviour expected for SRD, or uncorrelated, processes ($H = 0.5$) and for a LRD process with $H = 0.75$. The dashed vertical line marks the lower bound of scales considered. Straight line fits for various ranges of scales to the DFA1 fluctuation function of the ARMA$[3,2]$ realisation are shown in the right panel.

of both fluctuation functions being approximately constant with $H > 0.5$ for large scales s. For $\log s > 2$, for example, the plot suggests that in both cases the slopes are rather compatible with $H = 0.75$ than with $H = 0.5$. Thus, this provides no evidence for an SRD process in neither case. Following this line of argument, we would falsely infer a LRD process with $H = 0.75$ underlying both realisations. This demonstrates that a reliable discrimination of realisations from LRD or SRD processes is not easily achieved on the basis of DFA. Consequently such results should be interpreted with care.

A more detailed analysis of the fluctuation function of the ARMA$[3,2]$ realisation in the right panel of Figure 4.2 shows already that the slope systematically increases with decreasing lower bound of scales for the fit (i.e. an increasing range of scales considered). This should raise awareness and provides first evidence that the asymptotic behaviour has not been reached. If it had been reached, the slope should fluctuate around a constant value but should not exhibit a systematic variation. A critical review on the use of DFA to infer LRD explicitely addressing this point is given in Appendix A together with a detailed analysis of bias and variance for Hurst exponent estimation.

4.3 Log-Periodogram Regression

Contrary to a DFA-based estimator, a limiting distribution does exist for the estimator $\widehat{d}_{\text{lp}}$ based on the log-periodogram regression (Appendix B.3.2). We can thus provide confidence intervals and test whether the estimate for the fractional difference parameter is compatible with zero. For both realisations, we perform the log-periodogram regression for various bandwidths and give a 95% confidence interval for the estimate $\widehat{d}_{\text{lp}}$. Figure 4.3 illustrates the regression for the ARMA$[3,2]$ realisation (left) and the FAR$[1]$ realisation (right). The estimates $\widehat{d}_{\text{lp}}$ with the corresponding asymptotic 95% confidence intervals

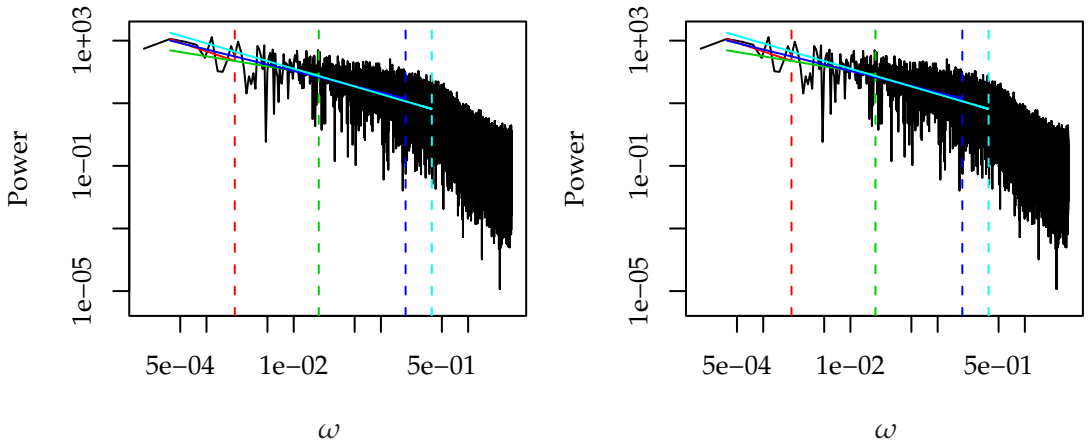

Figure 4.3: Log-periodogram regression for various bandwidths (different colours) for the ARMA$[3,2]$ (left) and the FAR$[1]$ realisation (right).

are shown in Figure 4.4. For the ARMA$[3,2]$ realisation we observe an increase of the

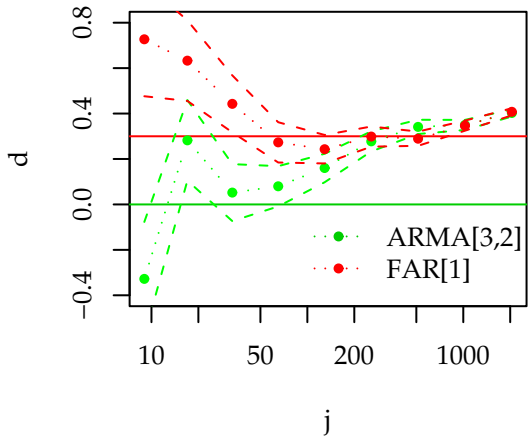

Figure 4.4: Log-periodogram regression estimates d_{lp} of the fractional difference parameter d for different bandwidth obtained from the ARMA$[3,2]$ realisation (green) and the FAR$[1]$ realisation (red). The abscissa specifies the bandwidths as index j of the largest Fourier frequency ω_j included in the fit. The green and red dashed lines specify the asymptotic 95% confidence intervals around the corresponding estimates. The corresponding solid lines mark the true value of the fractional difference parameter.

estimate $\widehat{d}_{\text{lp}}$ changing from negative to positive values. It thus passes $d = 0$. Furthermore, the 95% confidence interval for three of the smallest four bandwidths encloses the value $d = 0$ expected for SRD processes. This is different for the FAR[1] realisation: the estimates are all strictly positive and not compatible with $d = 0$ within a 95% confidence interval. In this case, we find, as expected, no evidence of SRD. Depending on the choice of the bandwidth we get ambiguous results for the ARMA$[3, 2]$ realisation. Here, the inference of the decay of the ACF (i.e. SRD or LRD) depends on the choice of the bandwidth. As mentioned before, the problem of choosing an optimal bandwidth is comparable to the model selection problem. Optimal bandwidth choice is discussed in, e.g., Robinson (2003).

4.4 Full-Parametric Modelling Approach

Instead of describing only the asymptotic behaviour for large time-scales/low frequencies from an empirical data set, a different approach for the detection of LRD is proposed in the following. We rephrase the question as a model selection problem: *Is the empirical record best represented by a LRD or a SRD model?* This implies that we need to impose some kind of principle to make the selection of a "best" model well defined. The principle we follow can be ascribed to *lex parsimoniae* – the principle of parameter parsimony. In a modelling framework, it can be expressed as follows: a model should be as complex as necessary but not more complex (Chapter 3). We thus have to decide about the necessary model complexity. This is accomplished by model selection strategies described in Chapter 3.

In order to discriminate a more complex SRD (ARMA$[p, q]$) from a simpler LRD (FARIMA$[p', d, q']$) model, we might require non-nested model selection, for example FARIMA$[1, d, 0]$ and ARMA$[3, 2]$. Alternatively, it is possible to use the simplest common model (here FARIMA$[3, d, 2]$) and test with the standard likelihood-ratio test whether one or the other model is an admissible simplification of the common model. According to Cox (1961) those tests have potentially lower power than a direct comparison of the two non-nested models. It is thus convenient to have a strategy for non-nested model selection at hand.

For both "observed" records, we estimate parameters for an ARMA$[3, 2]$, a FAR$[1]$, and a FARIMA$[3, d, 2]$ process. The two simple models are then tested for being an admissible simplification of the FARIMA$[3, d, 2]$ using the standard likelihood-ratio test (Section 3.2.1). A direct comparison of the ARMA$[3, 2]$ and the FAR$[1]$ is finally realised with the simulation-based model selection (Section 3.3).

4.4.1 Modelling the ARMA$[3, 2]$ Realisation

We begin with estimating model parameters for the three processes considered as potential models for the ARMA$[3, 2]$ realisation. The resulting parameters and p-values for the goodness-of-fit test (Section 3.1.3) are listed in Table 4.1. The parameter estimates for the FAR$[1]$ model are, within one standard deviation, compatible with the parameters of the original FAR$[1]$ process used to motivate the ARMA$[3, 2]$ model. Likewise, the ARMA$[3, 2]$ estimates are compatible with the parameters of the underlying process. The standard deviation of the FARIMA$[3, d, 2]$ parameter estimates is about one order of magnitude larger than for the other two models, except for d. Taking these standard deviations into account, the AR and MA parameters are compatible with those from the

Parameter	ARMA[3,2]	FAR[1]	FARIMA[3,d,2]
d	-	0.300(0.015)	0.316(0.022)
a_1	2.182(0.041)	0.705(0.014)	1.750(0.531)
a_2	−1.493(0.078)	-	−1.094(0.883)
a_3	0.309(0.036)	-	0.248(0.387)
b_1	−1.179(0.043)	-	−1.062(0.528)
b_2	0.210(0.042)	-	0.372(0.552)
p-val	0.946	0.931	0.935

Table 4.1: Parameters and asymptotic standard deviation in parentheses for the ARMA[3,2], FAR[1] and the FARIMA[3,d,2] model estimated from the ARMA[3,2] realisation.

ARMA[3,2]. Similarly, the estimate for the fractional difference parameter d is compatible with the corresponding parameter of the original FAR[1] process. None of the three models can be rejected on any reasonable level of significance according to the goodness-of-fit test. Furthermore, $d = 0$ is not within the asymptotic 95% confidence interval for the estimate $\widehat{d}_{\text{lp}}$ for both fractional models.

In Figure 4.5 (left), we compare the spectral densities of the two competing models, FAR[1] and the ARMA[3,2], to the periodogram of the "observed" series. The spectral

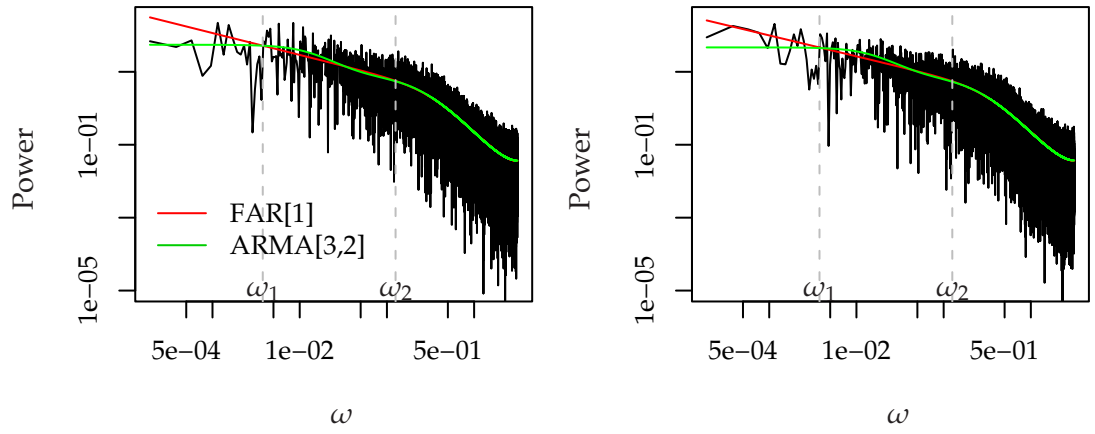

Figure 4.5: Periodogram of the ARMA[3,2] (left) and FAR[1] (right) realisation together with the spectral densities of the ARMA[3,2] (green) and FAR[1] (red) model fitted to these realisation. The dashed vertical lines mark regions on the frequency axis where the spectral densities diverge (left), are similar (middle) and are quasi identical (right).

densities are quasi indistinguishable for frequencies $\omega \gtrsim \omega_2 = 0.13$. Differences are visible in the range $4 \cdot 10^{-3} = \omega_1 \gtrsim \omega \gtrsim \omega_2$. Here, the ARMA[3,2] spectral density does not exactly reproduce the power-law like behaviour of the FAR[1] process. However, it mimics it to a certain extend. Divergence of the spectral densities can be observed for $\omega \lesssim \omega_1$: in this range the difference between the SRD and the LRD model manifests.

Likelihood-Ratio Test

Standard likelihood-ratio tests (Section 3.2.1) of the FARIMA[3,d,2] against the two simpler models yield p-values of $p = 1$ and $p = 0.793$ for the ARMA[3,2] and the FAR[1],

respectively. We can thus not reject neither model as being an admissible simplification of the FARIMA$[3, d, 2]$ on any reasonable level of significance. The power of testing FARIMA$[3, d, 2]$ against FAR$[1]$ is not large enough to reject the false simplification. We are thus in a situation where we cannot decide about the SRD or LRD on the basis of a likelihood-ratio test. Furthermore, the fractional difference parameter estimate for the FARIMA$[3, d, 2]$ model is also not compatible with zero, giving no indication for SRD. We consequently need a more specific test for a reliable decision and we consider the simulation-based approach introduced in Section 3.3.

Simulation-Based Model Selection

Analogously to the procedure outlined in Section 3.3.2, we now perform the simulation-based model selection with the likelihood ratio as test statistic. The two models we consider as candidates for the underlying process are the FAR$[1]$ and the ARMA$[3, 2]$. To facilitate the presentation, we denote in the following the ARMA$[3, 2]$ and FAR$[1]$ process as model f and g, respectively.

Figure 4.6 shows the cumulative distribution functions of the two log-likelihood-ratios. We find the value obtained for the "observed" record $\mathrm{lr}_{\mathrm{obs}}$ well outside the critical

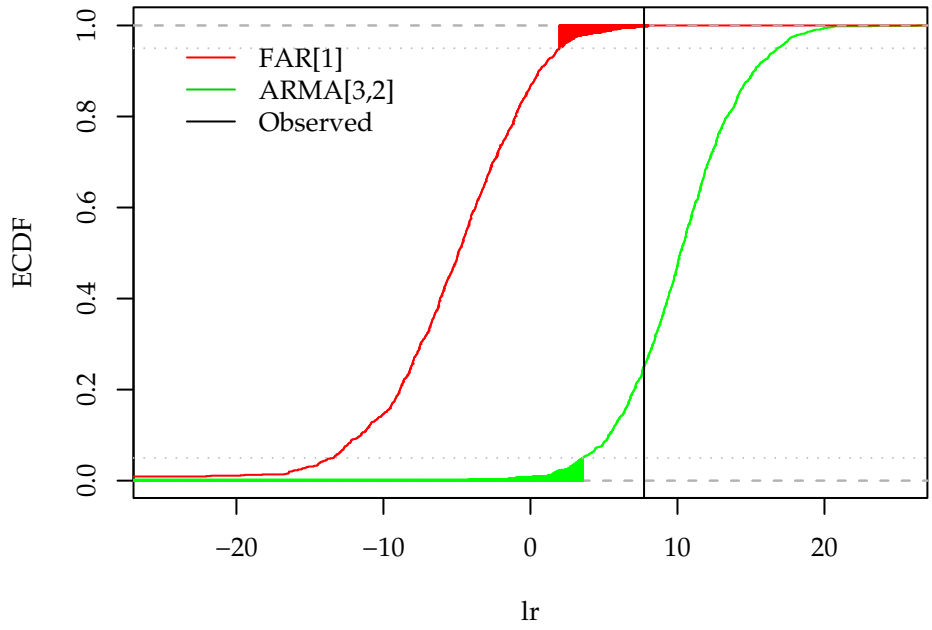

Figure 4.6: Empirical cumulative distribution function of log-likelihood-ratios $\mathrm{lr}_{f,r}$ (green) and $\mathrm{lr}_{g,r}$ (red) obtained from 1 100 simulations. The critical regions upto α-levels of 0.05 are filled with the respective color. The log-likelihood-ratio $\mathrm{lr}_{\mathrm{obs}}$ for the observed series is shown as a vertical line.

region of model f and inside the critical region of model g. The estimate of the power of testing H_f against H_g for the specified critical level $\alpha = 0.05$ yields $\widehat{\mathrm{pow}}_f(0.05, g) = 0.979$. For the inverse situation of testing H_g against H_f we find $\widehat{\mathrm{pow}}_g(0.05, f) = 0.977$. We can thus very confidently reject H_g (FAR$[1]$) in favour of H_f (ARMA$[3, 2]$) and correctly identify the underlying process.

4.4.2 Modelling the FAR[1] Realisation

Next, we estimate the model parameters obtained from the FAR[1] realisation. The resulting paramters and p-values for the goodness-of-fit test (Section 3.1.3) are listed in Table 4.2. Again, no model can be rejected on the basis of the goodness-of-fit test and

Parameter	ARMA$[3,2]$	FAR$[1]$	FARIMA$[3,d,2]$
d	-	$0.287(0.015)$	$0.296(0.019)$
a_1	$2.178(0.044)$	$0.710(0.014)$	$0.704(3.582)$
a_2	$-1.488(0.083)$	-	$0.001(1.298)$
a_3	$0.309(0.039)$	-	$-0.002(1.216)$
b_1	$-1.184(0.045)$	-	$-0.007(3.588)$
b_2	$0.218(0.045)$	-	$0.002(1.707)$
p-val	0.333	0.356	0.359

Table 4.2: Parameters and asymptotic standard deviation in parentheses for the ARMA$[3,2]$, FAR$[1]$ and the FARIMA$[3,d,2]$ model estimated from the FAR$[1]$ realisation.

the original FAR$[1]$ parameters are recovered within one standard deviation. Also for the FARMA$[3,d,2]$ model, the estimates are compatible with the parameters of the original FAR$[1]$ process.

Figure 4.5 (right) compares the ARMA$[3,2]$ and the FAR$[1]$ spectral densities to the periodogram of the FAR$[1]$ realisation. The picture is very similar to the left panel in the same figure. The two spectral densities are quasi identical in the high frequency region but differ for low frequencies. Especially for $\omega < \omega_1$ the difference between the LRD and SRD asymptotic behaviour is clearly visible.

Likelihood-Ratio Test

A standard likelihood-ratio test of the ARMA$[3,2]$ model being an admissible simplification of FARIMA$[3,d,2]$ is rejected on a 5%-level with a p-value of $p = 0.006$. The same test for FAR$[1]$ as simplified model cannot be rejected ($p = 0.889$). In this case, with the underlying process being LRD, we can discriminate ARMA$[3,2]$ and FAR$[1]$ already on the basis of standard nested-model selection. The simulation-based selection criterion for non-nested models is not needed. Thus, also for the second "observed" record, a correct identification of the underlying process could be achieved on the basis of a full-parametric model.

4.5 Summary

We illustrated three different strategies for the detection of LRD along the lines of a challenging example: one realisation of a FAR$[1]$ and one from an ARMA$[3,2]$ process with parameters chosen such that realisations with a specific length have very similar spectral characteristics.

With the heuristic DFA, the two realisations were not distinguishable. Furthermore, with the log-log plot as the central argument, we falsely "inferred" a LRD processes underlying both records. The log-periodogram regression revealed differences in the two series. For the FAR$[1]$ realisation, the asymptotic 95% confidence intervals did not enclose $d = 0$ for all the bandwidths used. Some did enclose this value for the ARMA$[3,2]$ realisation. Here, a suitable strategy for selecting the proper bandwidth is required.

The third approach to detect LRD is based on the full-parametric modelling and model selection strategies presented previously. We fitted FAR[1], ARMA[3, 2], and the smallest common model, FARIMA[3, d, 2], to both realisations. For the ARMA[3, 2] realisation, we could not reject neither the FAR[1] nor the ARMA[3, 2] as admissible simplifications of the FARIMA[3, d, 2] on the basis of a standard likelihood-ratio test. Here, we had to use the simulation-based approach to directly compare the two alternatives FAR[1] and ARMA[3, 2]. The likelihood-ratio test was, however, sufficient in the case of the FAR[1] realisation. For both "observed" records the proper model and, in particular, the correct dependence structure could be identified.

This example illustrated the reliability of the full-parametric modelling approach with respect to the detection of LRD. Furthermore, it turned out that the standard model selection, in some cases, such as the example at hand, does not have enough power to discriminate the two alternatives under consideration. In such cases, a suitable strategy for the discrimination of non-nested models, such as the simulation-based model selection, is required

The approach presented here, i.e. rephrasing the problem of detecting LRD as a model selection problem, is different from other approaches mainly because of the direct comparison of the most suitable SRD and the most suitable LRD process. A pivotal element in this model comparison is the model selection strategy based on the log-likelihood-ratio.

Chapter 5

Modelling Temperature Records

The fluctuation of temperature anomalies, i.e. a deviation from a long-run mean value, has been frequently discussed in recent years with respect to LRD (e.g., Smith, 1993; Koscielny-Bunde et al., 1996; Talkner and Weber, 2000; Caballero et al., 2002; Monetti et al., 2001; Smith, 2003; Eichner et al., 2003; Fraedrich and Blender, 2003; Cohn and Lins, 2005; Gil-Alana, 2005). The specific interest in this feature is partly motivated by a decreased sensitivity in the detection of trends and structural breaks compared to the uncorrelated or short-range correlated case (cf. Section B.1; e.g., Sibbertsen, 2001; Craigmile et al., 2004; Kallache, 2007). Furthermore, it is a feature – if detected for temperature records – that might be expected to be reproduced by climate models. The line of argument in this respect, proposed by Govindan et al. (2002), is the following: climate models do not show LRD but have similar spectral power on low frequencies as observed temperature records, they might compensate for this deficiency by overestimating the trend. Because trends in temperature are a politically highly relevant subject, the detection of LRD in temperature records was subjected to considerable interest and was also controversially debated (Fraedrich and Blender, 2003; Bunde et al., 2004; Fraedrich and Blender, 2004; Vyushin et al., 2004). A matter of scientific discussion were as well the methods used to infer a LRD process underlying the temperature records, particularly DFA (e.g., Metzler, 2003; Maraun et al., 2004). We discuss DFA with respect to the inference of a LRD underlying process in Chapter 4 and Appendix A.

One of the first series of temperature anomalies investigated for LRD with DFA was a record from Prague (Koscielny-Bunde et al., 1998). In the following, we analyse the same record using the full-parametric strategy developed in Section 4.4. A discussion on difficulties and pitfalls using DFA is illustrated with the very same temperature record in Appendix A.3 (cf. also Maraun, Rust, and Timmer, 2004).

Besides the Prague temperature series, we investigate the northern hemisphere mean temperature anomalies (Jones et al., 1999) for LRD. This series is suspected to stem from an underlying LRD process, too (Cohn and Lins, 2005).

During the following analysis, we assume the record Y_t to be a realisation of a process which we consider as a linear superposition of various influences, namely a trend T_t, periodic components S_t (seasonal cycle), and a stationary stochastic process X_t:

$$Y_t = T_t + S_t + X_t. \tag{5.1}$$

Under this assumption, we can separately estimate and remove the trend and seasonal components. The remaining noise term, or residual term, can then be analysed for LRD. A concise description of this preprocessing for the following analyses, i.e. the subtraction

of the seasonal cycle and trend components, as well as a static transformation function is given in Appendix D.

5.1 Prague Daily Maximum Temperature

We consider the daily maximum temperature series measured at the Clementinum in Prague (Czech Republic). The record as it has been used by Koscielny-Bunde et al. (1998) extends from 01/01/1800 to the 31/12/1992 and thus comprises almost 200 years (70 492 measurements). The first ten years are depicted in Figure 5.1.

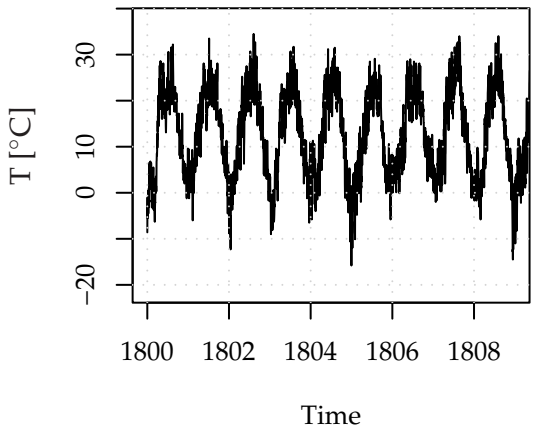

Figure 5.1: First ten years of the daily maximum temperature measured at the Clementinum in Prague.

Prior to the parametric modelling of the stochastic noise term, we estimate and remove the annual cycle. The latter is the most prominent signal in the record (Figure 5.1). Furthermore, we estimate and subtract a polynomial trend (Appendix D.3.1). The resulting residual series is regarded as a realisation of the stochastic process X_t in (5.1). For this series, we estimate model parameters for the FARIMA$[p,d,q]$ and ARMA$[p,q]$ class using the Whittle estimator (Section 2.3.2) and choose a suitable model (Chapter 3). Finally, we discuss the detection of LRD (Chapter 4) and present a simple trend test based on the parametric model found.

5.1.1 Stochastic Modelling

The autocorrelation sequence of the residuals is shown in Figure 5.2. It is outside the 95% confidence interval for white noise for lags smaller than 3 months. For larger lags the variance of the estimation renders any interpretation difficult. In the following, we pursue a full-parametric description of the ACF.

Consider a family of FARIMA$[p,d,q]$ processes with $0 \leq p \leq 8$ and $0 \leq q \leq \min(p,7)$ and additionally the ARMA$[p,q]$ class with orders $1 \leq p \leq 8$ and $0 \leq q \leq \min(p,7)$. Larger orders p and q lead to difficulties in the numerical estimation procedure. The reduction of this large set of models (here 87) is done in three steps:

1. goodness-of-fit test on a 5%-level of significance to eliminate inappropriate models (Section 3.1.3),

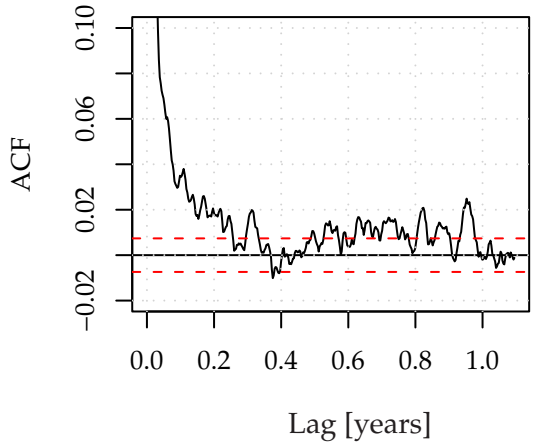

Figure 5.2: Autocorrelation sequence of the Prague temperature residuals. The dashed red lines mark the asymptotic 95% confidence intervals pertaining to white noise.

2. a preselection with the Hannan-Quinn Information Criterion (HIC, Section 3.2.2), to reduce the potentially large set to a small number[1] of models which are then mutually compared with a

3. likelihood-ratio test or simulation-based model selection (Sections 3.2.1 and 3.3).

Goodness-of-Fit Test

On the basis of the goodness-of-fit test (3.4) applied to the before mentioned set of models, we can reject the following processes on a 5%-level of significance:

- FARIMA$[p,d,q]$ with $(p,q) \in \{(1,0),(1,1),(2,0),(2,1)\}$,
- ARMA$[p,q]$ with $(p,q) \in \{(1,0),(2,0),(1,1),(2,1),(7,1)\}$.

HIC-Based Model Selection

We compare the remaining ARMA$[p,q]$ and FARIMA$[p,d,q]$ models by means of the HIC (Figure 5.3). In Section 3.2.2 we discussed the drawbacks of this approach and thus we consider it only as a rough guide used to restrain the set of potentially suitable processes. For the FARIMA$[p,d,q]$ processes, the lowest value is attained for FARIMA$[2,d,2]$ followed by the orders $[3,d,2]$, $[4,d,1]$, $[3,d,1]$, $[3,d,3]$, $[4,d,2]$, $[5,d,1]$, $[4,d,3]$, $[5,d,2]$ and $[5,d,0]$. Restricting the fractional difference parameter to $d = 0$, we find the ARMA$[5,3]$ process having the smallest HIC, followed by the $[6,1]$, $[8,1]$, $[5,1]$, $[4,1]$, $[8,2]$ and $[6,2]$. The models named here are retained and compared using the likelihood-ratio test or the simulation-based approach.

[1] In the following, we reduce the number of models to typically around three to ten models for each, the short-range as well as the long-range dependent processes. This number is fuzzy and depends on the situation at hand. As a rule of thumb, models within a range of about two units of HIC can be considered as comparable (Ripley, 2004). In principle, all models can be mutually compared with the likelihood-ratio test or the simulation-based approach. Involving the information criteria is just a matter of convenience.

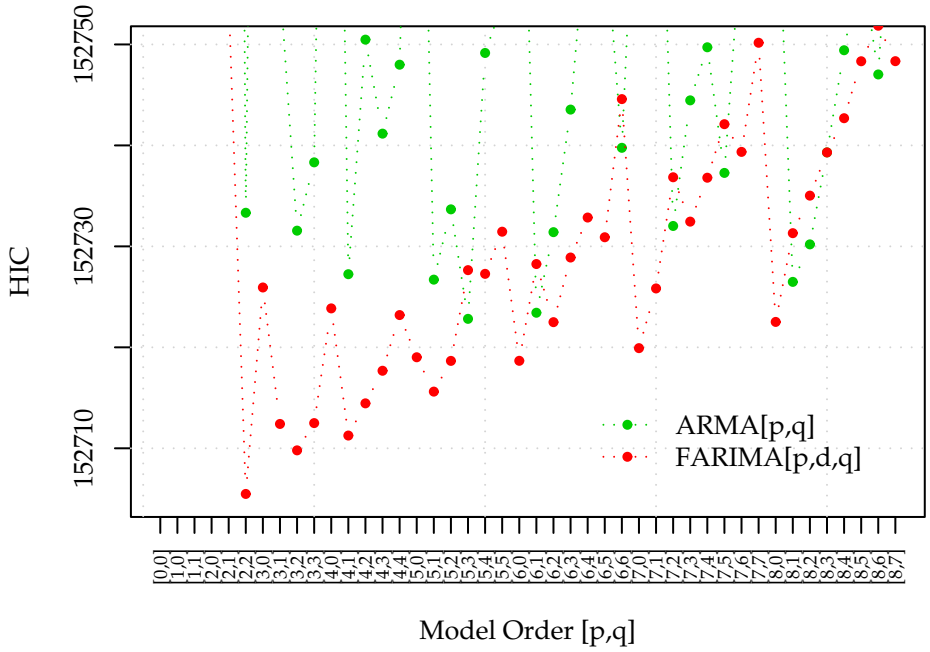

Figure 5.3: HIC for various ARMA$[p,q]$ (green) and FARIMA$[p,d,q]$ (red) models fitted to the Prague temperature anomalies. The model orders $[p,q]$ are plotted along the abscissa. The orders are discrete and the lines between the points are drawn only to guide the eye.

Likelihood-Ratio-Based Model Selection

As far as the models are nested, we use a likelihood-ratio test on a 5%-level of significance. The resulting p-values for tests within the ARMA$[p,q]$ and the FARIMA$[p,d,q]$ family are given in Tables 5.1.

From the first two entries in Table 5.1 (left), we find ARMA$[8,1]$ as an admissible simplification of ARMA$[8,2]$. ARMA$[6,2]$ is not suitable as such a simplification. Furthermore, ARMA$[8,1]$ cannot be simplified by ARMA$[6,1]$ (line 3) and ARMA$[5,3]$ does not find an admissible simplification in neither ARMA$[5,1]$ nor ARMA$[4,1]$. Thus we leave this series of tests with an ARMA$[5,3]$ and an ARMA$[8,1]$ model. These two models are not nested and we revert to ARMA$[8,3]$ as a common larger model. Both alternatives are admissible simplifications of the larger model (lines 6 and 7). We thus try to discriminate them using the simulation-based approach (Figure 5.4, left). The empirical cumulative distribution function of the likelihood ratios obtained by simulation of the ARMA$[8,1]$ and ARMA$[5,3]$ models show a large overlap and the observed value is in neither critical region. We thus cannot reject one of the two models and, for the moment being, we keep both.

As for the ARMA$[p,q]$ models, we compare the FARIMA$[p,d,q]$ models with lowest HIC on the basis of a likelihood-ratio test (Table 5.1, right). From the first three lines we find FARIMA$[5,d,1]$ and FARIMA$[3,d,2]$ as admissible simplification of FARIMA$[5,d,2]$. FARIMA$[5,d,1]$ can be further simplified by FARIMA$[4,d,1]$ (line 5) but not by $[5,d,0]$ (line 4). FARIMA$[4,d,1]$ is also an admissible simplification of FARIMA$[4,d,3]$ (line 6) but cannot be further simplified by FARIMA$[4,d,0]$ (line 7). The FARIMA$[3,d,2]$ resulting from line 3 is an admissible simplification of FARIMA$[3,d,3]$ and can be further simplified by FARIMA$[2,d,2]$ (lines 8 and 10). It cannot be simplified by a FARIMA$[3,d,1]$ (line

ARMA$[p,q]$

	Model f	Model g	p-val
1	$[8,2]$	$[8,1]$	0.292
2	$[8,2]$	$[6,2]$	0.004
3	$[8,1]$	$[6,1]$	0.037
4	$[5,3]$	$[5,1]$	0.001
5	$[5,3]$	$[4,1]$	$<$0.001
6	$[8,3]$	$[8,1]$	1
7	$[8,1]$	$[5,3]$	1

FARIMA$[p,d,q]$

	Model f	Model g	p-val
1	$[5,d,2]$	$[5,d,1]$	0.182
2	$[5,d,2]$	$[4,d,2]$	$<$0.001
3	$[5,d,2]$	$[3,d,2]$	0.673
4	$[5,d,1]$	$[5,d,0]$	0.004
5	$[5,d,1]$	$[4,d,1]$	0.492
6	$[4,d,3]$	$[4,d,1]$	0.198
7	$[4,d,1]$	$[4,d,0]$	$<$0.001
8	$[3,d,3]$	$[3,d,2]$	0.145
9	$[3,d,2]$	$[3,d,1]$	0.006
10	$[3,d,2]$	$[2,d,2]$	0.476
11	$[4,d,2]$	$[4,d,1]$	0.201
12	$[4,d,2]$	$[2,d,2]$	0.712

Table 5.1: p-values for a likelihood-ratio test of model g being an admissible simplification of model f.

9). We thus remain with two non-nested models, FARIMA$[2,d,2]$ and FARIMA$[4,d,1]$. Reverting to the common larger model, FARIMA$[4,d,2]$, results in both alternatives being an admissible simplification (lines 11 and 12). We thus need the simulation-based approach to discriminate them (Figure 5.4, right). The observed log-likelihood-ratio is

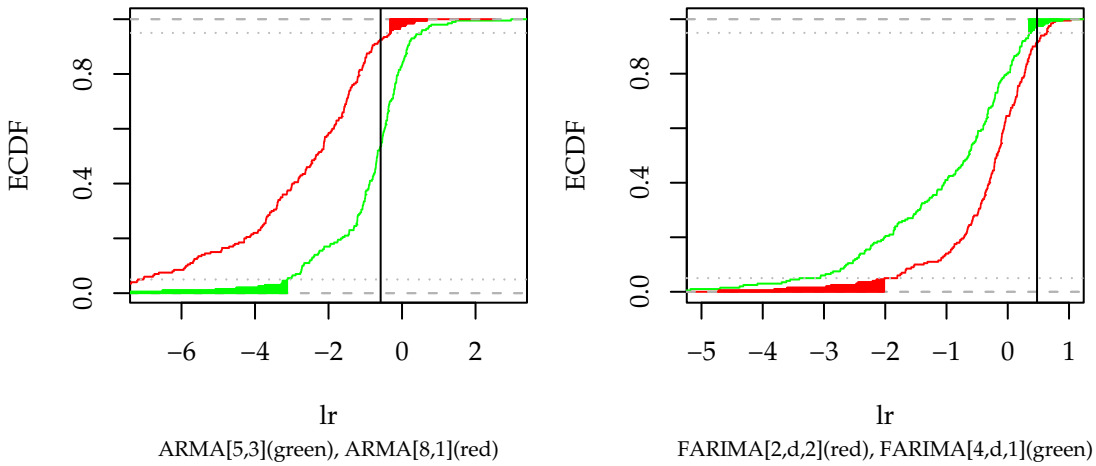

Figure 5.4: Simulation-based model selection (200 runs) for ARMA$[5,3]$ against ARMA$[8,1]$ (left) and FARIMA$[2,d,2]$ against FARIMA$[4,d,1]$ (right). The black solid line represents the corresponding log-likelihood-ratio for the observed series.

in the critical region of the distribution obtained from the FARIMA$[4,d,1]$. We thus reject the hypothesis that the latter is a more suitable model than FARIMA$[2,d,2]$. This leaves us with the FARIMA$[2,d,2]$ model as the most suitable LRD model found in the set of models we started with.

Parameter Estimates and Spectral Densities

The ML parameter estimates obtained for the three remaining models are summarised in Table 5.2. Three of the nine parameters estimated for the ARMA$[8,1]$ are compatible with

Parameter	FARIMA$[2,d,2]$	ARMA$[5,3]$	ARMA$[8,1]$
d	$0.109(0.017)$	–	–
a_1	$1.210(0.066)$	$1.548(0.275)$	$1.762(0.012)$
a_2	$-0.332(0.042)$	$-0.139(0.544)$	$-0.879(0.012)$
a_3	–	$-0.679(0.368)$	$0.145(0.008)$
a_4	–	$0.314(0.116)$	$-0.035(0.008)$
a_5	–	$-0.050(0.019)$	$0.005(0.008)$
a_6	–	–	$-0.003(0.008)$
a_7	–	–	$0.002(0.008)$
a_8	–	–	$-0.005(0.004)$
b_1	$-0.513(0.057)$	$-0.741(0.275)$	$-0.955(0.011)$
b_2	$-0.106(0.007)$	$-0.567(0.326)$	–
b_3	–	$0.342(0.095)$	–
p-val	0.074	0.071	0.075

Table 5.2: Maximum likelihood parameter estimates and asymptotic standard deviation in parentheses for the FARIMA$[2,d,2]$, ARMA$[5,3]$ and ARMA$[8,1]$ process obtained from the Prague temperature residuals. The last line gives the p-values of the goodness-of-fit test.

zero within one standard deviation, giving evidence that this model is over-parametrised. We take this as motivation to dismiss this process. The remaining two models are then FARIMA$[2,d,2]$ and ARMA$[5,3]$. Figure 5.5 compares the spectral densities of these most suitable LRD and SRD processes to the periodogram of the temperature residuals. Naturally, we find the strongest deviations in the low frequency range. The two processes are almost indistinguishable for frequencies $f > 0.01 = 1/100$ days. Discriminating these two models addresses the question whether the underlying process is either LRD or SRD.

5.1.2 Detecting Long-Range Dependence

We finally compare the most suitable SRD (ARMA$[5,3]$) and LRD (FARIMA$[2,d,2]$) model. Reverting to the smallest common model (FARIMA$[5,d,3]$) and performing a LRT leads to the ARMA$[5,3]$ and the FARIMA$[2,d,2]$ being both admissible simplifications. In this case the standard approach does not have enough power and we use the simulation-based model selection (Figure 5.6). We can reject the ARMA$[5,3]$ process with a power of $\widehat{\mathrm{pow}}_{\mathrm{ARMA}[5,3]}(\mathrm{FARIMA}[2,d,2], 0.05) = 0.884$. This leaves us with the FARIMA$[2,d,2]$ as the most suitable model for the Prague temperature residuals. Thus, on the basis of the full-parametric modelling approach, we can conclude an underlying LRD process.

To corroborate this argument, we analyse an artificial series generated from a SRD process constructed such that its realisations mimic the characteristics of the Prague series as closely as possible (Appendix A.3). We find that, a realisation of this process could not be distinguished from the original record by means of DFA. Both, the empirical Prague temperature residuals and this simulated record are classified as stemming from a LRD process. The full-parametric approach, instead, could unambiguously identify the underlying process of the artificial record as SRD. Thus, for this example, we find a larger

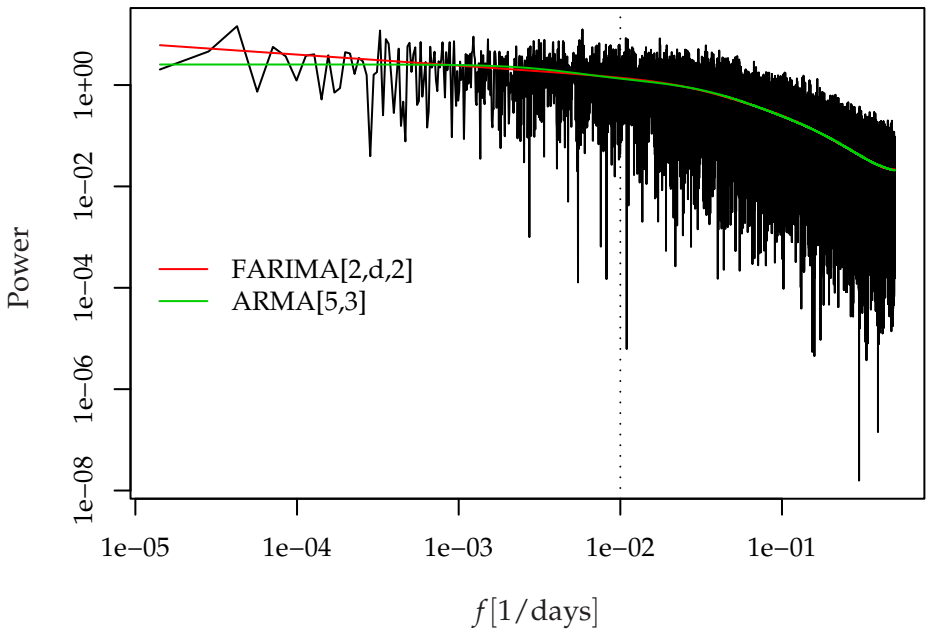

Figure 5.5: Spectral densities of the FARIMA$[2, d, 2]$ (red) and ARMA$[5, 3]$ (green) processes with optimised parameters in a double logarithmic plot together with the periodogram of the Prague temperature residuals. The dotted vertical line mark the frequency $f = 0.01$ (period of 100 days).

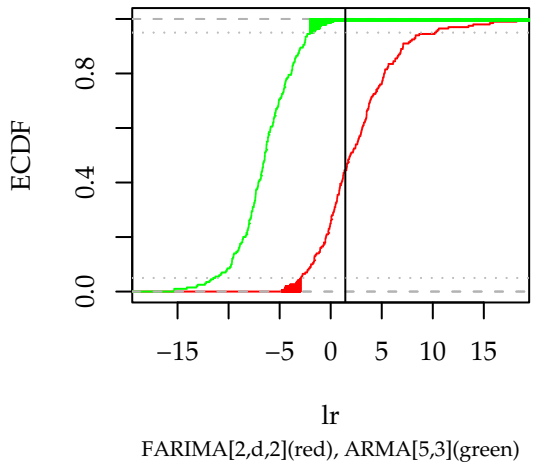

Figure 5.6: Simulation-based model selection (200 runs) for FARIMA$[2, d, 2]$ against ARMA$[5, 3]$.

specificity[2] for the full-parametric approach than for DFA.

5.1.3 Conservative Trend Test

We can straightforwardly perform a conservative test for a trend based on a parametric model. Assuming a potential trend does marginally influence the stochastic modelling of the natural variability, we can repeat the above model building procedure for the temperature anomalies, i.e. the transformed series with the seasonal cycle, but not with the trend, removed. This model can be used to generate an ensemble of records representing the natural variability, i.e. the hypothesis of no trend. To each ensemble member we can fit the polynomial of second order (D.5) and obtain distributions for the parameters a, b and c of the trend model. These can be used for testing H_0 "$a = 0, b = 0$ and $c = 0$". This test is conservative because a potential trend influences the FARIMA$[p, d, q]$ parameter estimation in the way that those parameters describing the low-frequency behaviour are biased towards larger values, especially the fractional difference parameter. Thus the stationary stochastic model derived exhibits low-frequency power which might mask a small trend. This results in a trend test with low power. A more refined test based on a semi-parametric estimation of the trend is discussed, e.g., by Kallache, Rust, and Kropp (2005) and Kallache (2007).

Model Selection and Parameter Estimates The model selection is essentially the same as for the detrended series with the exception that the differences between ARMA$[p, q]$ and FARIMA$[p, d, q]$ models in terms of HIC are larger. The model we find to be the most suitable also is the FARIMA$[2, d, 2]$ process with parameters and p-value for the goodness-of-fit test (3.3) given in Table 5.3. Compared to the estimates obtained from the

Parameter	FARIMA$[2, d, 2]$
d	$0.143(0.016)$
a_1	$1.127(0.081)$
a_2	$-0.282(0.048)$
b_1	$-0.463(0.072)$
b_2	$-0.099(0.008)$
$p-$val	0.044

Table 5.3: Maximum-Likelihood estimates and standard deviation in parentheses for the FARIMA$[2, d, 2]$ model for the normalised Prague temperature anomalies. The last line gives the p-value of the goodness-of-fit test.

temperature residuals in Table 5.2, a somewhat larger estimate for the fractional difference parameter is obtained. The other estimates are compatible within one standard deviation. The p-value for the goodness-of-fit test is smaller than the one obtained previously. This indicates that not subtracting the polynomial trend leads to the FARIMA$[2, d, 2]$ being not an adequate model, according to the goodness-of-fit test (3.4)

Ensemble Generation and Trend Test We use the FARIMA$[2, d, 2]$ model and simulate an ensemble of 1 000 records. The polynomial (D.5) is fitted to each ensemble member and the three coefficients a, b and c are recorded. Figure 5.7 shows the marginal empirical

[2] We require an appropriate method for the detection of LRD to be a) sensitive and detect the LRD realisation as such and b) specific, i.e. to classify *only* the LRD realisations as stemming from a LRD process.

distribution functions of the single parameters and the corresponding parameter of the fit to the empirical record as a vertical line. If we consider separate one-sided tests for $a > 0$,

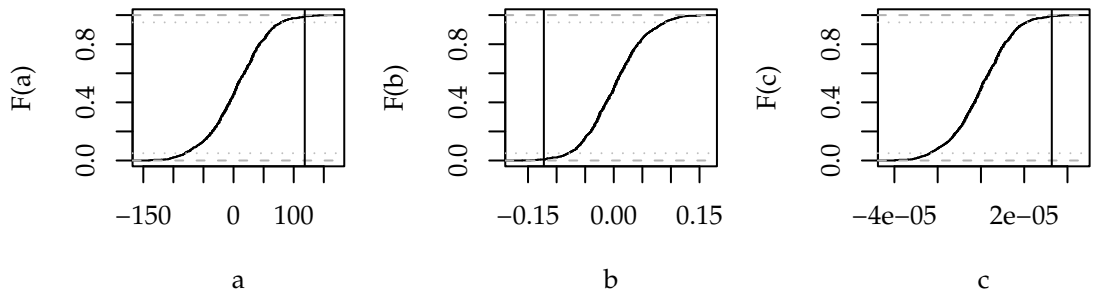

Figure 5.7: Marginal empirical cumulative distribution functions for the three parameters a, b and c from the second order polynomial fit. The vertical lines marks the corresponding parameter estimate for the empirical series. The horizontal dotted line mark the 5% and 95%-level.

$b < 0$ and $c > 0$, we find the estimates from the empirical record in a 5% critical region in all three cases. A joint p-value calculated by counting all members with a parameter estimate $\theta = (a, b, c)^\dagger$ more extreme than the observed results in $\widehat{p} = 0.002$. This leads to the rejection of H_0 and thus we find evidence against the hypothesis of no trend on a statistical basis.

This trend in the temperature record is probably a result of multiple effects. Besides the effect due to global warming (IPCC, 2001), an increasing temperature as a result of a growing city might also influence the record (Oke, 1973).

5.2 Northern Hemisphere Mean Temperature

In recent years the northern hemisphere mean surface air temperature has been investigated with respect to LRD (e.g., Smith, 2003; Cohn and Lins, 2005; Gil-Alana, 2005; Rybski et al., 2006). The basis of these analyses is a temperature record compiled at the Climate Research Unit (University of East Anglia) by Jones et al. (1999). The actual data set comprises monthly values from 1856 to 2005. A detailed description of the data set and some analyses are given by Jones et al. (1999) and Jones and Moberg (2003). While Gil-Alana (2005) studies seasonally adjusted monthly means, Cohn and Lins (2005) investigate annual means using low-order FARIMA$[p, d, q]$ models.

Gil-Alana as well as Cohn and Lins address the question whether the series shows an underlying trend in the mean, i.e. can be regarded as a realisation of

$$Y_t = a + bt + X_t, \tag{5.2}$$

were X_t is possibly a LRD process. Cohn and Lins (2005) proposed a trend test for the null hypothesis H_0 "$b = 0$" based on a modified likelihood-ratio test. Assuming a LRD process for X_t implies that the trend test is less sensitive than for SRD processes. In this case "small" trends are likely not to be detected. Particularly for X_t being a FARIMA$[1, d, 0]$ or FARIMA$[0, d, 1]$ process, H_0 cannot be rejected on a 5%-level of significance in their analysis. However, for the seasonally adjusted monthly record Gil-Alana proposed a LRD process including a linear trend in the mean, thus $b > 0$.

Here, we follow Cohn and Lins and study the annual mean series of temperature anomalies. The anomalies are taken as differences to the mean temperature of the pe-

riod from 1950 to 1979 (Figure 5.8, left). For the linear trend (5.2) parameter estimates

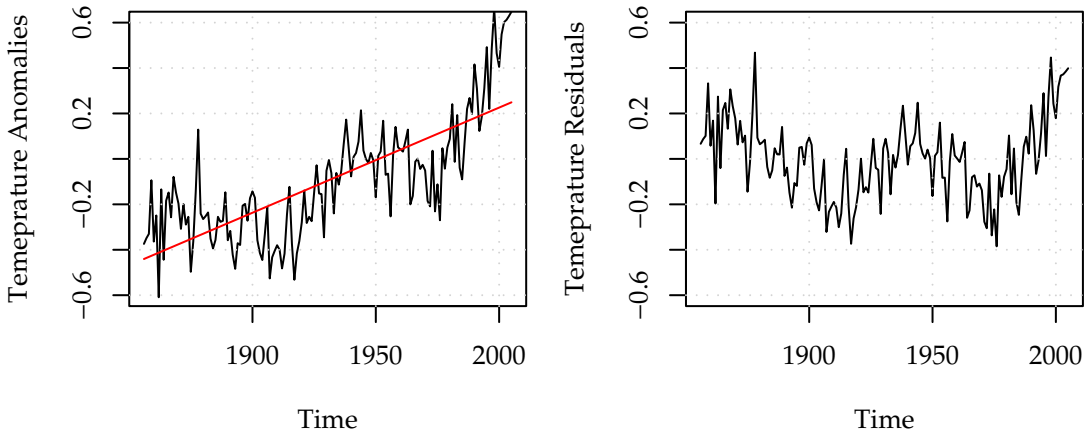

Figure 5.8: Northern hemisphere mean temperature anomalies (left) from 1856 to 2005 and a linear trend (red). The right panel shows the corresponding residuals.

$\widehat{a} = -9.031(0.638)$ and $\widehat{b} = 0.0046(0.0003)$ are obtained using ordinary least squares. The given standard deviation (in parentheses) is calculated under the assumption of an uncorrelated record. In the following, we use FARIMA$[p, d, q]$ models to describe the temperature anomalies as well as the residual series. The latter is obtained by subtracting the linear trend from the anomalies (Figure 5.8, right). The assumption of a simple linear trend made here is motivated by the analyses of Gil-Alana (2005) and Cohn and Lins (2005). A discussion of this assumption and of the use of LRD processes to model such a series closes this section.

5.2.1 Temperature Anomalies

As in Cohn and Lins (2005), we describe the temperature anomalies with FARIMA$[p, d, q]$ and ARMA$[p, q]$ models. In contrast to their approach, we use higher orders p and q, namely $0 < p < 4$ and $0 < q < \min(p, 3)$. Larger values for p and q are numerically not feasible.

Stochastic Modelling

The autocorrelation sequence of the northern hemisphere temperature anomalies are shown in Figure 5.9. For lags smaller than about 25 years, the autocorrelations are outside the 95% confidence interval for white noise. For a parametric description of the ACF, we proceed analogously to the strategy proposed in Section 5.1.1. First, we reduce the set of models with the goodness-of-fit test and the HIC criterion and compare the remaining models by means of the likelihood-ratio test or the simulation-based model selection.

Goodness-of-Fit Test On the basis of the goodness-of-fit test (3.4) with a 5%-level of significance, we cannot reject any model from the set and compare thus all models by means of the HIC criterion.

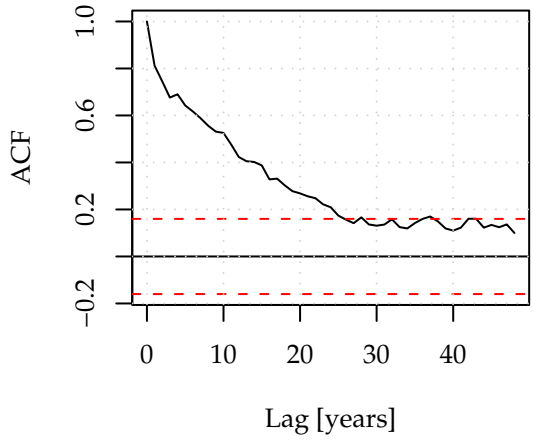

Figure 5.9: Autocorrelation sequence of the northern hemisphere temperature anomalies. The dashed red lines mark the asymptotic confidence intervals pertaining to white noise.

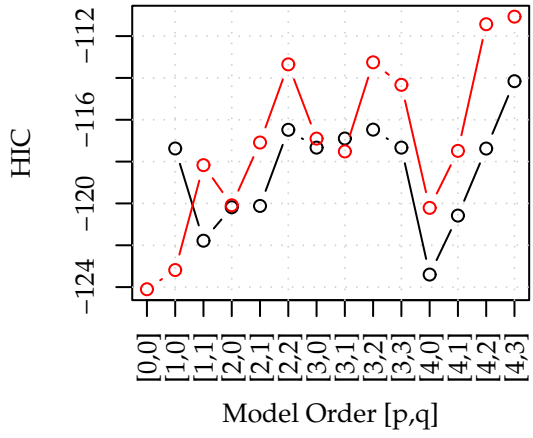

Figure 5.10: HIC for FARIMA$[p,d,q]$ (red) and ARMA$[p,q]$ (black) processes obtained for the mean northern hemisphere annual temperature anomalies.

ARMA$[p,q]$

	Model f	Model g	p-val
1	$[4,1]$	$[4,0]$	0.526
2	$[4,1]$	$[2,1]$	0.032
3	$[4,1]$	$[1,1]$	0.037
4	$[4,0]$	$[2,0]$	0.001

FARIMA$[p,d,q]$

	Model f	Model g	p-val
1	$[4,d,0]$	$[2,d,0]$	0.037
2	$[4,d,0]$	$[1,d,0]$	0.082
3	$[1,d,0]$	$[0,d,0]$	0.129

Table 5.4: p-values for a likelihood-ratio test of model g being an admissible simplification of model f. The ARMA$[p,q]$ (left) and FARIMA$[p,d,q]$ (right) models are represented only by their orders $[p,q]$.

HIC-Based Model Selection The HIC for the ARMA$[p,q]$ and the FARIMA$[p,d,q]$ models is shown in Figure 5.10. The FARIMA$[p,d,q]$ models with lowest HIC which we retain in the following are the FARIMA$[0,d,0]$ and the orders $[1,d,0]$, $[2,d,0]$ and $[4,d,0]$. Restricting the fractional difference parameter to zero, we find among the ARMA$[p,q]$ models ARMA$[4,0]$ and the orders $[1,1]$, $[4,1]$, $[2,0]$ and $[2,1]$ having the lowest HIC.

Likelihood-Ratio-Based Model Selection Among these models, we use the likelihood-ratio test at a 5%-level of significance to further reduce the set. Table 5.4 shows the p-values for tests within the ARMA$[p,q]$ and FARIMA$[p,d,q]$ class. Within the ARMA$[p,q]$ class (left table), we find the ARMA$[4,0]$ as an admissible simplification of the ARMA$[4,1]$ (line 1). Neither ARMA$[2,1]$ nor ARMA$[1,1]$ can be regarded as such an admissible simplification of ARMA$[4,1]$ (lines 2 and 3). According to the last line of the table we reject ARMA$[2,0]$. Thus ARMA$[4,0]$ remains as the most suitable SRD model here.

Among the FARIMA$[p,d,q]$ processes (right table), we find FARIMA$[1,d,0]$ as an admissible simplification of FARIMA$[4,d,0]$ (line 2), FARIMA$[2,d,0]$ has to be rejected as such (line 1). FARIMA$[1,d,0]$ can in turn be simplified with FARIMA$[0,d,0]$ (line 3). We thus remain with the latter as the most suitable LRD model.

Parameter Estimates and Spectral Densities

The corresponding parameter estimates of the two remaining models are given in Table 5.5. For the fractional difference parameter d in case of the FARIMA$[0,d,0]$ process, we

Parameter	FARIMA$[0,d,0]$	ARMA$[4,0]$
d	$0.622(0.064)$	–
a_1	–	$0.617(0.079)$
a_2	–	$0.126(0.093)$
a_3	–	$-0.104(0.093)$
a_4	–	$0.245(0.079)$
p-val	0.546	0.921

Table 5.5: Maximum-likelihood parameter estimates, asymptotic standard deviations and goodness-of-fit p-values for the FARIMA$[0,d,0]$ and ARMA$[4,0]$ process obtained from the northern hemisphere mean temperature anomalies.

find an estimate $\widehat{d} = 0.622(0.064)$ which is not in the stationary range ($-0.5 < d < 0.5$). If we thus choose to describe the northern hemisphere mean annual temperature with this mean stationary processes we are confronted with an infinite second moment (cf. Sec-

tion 2.2.2). Cohn and Lins do not give values for the FARIMA$[p, d, q]$ parameter estimates. It is thus not transparent whether they use stationary or variance instationary processes.

The roots z_i of the AR polynomial $\Phi(z)$ of the ARMA$[4, 0]$ model reveal that this process is stationary. There are two relaxating, and one oscillating component. With a modulus $\mod(z_1) = 1.071$ one root is close to the unit circle and thus the process is close to instationarity (cf. Section 2.2.1). The latter corresponds to a characteristic time scale of $\tau \approx 14.5$ years[3]. The dominating period of the oscillating component with about 4.8 years falls into the range of periods of the El Niño/Southern Oscillation which exhibits a broad peak with a maximum at about 4 years (Latif et al., 1998).

The periodogram and the spectral densities of the two models are shown in Figure 5.11. The periodogram exhibits a strong increase for frequencies $f < 0.1$, i.e. for peri-

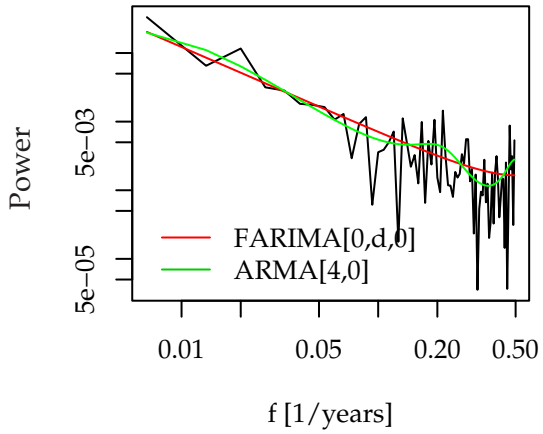

Figure 5.11: Periodogram of the northern hemisphere mean temperature anomalies (black) and spectral density of the ARMA$[4, 0]$ (green) and FARIMA$[0, d, 0]$ process (red) in double logarithmic representation.

ods larger about 1 year. Dominant are thus the first ten points of the periodogram. These increase is well described by the FARIMA$[0, d, 0]$ process. What is not well accounted for is the difference in power in the range $0.1 < f < 0.2$ (periods of 5 to 10 months) compared to the range $0.2 < f$. In this range the spectral density of the ARMA$[4]$ model shows a broad peak, representing the oscillating component and it follows the periodogram more closely than the FD spectral density.

Detection of Long-Range Dependence

We finally try to discriminate between the ARMA$[4, 0]$ and the FARIMA$[0, d, 0]$ as the most suitable SRD or LRD model, respectively. Because they are not nested, we refer to the simulation-based model selection depicted in Figure 5.12 (right). The observed value for the log-likelihood-ratio is in neither critical region, thus implying that we cannot discriminate the two models on the basis of this criterion. Both models perform equally well in describing the temperature anomalies. If we restrict, however, the analysis to stationary processes, we have to dismiss the FARIMA$[0, d, 0]$. In that case we would

[3] As characteristic time scale we denote the period τ during which the autocorrelation is reduce by a factor of $1/e$.

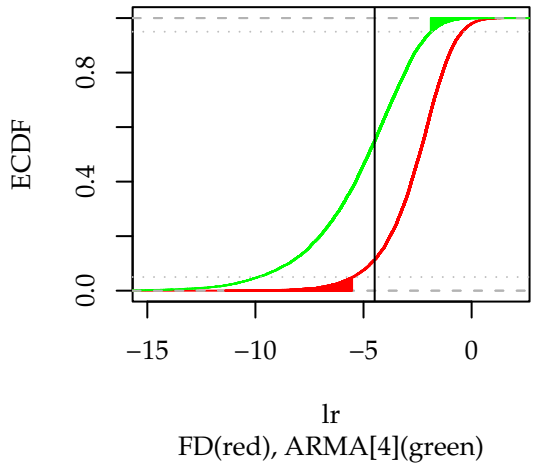

Figure 5.12: Simulation-based model selection (10 000 runs) between FARIMA$[0, d, 0]$ and ARMA$[4, 0]$ for the northern hemisphere temperature anomalies.

favour the ARMA$[4, 0]$ as the most suitable model for the northern hemisphere mean temperature record.

In the following, we repeat the analysis for the temperature series with a linear trend subtracted, i.e. the temperature residuals.

5.2.2 Temperature Residuals – Accounting for a Linear Trend

Stochastic Modelling

Figure 5.13 shows the autocorrelation sequence of the northern hemisphere temperature residuals. The autocorrelations enter the 95% confidence interval for white noise at about

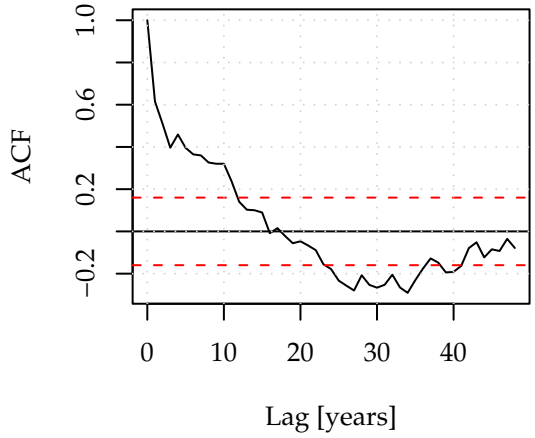

Figure 5.13: Autocorrelation sequence of the northern hemisphere temperature residuals. The dashed red lines mark the asymptotic confidence intervals pertaining to white noise.

11 years and thus earlier than for the temperature anomalies. For modelling the temperature residuals (Figure 5.8, right), we use the same orders of FARIMA$[p, d, q]$ and ARMA$[p, q]$ models and enter immediately into the model selection procedure.

	ARMA$[p,q]$		
	Model f	Model g	p-val
1	$[4,1]$	$[4,0]$	0.516
2	$[4,1]$	$[1,1]$	0.022

	FARIMA$[p,d,q]$		
	Model f	Model g	p-val
1	$[4,d,0]$	$[1,d,0]$	0.011
2	$[4,d,0]$	$[0,d,0]$	0.024

Table 5.6: p-values for a likelihood-ratio test of model g being an admissible simplification of model f.

Goodness-of-Fit Test Using the goodness-of-fit test at a 5%-level of significance, we can reject only the ARMA$[1,0]$ process as a suitable model.

HIC-Based Model Selection The HIC for the various processes is shown in Figure 5.14. Within the FARIMA$[p,d,q]$ class, we retain as models with lowest HIC the FARIMA$[0,d,0]$

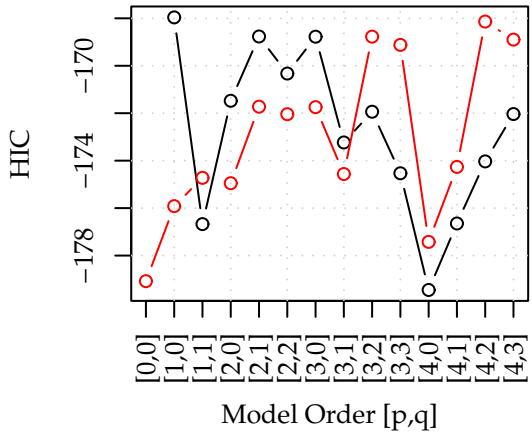

Figure 5.14: HIC for FARIMA$[p,d,q]$ (red) and ARMA$[p,q]$ (black) processes obtained for the northern hemisphere mean temperature residuals.

and orders $[4,d,0]$ and $[1,d,0]$. The ARMA$[p,q]$ models with low HIC which remain in the set for further analysis are the ARMA$[4,0]$ and orders $[1,1]$ and $[4,1]$.

Likelihood-Ratio-Based Model Selection The remaining models are compared using a likelihood-ratio test at a 5%-level of significance. The corresponding p-values for the ARMA$[p,q]$ and the FARIMA$[p,d,q]$ are given in Table 5.6. We find the ARMA$[4,0]$ being an admissible simplification of the ARMA$[4,1]$, while the ARMA$[1,1]$ is not. Thus, again the ARMA$[4,0]$ is the most suitable SRD model.

In case of the FARIMA$[p,d,q]$ family, we find neither the FARIMA$[1,d,0]$ nor the FARIMA$[0,d,0]$ being an admissible simplification of the FARIMA$[4,d,0]$. Thus in case of the temperature residuals, it is the latter which is the most suitable LRD model.

Parameter Estimates and Spectral Densities

The parameter estimates for these two models are given in Table 5.7. For FARIMA$[4,d,0]$, the estimate $\widehat{d}$ is compatible with $d=0$ within one standard deviation. We find, however,

Parameter	FARIMA$[4, d, 0]$	ARMA$[4, 0]$
d	$-0.186(0.200)$	
a_1	$0.628(0.198)$	$0.466(0.078)$
a_2	$0.112(0.097)$	$0.127(0.086)$
a_3	$-0.108(0.097)$	$-0.085(0.086)$
a_4	$0.279(0.084)$	$0.298(0.078)$
p-val	0.823	0.801

Table 5.7: Maximum likelihood parameter estimates, standard deviations, and goodness-of-fit p-values for the FARIMA$[4, d, 0]$ and the ARMA$[4, 0]$ process obtained from the mean northern hemisphere temperature residuals.

a pseudo-periodic component with a period of about 4.8 years. For the ARMA$[4, 0]$ process, we find approximately the same period of about 4.6 years. The longest relaxation time decreased from more than 14 years for the anomalies to below 10 years for the residual series. The interpretation of these components is to be treated with care because of the strong assumption of the underlying trend to be linear. This is discussed in Section 5.2.3.

The periodogram of the residuals and the spectral densities of the ARMA$[4, 0]$ and the FARIMA$[4, d, 0]$ process is depicted in Figure 5.15. Only the first three points of the peri-

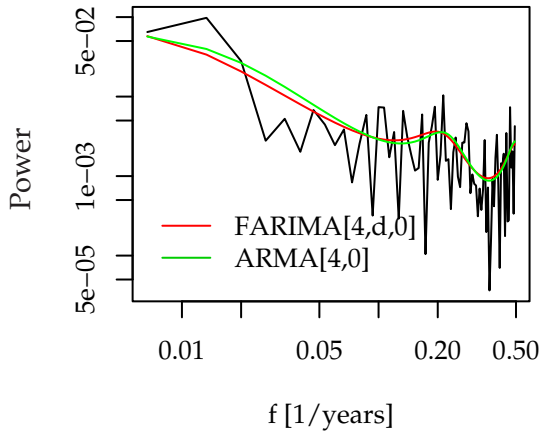

Figure 5.15: Periodogram of the northern hemisphere mean temperature residuals (black) and spectral density of the ARMA$[4, 0]$ (green) and FARIMA$[4, d, 0]$ process (red) in double logarithmic representation.

odogram exhibit an extreme increase in spectral power, different from the periodogram of the anomalies in Figure 5.11. As expected, the subtraction of the trend changed the low frequency power. Having this argument in mind, we might hypothesise, that subtracting a different – and possibly more adequate trend estimate – might also change the spectral power in the range of frequencies pertaining to the three points. The rivalry between LRD and a trend component becomes obvious here. For larger frequencies, the spectral densities for both models show the broad peak with the dominant frequency at about 4.6 years as already observed for the ARMA$[4, 0]$ model of the temperature anomalies.

Detection of Long-Range Dependence

FARIMA$[0, d, 0]$ and also ARMA$[4, 0]$ as models for northern hemisphere mean temperature residuals have been considered also by Bloomfield (1992). They carried out model selection using the AIC criterion (Section 3.2.2). In our analysis, we find the ARMA$[4, 0]$ and FARIMA$[4, d, 0]$ as the most suitable SRD and LRD models, respectively. They are nested and can be directly compared using the likelihood-ratio test. With a p-value of $p = 0.274$ we cannot reject the hypothesis that the ARMA$[4, 0]$ is an admissible simplification of the corresponding LRD process on any reasonable level of significance. Furthermore, the estimate for the fractional difference parameter for FARIMA$[4, d, 0]$ is compatible with zero within one standard deviation. Consequently, the northern hemisphere mean temperature residuals obtained from subtracting a linear trend are, according to this analysis, most suitably described with a SRD process.

5.2.3 Long-Range Dependence and the Linear-Trend Assumption

Cohn and Lins (2005) suggested a process allowing for a relatively large variability on large scales (low frequencies), i.e. a LRD process, to account for the variability of the residuals obtained from subtracting a linear trend. Due to its large variability for low frequencies, this process can also account for the full variations of the temperature series without subtraction of the linear trend. This does not necessarily imply the absence of a trend. It is also plausible that the misspecification of the trend form requires a highly variable (LRD) process to explain the residual variance.

The assumption of a linear trend in the mean made by Cohn and Lins is difficult to justify. On the one hand because there are various radiative forcing factors[4] being held responsible for a change in the global mean temperature (e.g. Kaufmann and Stern, 1997; Rahmstorf and Schellnhuber, 2006). Kaufmann and Stern (1997) account for these by explicitely including different forcing factors in a statistical description of the hemispheric mean temperatures. On the other hand one can expect a non-linear response of the temperature to the forcings. A way to account for this effect has been suggested by Smith et al. (2003). They include responses from physical models to the individual forcings in their analysis. The latter approach is a central idea of the fingerprinting put forward by, e.g, Santer et al. (1994) and Hasselmann (1997). Smith et al. (2003) used, however, only a bivariate temperature series, while fingerprinting aims at identifying multivariate space-time patterns resulting from climate models with different combinations of forcings included.

Without any plausible hypothesis for the underlying trend it is impossible to reliably identify the residual process. Especially the discrimination of slow variations, which can be considered as trend, and a LRD process is challenging. Many theoretical works and also applications with respect to this problem can be found in the recent literature (e.g., Teverovsky and Taqqu, 1997; Deo and Hurvich, 1998; Giratis et al., 2001; Craigmile et al., 2000, 2004; Kallache et al., 2005). In the example considered here, the long-time excursions from the linear trend curve depicted in Figure 5.8 is attributed to a LRD process by Cohn and Lins (2005). This decision is, however, not unambiguous. We showed that also an ARMA$[4, 0]$ process can account for that variability. Furthermore, if the trend component is constructed including non-linear responses from climate models to the forcings,

[4]Forcing factors related to anthropogenic activity are, e.g., changes in greenhouse gas concentration, tropospheric ozone and aerosols. Naturally variable factors are changes in the solar radiation and volcanic eruptions leading to spontaneous rise in aerosol concentration.

the northern hemisphere mean temperature can be well described without the use of LRD processes (Bloomfield, 1992).

5.3 Summary

We investigated the Prague daily maximum temperature and the annual northern hemisphere mean temperature for an underlying LRD process. With the parametric modelling approach (Section 4.4), we find a LRD model (FARIMA$[2, d, 2]$) to describe the Prague temperature anomalies best. The underlying LRD model is in line with other analyses of this record using DFA (Koscielny-Bunde et al., 1998). To challenge the strategies to detect LRD, we used the Prague temperature anomalies and constructed a SRD process which mimics the observed record as closely as possible. The analysis is then repeated for a realisation of this artifical process (Appendix A.3). Contrary to results from the DFA, the parametric modelling approach could correctly identify the realisation stemming from an SRD process. In this case, the full-parametric modelling approach with a proper model selection strategy turned out to be more specific than DFA.

With a parametric model at hand it was straightforward to construct a trend test based on simulations. The second order polynomial trend was found to be significant using a test which is rather conservative. Apart from global warming, the temperature increase in the Prague region might be partially attributed also to the urban heat island effect.

For the northern hemisphere mean temperature anomalies we found an ARMA$[4, 0]$ and a FARIMA$[0, d, 0]$ process to most suitably describe the record. For the latter, we found an estimate for the fractional difference parameter which is not in the stationary domain anymore. After subtraction of a linear trend, i.e. the temperature residuals, we could clearly identify the SRD ARMA$[4, 0]$ process as the most suitable model. The hypothesis of a linear trend is, however, problematic because it is likely that the warming in the last decades follows a more complex pattern.

Chapter 6

Modelling Run-off Records

The type of models commonly associated with run-off records are motivated by physical processes and are either conceptual models based on a cascade of reservoirs (e.g., Romano et al., 2001), or so-called distributed models describing the water balance in the soil on the basis of spatially distributed grid cells (e.g., Beven, 1998). The objective of these modelling approaches is, on the one hand, the monitoring and prediction of the water balance, especially the prediction of river run-off given a certain input, such as precipitation and temperature. On the other hand, process-based modelling is a means of gaining insight into the processes governing the dynamics. This type of models allows to study the response of a system to a possible change in external conditions, such as a change in temperature or precipitation patterns, e.g., due to climate change.

The objective of the stochastic modelling approach pursued here is of different nature. The goal is a suitable description of the autocorrelation function of the daily run-off. Such a description enables to obtain statistical estimates on the basis of an assumption which is more general than independent observations, i.e. the assumption of an underlying linear stochastic process. This influences the estimation of statistical characteristics, such as mean flows, return levels, trend parameters and the detection of structural breaks, such as change points or regime shifts. The main effect is a more appropriate specification of the uncertainty which is highly relevant especially with respect to return level and thus flood-risk estimation (e.g., Koutsoyiannis, 2003; Apel et al., 2004; Merz and Thieken, 2005).

We use models formulated in the framework of linear modelling and seek a parametric description of the ACF of run-off records. Three records are investigated in the following: daily mean run-off measured at the gauges Achleiten at the Danube River and Vilsbiburg at the river Große Vils and further monthly mean run-off at Tczew at the River Wisla. As well as for the temperature records, we assume periodic seasonal cycles, a possible trend and the so-called run-off anomalies being linearly superposed as specified in (5.1). A similar preprocessing is applied also to the run-off data (Appendix D.2): We use a Box-Cox transformation advocated for river run-off by Hipel and McLeod (1994) (Appendix D.2.2) to obtain a marginal distribution being closer to a Gaussian. We further subtract annual and, if present, weekly cycles as well as a polynomial trend. Weekly cycles in the run-off series are of anthropogenic origin. They are due to, e.g., industrial use, such as pumped-storage power station. Before addressing the run-off records, we discuss the motivation for describing run-off records using ARMA-type models.

ARMA-Type Models for Run-Off Records

Quite general arguments for using the linear stochastic models of the ARMA-type are given in Section 2.2.3. A more elaborated physically based motivation for river run-off was proposed by Salas and Smith (1981). We follow the representation given by Hipel and McLeod (1994) and denote the precipitation in year t as X_t. According to the conceptual representation sketched in Figure 6.1, we assume evaporation bX_t and infiltration aX_t through the soil into the groundwater storage depending linearly on the precipitation X_t. Thus the surface run-off flowing into the river is $(1-a-b)X_t = dX_t$. The groundwater storage at the start of year t is denoted as S_{t-1} and the groundwater contribution to the runoff Z_t is cS_{t-1}. It is now necessary to assume ranges for the con-

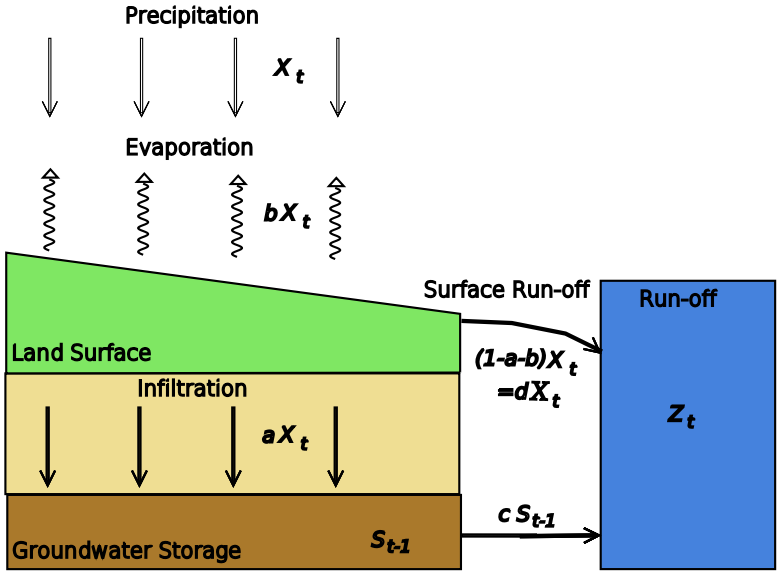

Figure 6.1: Conceptual model of a watershed, taken from Hipel and McLeod (1994).

stants: $0 \leq a, b, c, d \leq 1$ and $0 \leq a + b \leq 1$. The total runoff Z_t is now the sum of the direct surface flow and the groundwater contribution

$$Z_t = cS_{t-1} + dX_t. \tag{6.1}$$

This combined with the mass balance equation

$$S_t = (1-c)S_{t-1} + aX_t \tag{6.2}$$

yields

$$Z_t = (1-c)Z_{t-1} + dX_t - [d(1-c) - ac]X_{t-1}. \tag{6.3}$$

If we now subtract the means μ_Z and μ_X from the variables Z_t and X_t, respectively, we obtain

$$(Z_t - \mu_Z) = (1-c)(Z_{t-1} - \mu_Z) + d(X_t - \mu_X) - [d(1-c) - ac](X_{t-1} - \mu_X). \tag{6.4}$$

Assuming now the precipitation X_t being IID variables, (6.4) describes an ARMA$[1,1]$ process. Hipel and McLeod (1994) show also the results for the alternative assumptions of an AR$[1]$ (or ARMA$[1,1]$) model for the precipitation X_t. This results in an ARMA$[2,1]$ (or ARMA$[2,2]$) process for the runoff Z_t. Thus, a simple conceptual approach to modelling a watershed motivates on physical grounds ARMA$[p,q]$ processes as models for run-off records.

6.1 Danube Daily Mean Run-off at Achleiten

The gauge Achleiten at the Danube River is located close to the city of Passau in the south-east of Germany. The total catchment area of this gauge extends to 76 653 km^2. Daily mean run-off has been recorded from 01/01/1901 to 31/12/2003 and thus comprises $N_{\text{years}} = 103$ years or $N = 37\,620$ days. The average run-off taken over the whole observation period is about 1 430 m^3/s. Figure 6.2 shows the daily observed run-off at the gauge as time series (left) and as histogram (right). Before attempting to achieve a

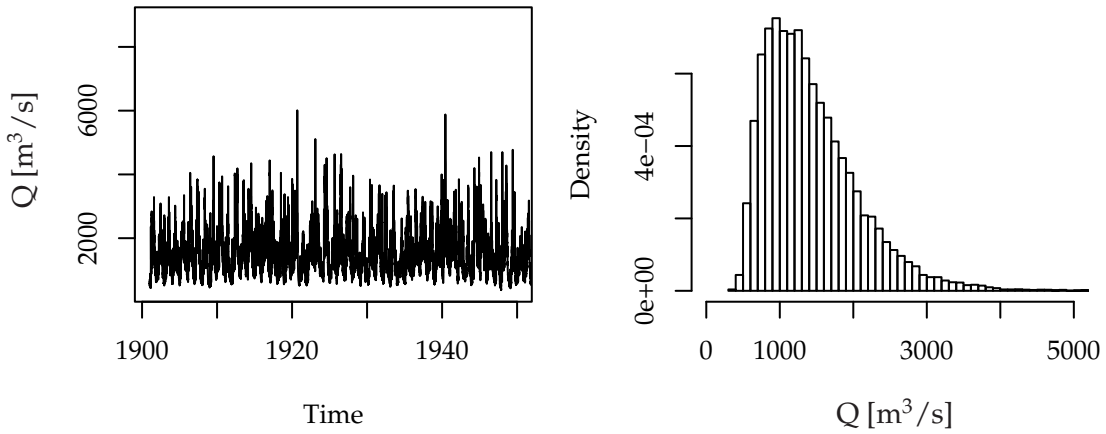

Figure 6.2: Daily mean run-off at the gauge Achleiten at the Danube River as time series (left) and histogram (right). For illustration, only for the period from 1901 to 1959 is shown in the left plot. The histogram is depicted for $Q < 5\,000$ m^3/s.

description using the stochastic models, we subject the empirical record to the preprocessing as mentioned above (cf. Appendix D.3.2). The record after the transformation and subtraction of the periodic cycles will be denoted as the anomaly series or simply anomalies.

6.1.1 Stochastic Modelling

The autocorrelation series estimated from the anomalies (Figure 6.3, left) reveals significant autocorrelations. These correlations are not falling inside the 95% confidence interval for white noise in the first one and a half years.

Additionally to estimating the ACF, we perform a first and second order detrended fluctuation analysis, DFA1 and DFA2 (cf. Section 2.3.3). Such an analysis has been carried out for the same record also by Koscielny-Bunde et al. (2006). We compare the result in a double-logarithmic plot to the asymptotic slopes of a SRD process ($H = 0.5$) or a LRD process with Hurst coefficient $H = 0.75$, corresponding to a fractional difference parameter $d = 0.25$ (Figure 6.3, right). For large scales, the plot rather suggests compatibility with the latter. As discussed in Appendix A.2 we cannot infer an underlying LRD process on the basis of DFA. Thus we use the FARIMA modelling approach in the following to achieve a description of the ACF.

The set of processes we start with contains FARIMA$[p,d,q]$ models with $0 \leq p \leq 8$ and $0 \leq q \leq \min(p,7)$ and additionally ARMA$[p,q]$ with orders $1 \leq p \leq 8$ and $0 \leq q \leq \min(p,7)$. Higher orders p and q are numerically not feasible. We estimate parameters for these models using the Whittle estimator (Section 2.3.2) and reduce the set of models

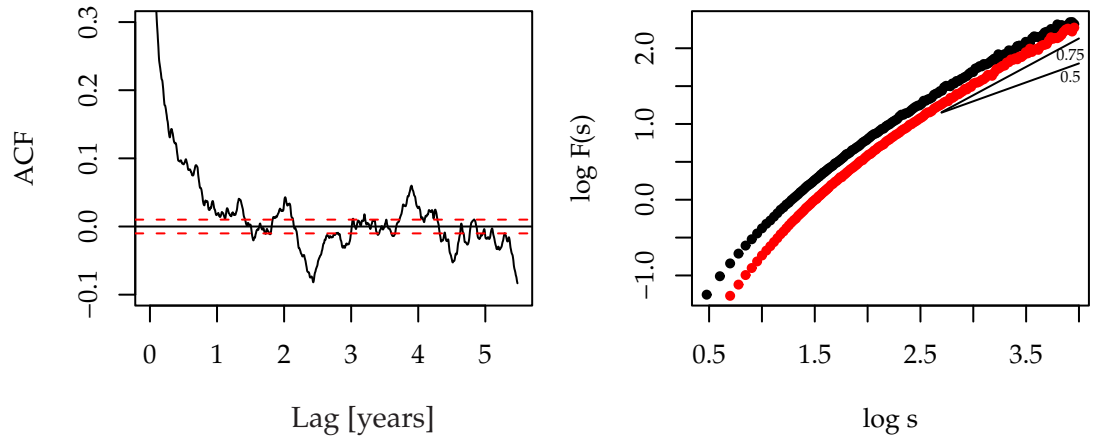

Figure 6.3: Autocorrelation series for the Danube run-off anomalies. The dashed red lines mark the asymptotic 95% confidence intervals for white noise (left). The right panel shows the DFA1 (black) and DFA2 (red) fluctuation functions in a double-logarithmic representation. The two solid lines mark the asymptotic slopes of a LRD ($H = 0.75$) and a SRD ($H = 0.5$) process.

according to the three steps outlined in Section 5.1.1: the goodness-of-fit test, the HIC-based model selection and, finally, the likelihood-ratio-based model selection.

Goodness-of-Fit

Initially, we reduce the model canon using the goodness-of-fit test (3.3). On a 5%-level of significance, we can reject the following processes:

- FARIMA$[0, d, 0]$ and
- ARMA$[p, q]$ with $(p, q) \in \{(p', q') \in \mathbb{N} \times \mathbb{N}^0 \mid (1 \leq p' \leq 4, q'), (5 \leq p' \leq 8, q' = 0), (6 \leq p' \leq 8, q' = 1), (p' = 8, q' = 2)\}$.

HIC-Based Model Selection

The remaining models are compared by means of the Hannan-Quinn information criterion (HIC, Section 3.2.2), see Figure 6.4. The models with lowest HIC we retain are the ARMA$[p, q]$ (green) with $(p, q) \in \{(7,6); (5,2); (7,5); (8,5); (8,6); (8,6); (5,3)\}$. Among the FARIMA$[p, d, q]$ models (red), we retain orders $(p, q) \in \{(7,6); (5,2); (7,5); (3,1); (2,2); (2,1); (8,5); (8,6)\}$.

Likelihood-Ratio-Based Model Selection

Among the models remaining, we use the likelihood-ratio test on a 5%-level of significance to further reduce the set of models. The p-values corresponding to the ARMA$[p, q]$ and FARIMA$[p, d, q]$ class are given in the left and right part of Table 6.1, respectively.

From the first two lines in the left table, we see that ARMA$[8, 6]$ finds an admissible simplification in ARMA$[7, 6]$ and ARMA$[8, 5]$. Both cannot be simplified by ARMA$[7, 5]$ (lines 3 and 4). ARMA$[5, 3]$ finds an admissible simplification in ARMA$[5, 2]$ (line 5), but the latter cannot be used as a simplification of neither ARMA$[7, 6]$ nor ARMA$[8, 5]$. We use the bootstrap-based model selection to discriminate between the two non-nested

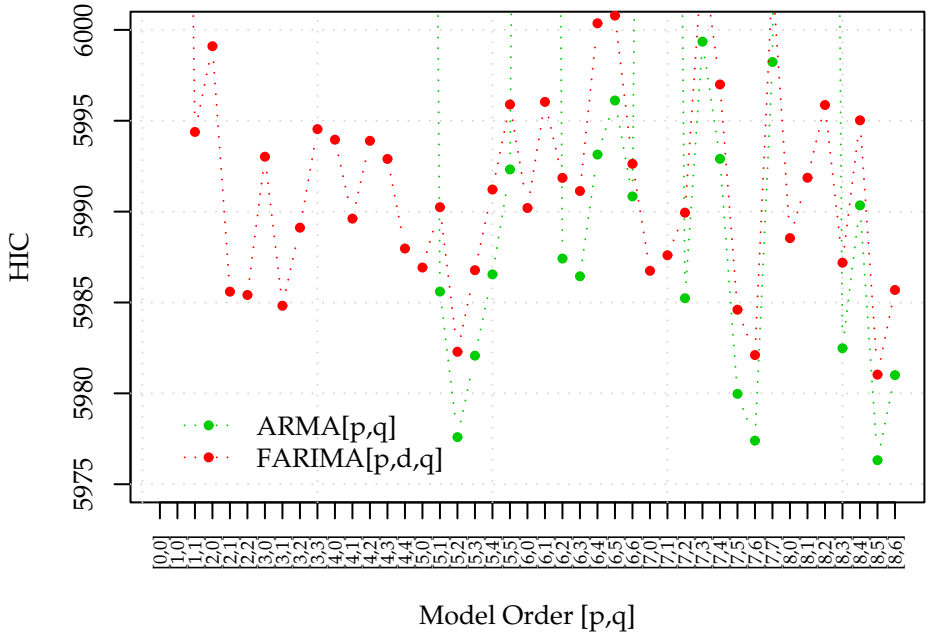

Figure 6.4: HIC for various ARMA$[p,q]$ (green) and FARIMA$[p,d,q]$ (red) models obtained for the Danube run-off anomalies. The model orders $[p,q]$ are plotted along the abscissa.

ARMA$[p,q]$

	Model f	Model g	p-val
1	$[8,6]$	$[8,5]$	0.868
2	$[8,6]$	$[7,6]$	0.293
3	$[8,5]$	$[7,5]$	0.003
4	$[7,6]$	$[7,5]$	0.007
5	$[5,3]$	$[5,2]$	0.635
6	$[7,6]$	$[5,2]$	$<$0.001
7	$[8,5]$	$[5,2]$	$<$0.001

FARIMA$[p,d,q]$

	Model f	Model g	p-val
1	$[8,d,6]$	$[8,d,5]$	0.818
2	$[8,d,6]$	$[7,d,6]$	0.288
3	$[8,d,5]$	$[7,d,5]$	0.004
4	$[7,d,6]$	$[7,d,5]$	0.007
5	$[5,d,3]$	$[5,d,2]$	0.638
6	$[7,d,6]$	$[5,d,2]$	$<$0.001
7	$[8,d,5]$	$[5,d,2]$	$<$0.001
8	$[7,d,6]$	$[3,d,1]$	$<$0.001
9	$[8,d,5]$	$[3,d,1]$	$<$0.001
10	$[7,d,6]$	$[2,d,2]$	$<$0.001
11	$[8,d,5]$	$[2,d,2]$	$<$0.001

Table 6.1: p-values for a likelihood-ratio test of model g being an admissible simplification of model f.

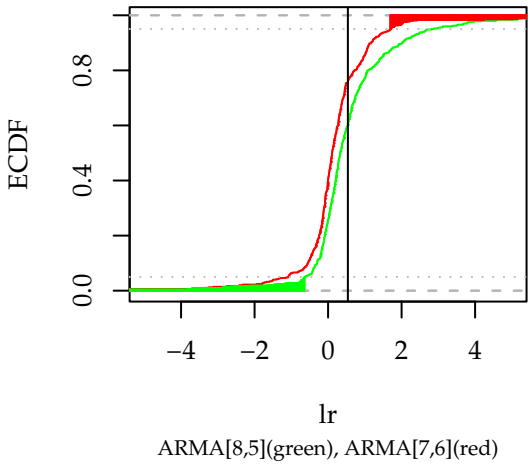

Figure 6.5: Bootstrap based model selection (500 runs) for ARMA$[8,5]$ against ARMA$[7,6]$. The vertical line mark the log-likelihood-ratio obtained for the observed series.

models (Figure 6.5). The distributions show a large overlap indicating, that both models perform equally well. We thus cannot discriminate the two models on the basis of the simulation-based model selection.

With the first 7 lines of the right table, we can follow the line of argument made for the ARMA$[p,q]$ models and find FARIMA$[7,d,6]$ and FARIMA$[8,d,5]$ as suitable LRD models. Those two models cannot be simplified by neither FARIMA$[3,d,1]$ (lines 8 and 9) nor FARIMA$[2,d,2]$ (lines 10 and 11) and thus not by FARIMA$[2,d,1]$. We thus keep FARIMA$[8,d,5]$ and FARIMA$[7,d,6]$ as the most suitable LRD models.

The likelihood-ratio test for the two ARMA$[p,q]$ models being admissible simplifications of the corresponding FARIMA$[p,d,q]$ yield p-values close to one. We can thus reduce the set of models to ARMA$[7,6]$ and ARMA$[8,5]$ and conclude that a SRD process is most suitable to describe the ACF of the run-off anomalies.

Parameter Estimates and Spectral Densities

Table 6.2 gives the corresponding parameter values for the two ARMA$[p,q]$ models remaining. The difference in the parameter estimate is largest for a_1. For the ARMA$[8,5]$, we find a_1 consistent with zero within one standard deviation. For ARMA$[7,6]$, it is the MA parameter b_6 which is not significantly different from zero. From the reciprocal roots of the autoregressive polynomial, we infer two pseudo-periodic components for each model (cf. Section 2.2.1). In both cases, we find frequencies (given in $1/\text{days}$) close to $f = 2/7 \approx 0.285$ and $f = 3/7 \approx 0.425$, which are related to a weekly cycle[1].

Figure 6.6 compares the periodogram of the run-off anomalies and the spectral densities of the ARMA$[7,6]$ (red) and ARMA$[8,5]$ model (green). For frequencies $f = 2/7$ and $f = 3/7$ one can observe the peaks in the spectral densities of the models, which are related to the two cyclic components (enlarge part in the right panel). Because these processes involve higher order autoregressive components modelling of a weekly periodicity is possible. A brief discussion on weekly cycles in river run-off in given in Section 6.2.3.

[1] Because the weekly cycle in general cannot be expected to be of sinusoidal form, higher harmonic components with frequencies $f = k/(7\ \text{days})$ are needed.

	ARMA[7,6]	ARMA[8,5]
a_1	0.135(0.079)	0.022(0.068)
a_2	0.633(0.088)	0.706(0.066)
a_3	0.572(0.062)	0.617(0.049)
a_4	0.453(0.066)	0.505(0.060)
a_5	−0.431(0.060)	−0.449(0.044)
a_6	−0.605(0.067)	−0.638(0.054)
a_7	0.232(0.027)	0.215(0.014)
a_8	–	0.011(0.007)
b_1	1.070(0.079)	1.184(0.067)
b_2	0.383(0.050)	0.446(0.042)
b_3	−0.381(0.033)	−0.381(0.033)
b_4	−1.030(0.048)	−1.093(0.046)
b_5	−0.707(0.068)	−0.769(0.052)
b_6	0.025(0.024)	–
p-val	0.204	0.207

Table 6.2: Maximum-likelihood parameter estimates, standard deviation and p-value for the goodness-of-fit test for the ARMA$[7,6]$ and ARMA$[8,5]$ processes obtained from the Danube run-off anomalies.

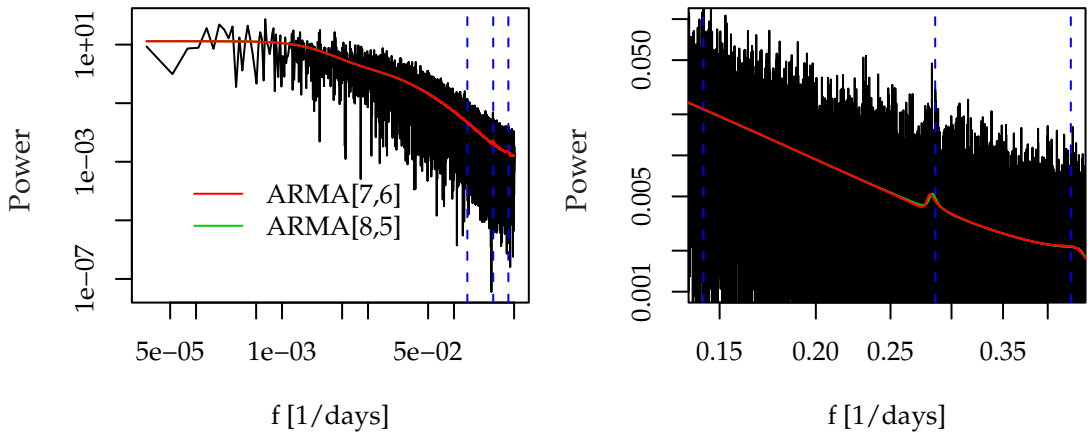

Figure 6.6: Periodogram of the Danube run-off anomalies (black) and spectral density of the ARMA$[7,6]$ model (red) and ARMA$[8,5]$ (green). The vertical blue lines indicate the frequencies corresponding to a weekly cycle: $f = 1/7, f = 2/7$ and $f = 3/7$. Note, that the green line is basically covered by the red one. The right plot shows only frequencies $1/7 \lesssim f \lesssim 3/7$.

6.1.2 Detecting Long-Range Dependence

Irrespective of the choice between the two ARMA$[p,q]$ processes, we can infer the underlying process as being SRD. This contradicts the impression obtained from the DFA result in Figure 6.3 and the results from Koscielny-Bunde et al. (2006). They report a Hurst exponent of $H = 0.82$ (corresponding to $d = 0.32$) and thus a LRD process underlying the Danube run-off anomalies at Achleiten obtained using DFA.

6.2 Große Vils Daily Mean Run-off at Vilsbiburg

The gauge Vilsbiburg at the river Große Vils is located in the Danube River catchment in the south-east of Germany about 80 km north-east of Munich. The total catchment area of this gauge extends to 320 km^2 and is thus about 240 times smaller than the Danube catchment drained at Achleiten. The daily mean run-off has been recorded from 01/11/1939 to 07/01/2002 and thus comprises $N = 22\,714$ days or approximately $N_{\text{years}} = 62$ years. The run-off averaged over the whole observation period is about 2.67 m^3/s. Figure 6.7 shows the daily observations at the gauge as time series (left) and as histogram (right). Again, the preprocessing described in Appendix D.3.3 is used prior to modelling. Be-

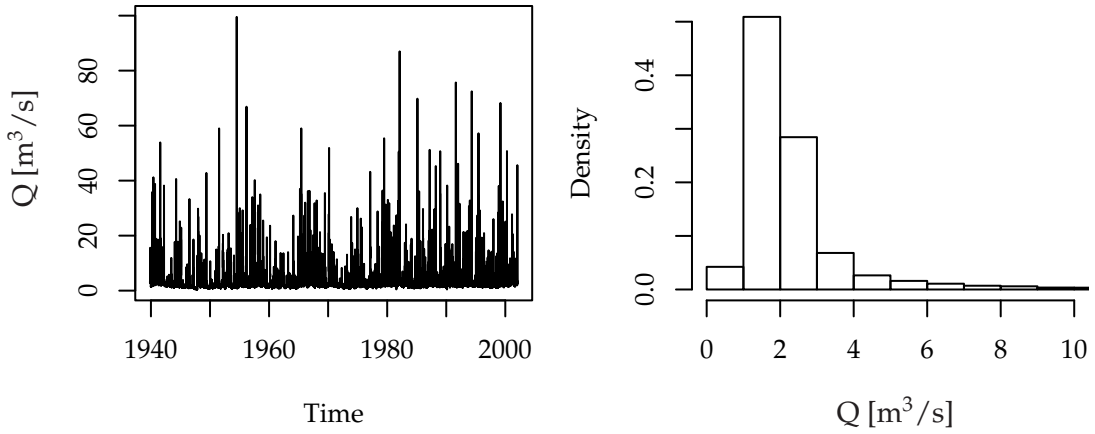

Figure 6.7: Daily run-off of the river Große Vils at the gauge Vilsbiburg as time series (left) and histogram for $Q < 10$ m^3/s (right).

cause we subtract a trend during the preprocessing, we refer to the resulting data set as residual series.

6.2.1 Stochastic Modelling

The estimated autocorrelation series (Figure 6.8) reveals significant autocorrelations not falling inside the 95% confidence interval for white noise in the first two years.

For a parametric description of the autocorrelation function, we start with a set of FARIMA$[p,d,q]$ processes with $0 \leq p \leq 8$ and $0 \leq q \leq \min(p,7)$ and additionally ARMA$[p,q]$ with orders $1 \leq p \leq 8$ and $0 \leq q \leq \min(p,7)$. Larger model orders are numerically not practical. The reduction of this set of models is carried out in the usual way (Section 5.1.1): the goodness-of-fit test, the HIC-based model selection and, finally, the likelihood-ratio-based model selection.

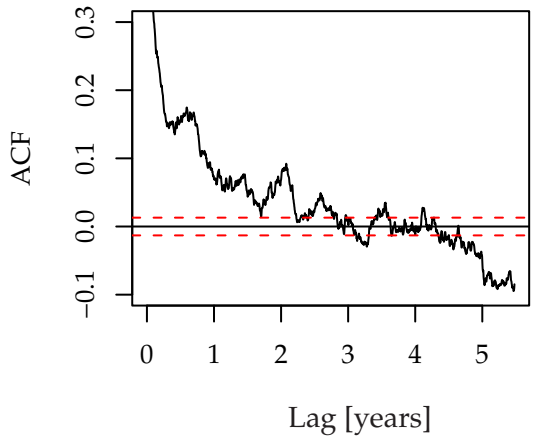

Figure 6.8: Autocorrelation series for the Große Vils run-off residuals. The dashed red lines mark the asymptotic 95% confidence intervals pertaining to white noise.

Goodness-of-Fit

In a first step we reduce the model canon using the goodness-of-test with a 1%-level of significance. A 5%-level would lead to a rejection of all models in the set. On the basis of this test, we can reject the following processes:

- FARIMA$[p,d,q]$ with $(p,q) \in \{(1,0),(2,0),(3,1),(4,4),(5,3)\}$ and
- all ARMA$[p,q]$ besides orders $(p,q) \in \{(7,6),(8,2),(8,4),(8,6)\}$.

HIC-Based Model Selection

Figure 6.9 compares the HIC for all ARMA$[p,q]$ (green) and FARIMA$[p,d,q]$ (red) processes in the set. We restrict the further analysis, however, to those models passing the goodness-of-fit test on a 1%-level. The four ARMA$[p,q]$ processes (green) which pass the goodness-of-fit test are marked with an outer green circle. Since these are only four models, we consider all of them in the following.

The FARIMA models with lowest HIC which we retain for further analysis are certainly FARIMA$[3,d,0]$ and in addition the model orders $[4,d,0]$, $[2,d,1]$,$[4,d,1]$, $[1,d,1]$ and $[2,d,2]$.

Likelihood-Ratio-Based Model Selection

Among the four ARMA$[p,q]$ and six FARIMA$[p,d,q]$ models with the lowest HIC, we perform the likelihood-ratio test or, if non-nested, a simulation-based model selection to obtain the most suitable SRD and LRD model. The p-values for the log-likelihood-ratio test are summarised in Table 6.3. We test ARMA$[7,6]$ and ARMA$[8,4]$ for being admissible simplifications of ARMA$[8,6]$. In both cases we cannot reject this null hypothesis (left table, lines 1 and 2). Furthermore, ARMA$[8,2]$ is an admissible simplification of ARMA$[8,4]$ (line 3). We thus remain with ARMA$[7,6]$ and ARMA$[8,2]$. These models are non-nested and are compared using the simulation-based approach (Figure 6.10, left). The distribution of the log-likelihood-ratios obtained for these models show a large overlap. We can, however, clearly reject the ARMA$[8,2]$ as the more suitable model compared to the ARMA$[7,6]$. Thus the latter remains as the most suitable SRD model.

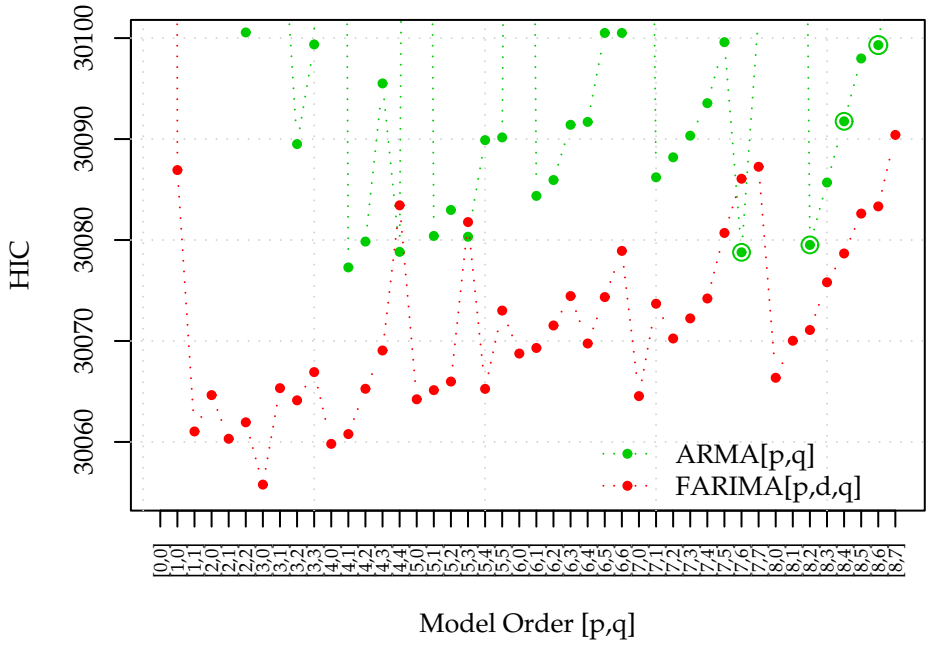

Figure 6.9: HIC for various ARMA$[p,q]$ (green) and FARIMA$[p,d,q]$ (red) models fitted to the Große Vils run-off anomalies. The model orders $[p,q]$ are plotted along the abscissa. The green circles mark the ARMA$[p,q]$ processes which pass the goodness-of-fit test.

ARMA$[p,q]$

	Model f	Model g	p-val
1	$[8,6]$	$[8,4]$	0.433
2	$[8,6]$	$[7,6]$	1
3	$[8,4]$	$[8,2]$	1

FARIMA$[p,d,q]$

	Model f	Model g	p-val
1	$[4,d,1]$	$[4,d,0]$	0.056
2	$[4,d,1]$	$[2,d,1]$	0.013
3	$[4,d,1]$	$[1,d,1]$	0.003
4	$[4,d,0]$	$[3,d,0]$	0.448
5	$[2,d,2]$	$[2,d,1]$	0.084
6	$[2,d,1]$	$[1,d,1]$	0.021

Table 6.3: p-values for a likelihood-ratio test of model g being an admissible simplification of model f.

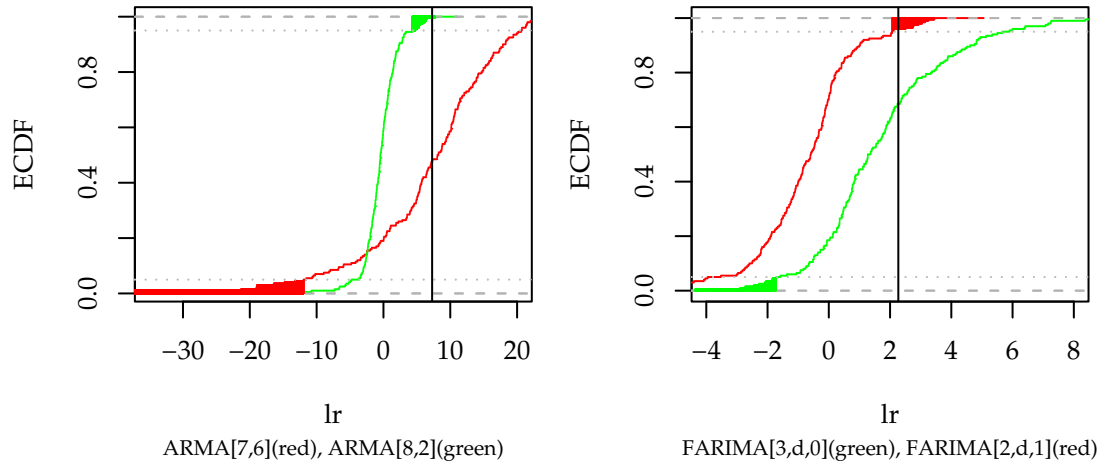

Figure 6.10: Bootstrap based model selection (200 runs) for ARMA$[7,6]$ against ARMA$[8,2]$ (left) and FARIMA$[3,d,0]$ against FARIMA$[2,d,1]$ (right). The vertical line mark the log-likelihood-ratio obtained for the observed series.

We further test FARIMA$[4,d,0]$, FARIMA$[2,d,1]$ and FARIMA$[1,d,1]$ for being admissible simplifications of FARIMA$[4,d,1]$ (right table, lines 1 - 3). This leaves us with FARIMA$[4,d,0]$ as the only admissible simplification. This model can be further simplified by FARIMA$[3,d,0]$ (line 4). FARIMA$[2,d,2]$ finds an admissible simplification in FARIMA$[2,d,1]$ (line 5) which cannot be further simplified by FARIMA$[1,d,1]$ (line 6). Thus, we remain with the two non-nested models, FARIMA$[3,d,0]$ and FARIMA$[2,d,1]$, which are compared using the simulation-based approach (Figure 5.4, right). The distributions of likelihood-ratios obtained for the two models largely overlap but we can reject the FARIMA$[2,d,1]$ in favour of the FARIMA$[3,d,0]$. The latter is thus the most suitable LRD model found.

Parameter Estimates and Spectral Densities

The parameter estimates and asymptotic standard deviation for the ARMA$[7,6]$ and FARIMA$[3,d,0]$ models are listed in Tables 6.4. Five parameter estimates obtained for the ARMA$[7,6]$ process are compatible with zero within one standard deviation. This indicates an over-parameterisation of this process. The largest relaxation has a characteristic time scale of about 1 year. Additionally, a pseudo-periodic component with a frequency close to $f = 2/7$ can be identified, indicating again that this model compensates for the remnants of the weekly cyle. For the FARIMA$[3,d,0]$ processes, we find a fractional difference parameter significantly different from zero and with $\widehat{d} \approx 0.44$, this LRD parameter is in the upper third of the admissible range for stationary processes.

Figure 6.11 shows the ARMA$[7,6]$ and the FARIMA$[3,d,0]$ spectral densities together with the periodogram of the Große Vils run-off residuals. The divergence between SRD and LRD manifests in the low frequency domain $f < 5 \cdot 10^{-4}$ (periods larger 54 years) and affects only about ten points in the periodogram. For frequencies $5 \cdot 10^{-4} < f < 5 \cdot 10^{-2}$ the ARMA$[7,6]$ spectral density closely follows the algebraic decay of the FARIMA$[3,d,0]$ model. A striking difference is in the vicinity of the weekly periodicities (Figure 6.11, right). Similar to the Danube record (Section 6.1.1), here the ARMA$[7,6]$ model shows as well a peak close to $f = 2/7$. This peak cannot be reproduced by the FARIMA$[3,d,0]$ and the remnants of the weekly cycle are not accounted for by this model. In the following,

	ARMA$[7,6]$	FARIMA$[3,d,0]$
d	-	0.439(0.016)
a_1	1.219(0.295)	0.415(0.017)
a_2	0.092(0.414)	−0.043(0.007)
a_3	0.048(0.210)	0.028(0.008)
a_4	0.123(0.205)	-
a_5	−0.912(0.187)	-
a_6	0.408(0.146)	-
a_7	0.022(0.100)	-
a_8	-	-
b_1	−0.364(0.295)	-
b_2	−0.506(0.186)	-
b_3	−0.382(0.226)	-
b_4	−0.413(0.180)	-
b_5	0.574(0.069)	-
b_6	0.096(0.137)	-
p-val	0.021	0.014

Table 6.4: Maximum likelihood parameter estimates, standard deviation and the p-value of the goodness-of-fit test for the ARMA$[7,6]$ and FARIMA$[3,d,0]$ processes obtained from the Große Vils run-off residuals.

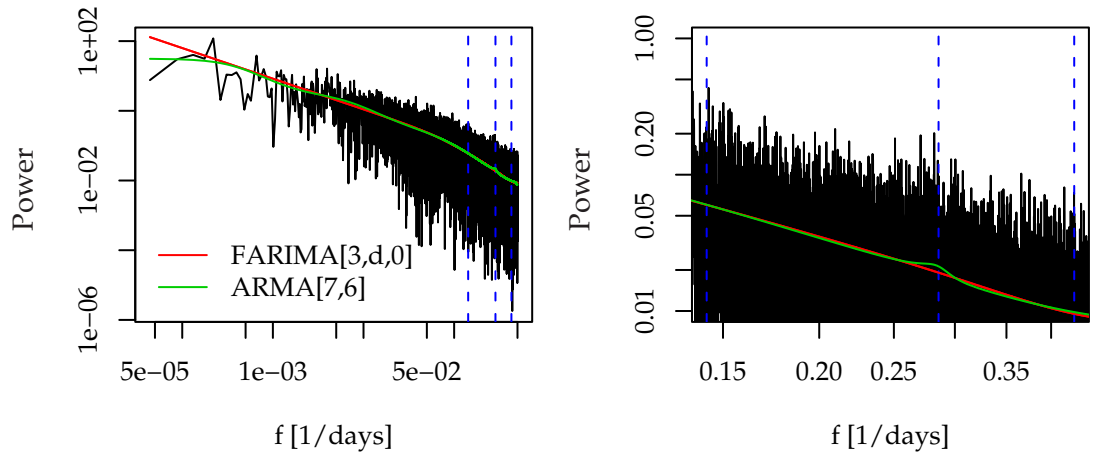

Figure 6.11: Spectral representation of the SRD (ARMA$[7,6]$, green) and LRD (FARIMA$[3,d,0]$, red) model together with the periodogram of the Große Vils run-off residuals. The vertical dotted lines mark the frequencies $f = \{1/7, 2/7, 3/7\}$. The right plot shows only frequencies $1/7 < f < 3/7$.

we try to discriminate these two remaining models.

6.2.2 Detecting Long-Range Dependence

As the most suitable SRD model, we identify ARMA$[7,6]$; as the most suitable LRD model the FARIMA$[3,d,0]$ process. Formulating the detection of LRD as a model selection problem, we aim to discriminate between the two alternatives and use the simulation-based model selection (Figure 6.12). The observed likelihood-ratio can be found in the 5% critical region of the distribution originating from the FARIMA$[3,d,0]$. The ARMA$[7,6]$ yields a very broad distribution of likelihood-ratio and the observed value is not near to the critical region. According to this result we are tempted to consider the ARMA$[7,6]$ as the

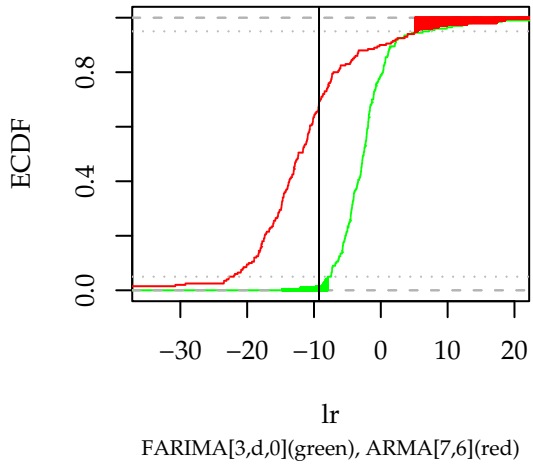

Figure 6.12: Simulation-based model selection (200 runs) for FARIMA$[3,d,0]$ against ARMA$[7,6]$. The vertical line marks the likelihood-ratio obtained for the observed series.

most suitable model for the Große Vils run-off anomalies. Having in mind, that five of the parameter estimates are compatible with zero within one standard deviation, we might suspect that this model is over-parametrised. It is furthermore difficult to decide whether the superior fit of the ARMA$[7,6]$ is a result of the difference in the range of the weekly periodicities (Figure 6.11) or because the record is rather compatible with a SRD model than a LRD model. Consequently, the result of the simulation-based model selection has to be interpreted with care in this case. discuss the problems resulting from the weekly cycle in the subsequent section.

6.2.3 Weekly Cycles in River Run-Off

In the periodogram of the daily run-off anomalies, obtained after subtraction of a mean annual cycle, we can observe an increased power at frequencies related to a weekly cycle. Being of anthropogenic origin, i.e. an influence resulting from adding and subtracting water according to human needs, this periodic component is not estimated and subtracted as easily as the annual one. As a consequences, we can still observe power at the weekly frequencies which influence the stochastic modelling. Models with a complexity (that allow for enhanced spectral power at the weekly frequencies, i.e. a pseudo-periodic component, are then preferred in the model selection. This might be problematic with respect to our modelling objectives (the detection of LRD).

So far, we have been mainly interested in the type of decay of the power at low frequencies. With the periodic components at a weekly scale, a second characteristic emerges, which we ask to be adequately represented by the model. A model capable of doing that might outperform one which adequately describes the decay at low frequencies but not not the weekly periodic components. This illustrates a typical problem of full-parametric modelling.

A simple alternative to avoid the difficulties introduced by a weekly periodic component is the investigation of monthly mean run-off, exemplified for a series of the Wisla River in Poland.

6.3 Wisla Monthly Mean Run-off at Tczew

The gauge Tczew at the Wisla (Vistula) River is located in the north of Poland, south of Gdansk. The total catchment area of this gauge extends to 194 376 km^2 and is about 2.5 times larger than the area drained at Achleiten. The monthly mean run-off has been recorded from November 1900 to October 1994 and thus comprises $N_{\text{years}} = 94$ years or $N = 1\,128$ months. Averaging over the whole observation period yields a mean run-off of about 1 042 m^3/s. Figure 6.13 shows the observed Wisla run-off as a time series (left) and as a histogram (right). For monthly mean run-off, the only periodic cycle to be removed

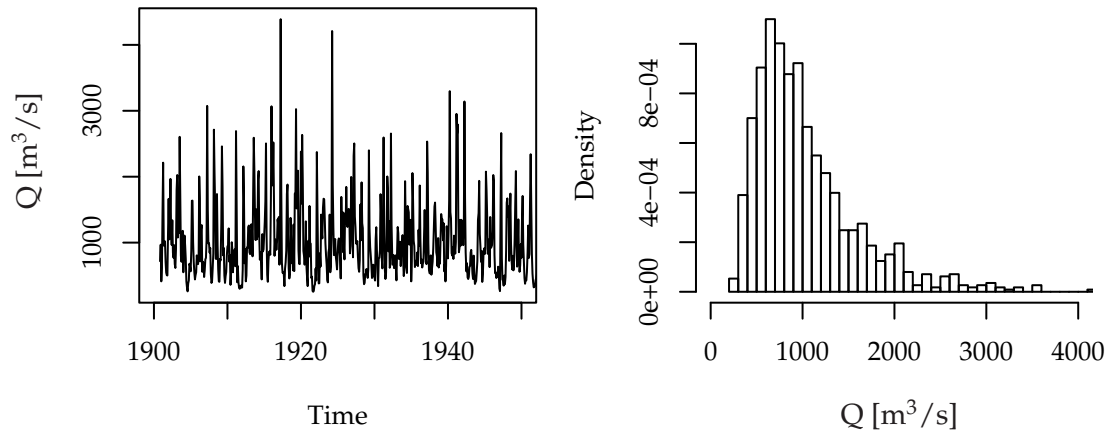

Figure 6.13: Monthly mean run-off for the Wisla River at the gauge Tczew as time series from the period 1901 to 1950 (left) and histogram (full record from 1900 to 1994) for $Q < 4\,000$ m^3/s (right).

is the annual one. Further preprocessing contains the Box-Cox transformation and the subtraction of a possible trend (Appendix D.3.4). The resulting data set is referred to as the residual series.

6.3.1 Stochastic Modelling

The autocorrelation series estimated from the residual record (Figure 6.14), reveals significant correlations not falling within the 95% confidence interval for white noise for lags smaller than 10 months.

To achieve a parametric description of the ACF, we start with a set of models containing FARIMA$[p, d, q]$ processes with orders $0 \leq p \leq 7$ and $0 \leq q \leq \min(p, 6)$ and additionally ARMA$[p, q]$ with orders $1 \leq p \leq 7$ and $0 \leq q \leq \min(p, 6)$, but the fractional

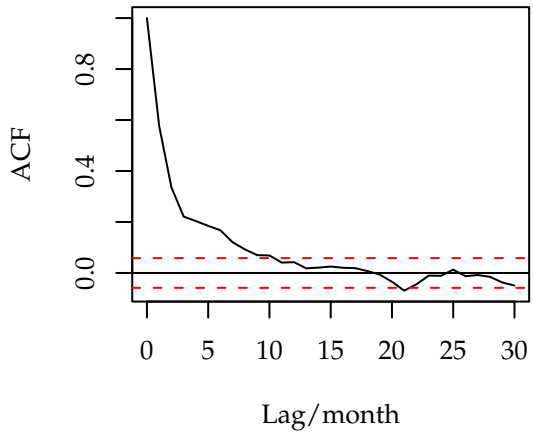

Figure 6.14: Autocorrelation sequence for the Wisla run-off residuals. The dashed red lines mark the asymptotic 95% confidence intervals pertaining to white noise.

difference parameter restricted to $d = 0$. Larger orders are not feasible due to numerical constrains. We estimate parameters for these models using the Whittle estimator (Section 2.3.2) and reduce the set of models in the same way as in the previous analyses (Section 5.1.1): the goodness-of-fit test, the HIC-based model selection and, finally, the likelihood-ratio-based model selection.

Goodness-of-Fit

Initially, we use the goodness-of-fit test (3.3) with a 5%-level of significance to reduce the set of models. On the basis of this test we can reject the following processes:

- None of the FARIMA$[p, d, q]$ but
- ARMA$[p, q]$ with $(p, q) \in \{(p', q') \in \mathbb{N} \times \mathbb{N}^0 \mid (5 \leq p' \leq 7, q' = 5), (6 \leq p' \leq 7, q' = 6), (p' = 7, q' = 3), (p' = 7, q' = 4)\}$.

HIC-Based Model Selection

The remaining models are compared by means of the Hannan-Quinn information criterion. Figure 6.15 shows the HIC for all ARMA$[p, q]$ (green) and FARIMA$[p, d, q]$ processes (red) in the set. We consider, however, only those passing the goodness-of-fit test. For orders $p < 3$ the FARIMA$[p, d, q]$ models show a smaller HIC than their ARMA counterparts. This changes for $p \geq 3$, indicating that for more complex models, in terms of the AR order p, the fractional difference component is not needed. The lowest HIC is found for the FARIMA$[1, d, 0]$ process. Within the fractional family of models it is followed by FARIMA$[1, d, 1]$ and FARIMA$[2, d, 0]$. For the ARMA family, the smallest HIC is attained by the ARMA$[1, 0]$ followed by the orders $[4, 0]$, $[3, 1]$, $[4, 1]$, $[5, 0]$ and $[3, 2]$.

Likelihood-Ratio-Based Model Selection

Among the models remaining, we use the likelihood-ratio test on a 5%-level of significance to further reduce the set of models. The p-values corresponding to the ARMA$[p, q]$ and FARIMA$[p, d, q]$ class are given in the left and right part of Table 6.5, respectively.

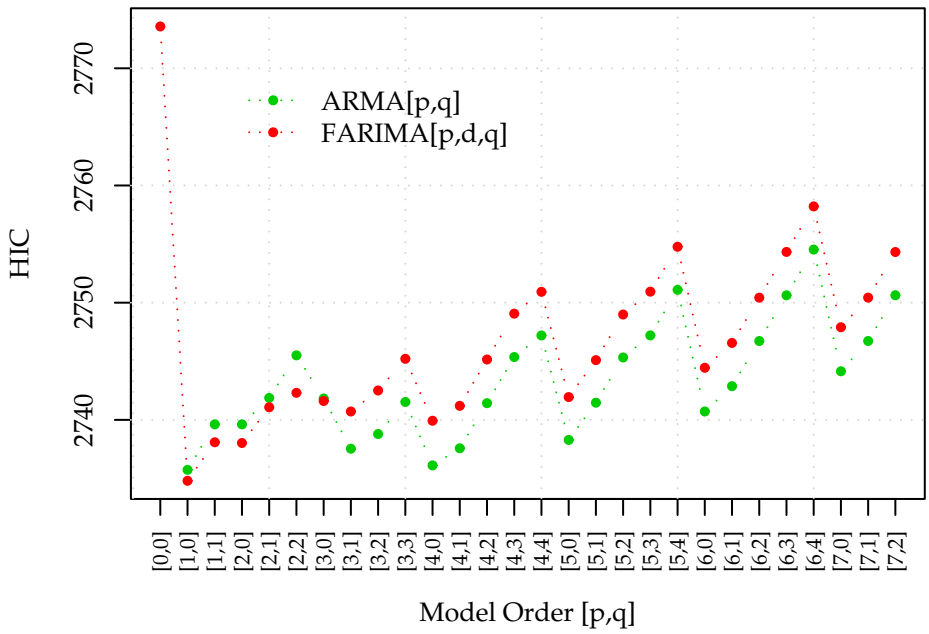

Figure 6.15: HIC for various ARMA$[p,q]$ (green) and FARIMA$[p,d,q]$ (red) models fitted to Wisla run-off residuals. The model orders $[p,q]$ are plotted along the abscissa.

Within the ARMA$[p,q]$ class (left Table), we find the ARMA$[4,0]$ as an admissible simplification of the ARMA$[5,0]$ (line 1) and also of the ARMA$[4,1]$ (line 3). The ARMA$[4,1]$ cannot be simplified by ARMA$[3,1]$ (line 4). The latter is, instead, an admissible simplification of ARMA$[3,2]$ (line 5). The remaining ARMA$[4,0]$ and ARMA$[3,1]$ cannot be further simplified by ARMA$[1,0]$ (line 2 and 6). Those two models are compared using the simulation-based model selection (Figure 6.16). The two distributions show a large overlap and we cannot discriminate the models. We thus retain both as suitable SRD models.

Among the three FARIMA$[p,d,q]$ models, we find the FARIMA$[1,d,0]$ as an admissible simplification of the two alternatives, the FARIMA$[2,d,0]$ and the FARIMA$[1,d,1]$. We thus retain it as the most suitable LRD model.

ARMA$[p,q]$

	Model f	Model g	p-val
1	$[5,0]$	$[4,0]$	0.189
2	$[4,0]$	$[1,0]$	0.010
3	$[4,1]$	$[4,0]$	0.118
4	$[4,1]$	$[3,1]$	0.049
5	$[3,2]$	$[3,1]$	0.104
6	$[3,1]$	$[1,0]$	0.019

FARIMA$[p,d,q]$

	Model f	Model g	p-val
1	$[2,d,0]$	$[1,d,0]$	0.411
2	$[1,d,1]$	$[1,d,0]$	0.434

Table 6.5: p-values for a likelihood-ratio test of model g being an admissible simplification of model f.

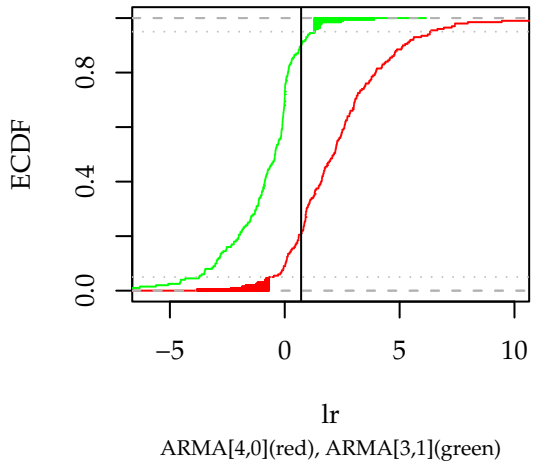

Figure 6.16: Simulation-based model selection (200 runs) for ARMA$[4,0]$ (red) against ARMA$[3,1]$ (green). The vertical line mark the log-likelihood-ratio obtained for the observed series.

Parameter Estimates and Spectral Densities

Table 6.4 gives the parameter estimates and corresponding standard deviation for the ARMA$[4,0]$, the ARMA$[3,1]$ and FARIMA$[1,d,0]$ processes. The estimates for the param-

	ARMA$[4,0]$	ARMA$[3,1]$	FARIMA$[1,d,0]$
d	-	-	$0.161(0.058)$
a_1	$0.571(0.030)$	$1.266(0.166)$	$0.401(0.068)$
a_2	$-0.016(0.034)$	$-0.431(0.097)$	-
a_3	$-0.014(0.034)$	$0.064(0.034)$	-
a_4	$0.092(0.030)$	-	-
b_1	-	$-0.698(0.165)$	-
p-val	0.646	0.696	0.571

Table 6.6: Maximum-likelihood parameter estimates, standard deviation and goodness-of-fit p-value for the ARMA$[4,0]$, ARMA$[3,1]$ and FARIMA$[1,d,0]$ processes obtained from the Wisla run-off residuals.

eter a_2 and a_3 for the ARMA$[4,0]$ model (Table 6.4) are consistent with zero within one standard deviation. This indicates that the model might be over-parametrised. From the reciprocal roots of the autoregressive polynomial (2.29) (Section 2.2.1) we find a pseudo-periodic component with a prevailling period of about five months ($f \approx 0.21$). This might be a result of an increased spectral power in the frequency range close to the semi-annual cycle ($f = 0.17$). The largest characteristic time scale of the relaxations is about 3.5 months. For the ARMA$[3,1]$ we find a pseudo-periodic component with a period of about nine months ($f \approx 0.11$). The largest characteristic time scale of this model is about six months. For the FARIMA$[1,d,0]$ model, we find the estimate for the fractional difference parameter $\widehat{d} = 0.161$ significantly different from zero but small compared to the LRD model for the Große Vils run-off residuals.

Figure 6.17 shows the spectral density of the three competing models (ARMA$[4,0]$, ARMA$[3,1]$ and FARIMA$[1,d,0]$ with the periodogram of the record in a double-logarithmic plot. The oscillating component manifests as a broad peak at $f \approx 0.2$ in the spectral density of the ARMA$[4,0]$ model. This oscillation with a period of about 5 months is not

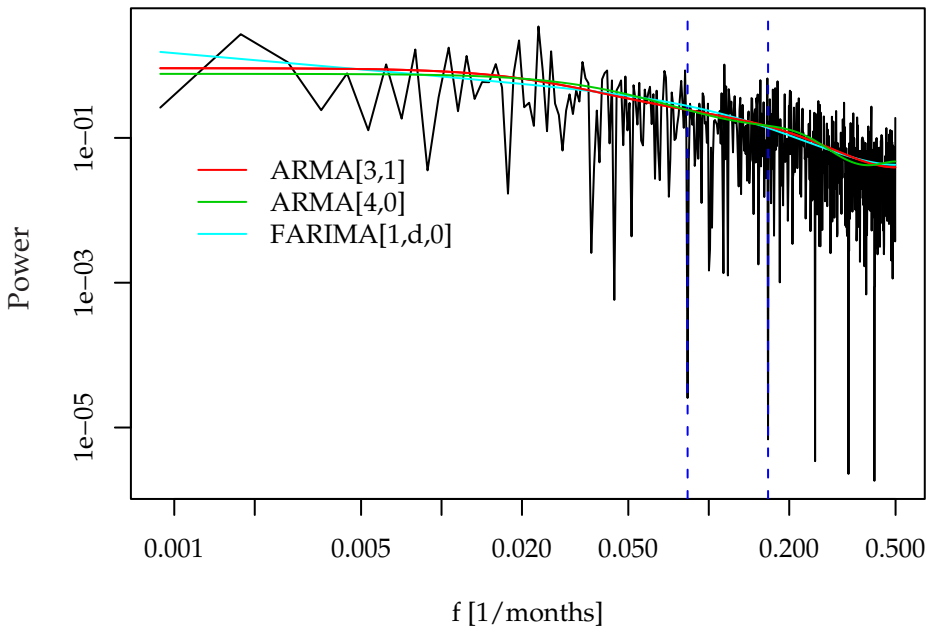

Figure 6.17: ARMA$[3,1]$, ARMA$[4,0]$ and FARIMA$[1,d,0]$ spectral density with the periodogram of the Wisla run-off residuals in a double logarithmic plot. The blue dashed lines mark the annual and semi-annual frequency.

reproduced by the other two models. The pseudo-periodic component of the ARMA$[3,1]$ model, instead, has a predominant frequency of $f \approx 0.11$ (period of approximately nine months). This feature can hardly be identified in the plot. Whereas there is no physical argument for the nine month periodicity, an increased spectral power in a semi-anual range is plausible. It might be either due to a remnant of the annual cycle or due to time periods of increased run-off being separated in time by about six months, e.g., snow melt in spring and increased rain-fall in autumn.

6.3.2 Detecting Long-Range Dependence

The remaining task is to discriminate the FARIMA$[1,d,0]$ from the ARMA$[3,1]$ and the ARMA$[4,0]$, respectively. Comparison is carried out using the simulation-based model selection (Figure 6.18). In both cases, the overlap of the distributions is large and we find the observed value close to the critical value of the FARIMA$[1,d,0]$ process. However, only in the comparison to the ARMA$[4,0]$, we find it inside the critical region. Thus, strickly speaking, only if we decide to choose the ARMA$[4,0]$ as the most suitable SRD model, we can reject the FARIMA$[1,d,0]$ and thus an underlying LRD process. As an additional argument, we might consider that two of the four parameters of the ARMA$[4,0]$ are compatible with zero within one standard deviation (Table 6.6). This might be taken as indication for an over-parametrisation of the ARMA$[4,0]$. On the basis of this argument, we can decide to prefer the ARMA$[3,1]$ as the most suitable SRD model. This decision further implies that we cannot discriminate between an underlying SRD and LRD process because ARMA$[3,1]$ and FARIMA$[1,d,0]$ are not distinguishable with the simulation-based model selection at a 5%-level of significance.

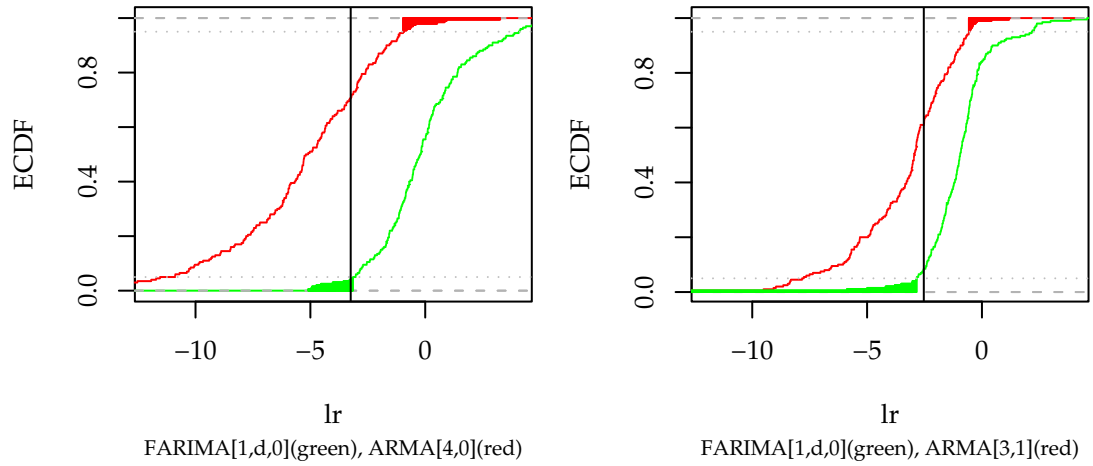

Figure 6.18: Simulation-based model selection (200 runs) for FARIMA$[1, d, 0]$ (green) against ARMA$[4, 0]$ (red) (left plot) and against ARMA$[3, 1]$ (red) (right plot). The vertical lines mark the likelihood-ratio obtained for the observed series.

6.4 Summary

We investigated run-off records from three different rivers and modelled their ACF with processes from the FARIMA$[p, d, q]$ class. Prior to modelling, a Box-Cox transformation was applied and periodic components, such as the annual and week cycle, were subtracted as well as low order polynomial trends. A description of the resulting run-off residual (or anomaly) series was achieved using various ARMA$[p, q]$ and FARIMA$[p, d, q]$ processes. The selection of the most suitable model was carried out sequentially by first eliminating those models which do not pass a goodness-of-fit test, second by reducing the set of models according to the Hannan-Quinn information criterion and finally by using the likelihood-ratio test or the simulation-based model selection for choosing between nested or non-nested models, respectively.

Further, we focused on the detection of long-range dependence, i.e. inferring an underlying LRD process from the data. For the gauge Achleiten, the result obtained with the parametric modelling approach is unambiguous: the most suitable model found is an SRD process, contrary to earlier claims based on DFA (Koscielny-Bunde et al., 2006). For the other two records, Große Vils at the gauge Vilsbiburg and Wisla at the gauge Tczew, the situation is ambivalent. For the Große Vils we find an ARMA$[7, 6]$ process as the most suitable SRD model and the most suitable LRD model is a FARIMA$[3, d, 0]$. Using the simulation-based model selection, we find the ARMA$[7, 6]$ as more suitable to represent the Große Vils run-off residuals. However, this process is able to describe spectral power at frequencies related to a weekly cycle. Thus its superiority might be due to an inadequate removal of this periodic component rather than due to the underlying process being SRD. The problem encountered for the Wisla record is different because data with a monthly resolution has been used and the weekly periodic component do not interfere. Here, the competing processes, ARMA$[3, 1]$ and FARIMA$[1, d, 0]$, are simply not discriminable with the simulation-based model selection at the chosen level of significance

Nevertheless, we can conclude from this analysis that LRD is not an ubiquitous and unambiguous phenomenon in river run-off records. With the Achleiten record, we further demonstrated that using a DFA log-log plot might yield deceptive results, i.e. suggesting LRD where SRD processes are more suitable to describe the observed record

(cf. also Appendices A.2 and A.3 and (Maraun, Rust, and Timmer, 2004; Metzler, 2003).

A further aspect that emerged during this study is the influence of a weekly periodic component. Due to its anthropogenic nature it is not as easily estimated and subtracted as the annual cycle. The periodicity is fixed but the amplitude might vary in an irregular way. Remnants of the weekly cycle can be described with autoregressive components of higher order. This feature can bias the detection of LRD.

Long-range dependence in run-off records is a characteristic which needs to be accounted for in trend tests (e.g., Kallache et al. (2005) and Craigmile et al. (2004)), structural breaks (e.g., Krämer and Sibbertsen, 2002), and also for determining confidence intervals for statistics obtained from the record (e.g., Koutsoyiannis, 2003). Those statistics can be the mean run-off or certain quantiles, such as the 0.99-quantile of the annual maxima records. The latter is denoted as the 100-year return level and is an important quantity for setting design values for dams, dykes and other water related constructions. In the following chapter we propose a bootstrap-based strategy to derive confidence intervals for these return level estimates. This strategy is based on the models found for the daily or monthly mean run-off records.

Chapter 7

Bootstrap Based Confidence Intervals for Return Level Estimates

The need for a further development of concepts and methods in risk assessment with respect to flooding has been expressed, among others, by Merz (2006): "Upcoming damage due to natural disasters can be seen as a consequence of todays decisions. In this respect, and taking rising damage due to natural disasters during the last decades into account (Münchner Rück, 2003), the contemporary methods for risk management are to be validated and further developed. This holds also for the risk of flooding. Globally, about 196 million people are affected by flooding; only between 1980 and 2000 about 170 000 people died (UNDP, 2004). Approximately half the losses due to natural disasters can be traced back to floodings (Kron and Tumerer, 2002).[1]" A fundamental building block in flood risk assessment is the determination of water levels exceeded on average once in a specified period. The estimation of these so-called return levels and their uncertainty plays an important role in hydrological engineering and decision making. Particularly the 100-year return level is frequently used as the basis for setting design values for hydraulic structures, such as bridges or flood protection buildings like dams or dykes. Because those constructions protect facilities of substantial value or are by themselves costly objects, it is certainly of considerable importance to have appropriate concepts of estimation and uncertainty assessment at hand. Otherwise severe damages, misallocation of public funds, or large claims against insurance companies might be possible.

In situations where common statistical approaches might not be applicable as usual, e.g., dependent records, specification of uncertainty bounds for a return level estimate cannot be made on the basis of the mathematically founded asymptotic theory for IID observations. An unjustified assumption of independent observations implies an underestimation of this uncertainty (e.g., Beran, 1994; Coles and Dixon, 1999; Koutsoyiannis, 2003). In the following, we focus on an improvement of uncertainty assessment of common statistical methods used for return level estimation. The strategy we suggest here is formulated within the framework of the block maxima approach (Appendix C.1) and is based on a bootstrap strategy. In Appendix C we suggest four different ways of generating bootstrap ensembles which can be used to evaluate the variability of a return level estimate. In a validation study, it turns out that one of these approaches performs reasonably well. It is studied in more detail regarding the variation of ensemble size ibidem.

The proposed concept explicitly takes autocorrelation into account and leads to an acceptable estimation of confidence intervals. It is based on a semi-parametric bootstrap

[1] This paragraph was translated from German.

approach involving a model for the ACF and a resampling strategy of the maxima series. The approach is briefly outlined in the following and exemplified with a case study: we estimate a 100-year return level with confidence intervals for the daily run-off series measured at the gauge Vilsbiburg at the river Große Vils in the Danube River catchment.

7.1 Sketch of the Bootstrap Approach

We initially estimate a return level from the annual maxima of the record under investigation. This is achieved by fitting a general extreme value distribution (GEV) to the maxima series (Appendix C.1). The resulting ML estimates of the GEV parameters can then be used to calculate a m-year return level (Appendix C.1.3). In the following, we focus on the estimation of a 100-year return level ($m = 100$). The likelihood approach allows to derive asymptotic confidence intervals for this estimate under the assumption of independent observations (Appendix C.1.3). We compare these asymptotic confidence intervals to the uncertainty bounds obtained using the semi-parametric bootstrap approach denoted as $bootstrap_{\mathrm{sp}}$, which is outlined in the following.

Sampling from a dependent process yields less information per observation regarding the marginal distribution than sampling from an independent process. This follows directly from the definition of dependence: given an observation we can derive from the dependence structure what the following observation is likely to be (Appendix C.2.1). Consider now the observed maxima series as a sample of a process with ACF $\rho(\tau)$ and a marginal distribution function F. If we knew this process, we could obtain many samples of equal length, estimate a 100-year return level from each sample, and study the variability of these estimate. This would lead to a distribution of return level estimates under the given conditions and thus opens up the possibility to obtain confidence intervals. Since we neither know the ACF nor the marginal distribution function, we replace both with their estimates. These are used to generate a long series from which samples of appropriate length are drawn. The distribution of return level estimates, which are obtained from each of the samples, is then used to quantify the variability of the return level estimator in terms of confidence intervals.

Once we have models for the ACF and the distribution of the maxima series (length N_{max}), we can generate the bootstrap ensemble (ensemble size N_{ens}) of artificial data sets according to the semi-parametric strategy. It is described in detail in Appendix C.2.5 and briefly sketched of the following:

1. Generate a long realisation with length $N = N_{\mathrm{ens}} N_{\mathrm{max}}$ from the model describing the ACF (FARIMA$[p, d, q]$).

2. Generate a series of same length by sampling with replacement from the empirical maxima series. If the original series was transformed prior to modelling its ACF, transform also the resampled series with the same transformation, e.g., a Box-Cox transformation (Appendix D.2.2) with the same parameter λ as obtained for the original series.

3. Use the iterative amplitude adjusted Fourier transform (IAAFT, Appendix C.2.4) to combine the two realisations.

4. Restore the original scale of measurement using the inverse Box-Cox transformation.

5. Split the long series in N_{ens} records of length N_{max}.

6. Estimate the 100-year return level (or 0.99 quantile) from each ensemble member.

7. Utilise the distribution of return level estimates to estimate a confidence interval.

In the following a case study exemplifies this strategy.

7.2 Case Study: Vilsbiburg/Große Vils

Consider the run-off record from the gauge Vilsbiburg at the river Große Vils. We studied this series already with respect to its ACF in Section 6.2. Here, we initially perform an extreme value analysis and model in a second step the ACF of the annual maxima series. Finally, we obtain and compare results from the $bootstrap_{\text{sp}}$ for two different model assumptions made for the ACF of the daily record: the most suitable SRD and LRD model found in Section 6.2.

7.2.1 Modelling the Maxima Distribution

In order to obtain an estimate for the 100-year return level, we carry out an extreme value analysis as described in Appendix C.1.3. This means to extract the annual maxima series and use ML estimation to determine the parameters and the asymptotic standard deviation of a GEV distribution. In order to test whether the estimated shape parameter $\widehat{\xi} = 0.04$ is significantly different from zero, we compare the result to a Gumbel fit by means of a likelihood-ratio test (Coles, 2001). With a p-value of 0.74 we cannot reject on any reasonable level the Gumbel distribution being a suitable model. The associated quantile and return level plots are shown in Figure 7.1 together with their 95% asymptotic confidence limits based on the assumption of independent observations.

According to (C.13), setting $m = 100$, we calculate a 100-year return level r_{100} and use the delta method (C.15) to approximate the standard deviation: $r_{100} = 97.7(7.9)$.

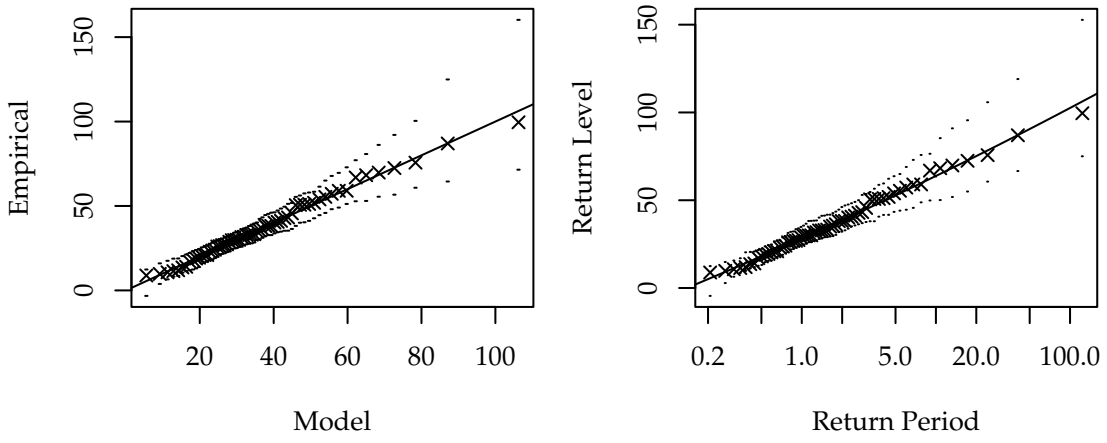

Figure 7.1: Result of the maximum-likelihood estimation of the Gumbel location and scale parameters ($\mu = 28.5(2.0)\text{m}^3/\text{s}$, $\sigma = 15.0(1.5)\text{m}^3/\text{s}$.) for the Vilsbiburg annual maxima series compared to the empirical maxima series in a quantile plot (left panel) and return level plot (right panel).

7.2.2 Modelling the ACF of the Maxima Series

For modelling the correlation structure of the maxima series, we revert to the results from Section 6.2: we found a FARIMA$[3, d, 0]$ as the most suitable LRD and an ARMA$[7, 6]$ as the most suitable SRD model. We exemplify the procedure using the FARIMA$[3, d, 0]$ process. The result obtained from the bootstrap is compared to the results obtained when assuming an ARMA$[7, 6]$ process underlying the daily run-off observations.

Using the specified model (FARIMA$[3, d, 0]$), a long series is generated with $N_{\text{long}} = 100 N_{\text{data}}$. This simulated daily series is partially back-transformed according to the transformations made in Appendix D.3.3: the overall mean and the seasonal cycles in mean and variance are added. The Box-Cox transformation is not inverted in this step. From the resulting record, we extract the "annual" maxima series using a block size of 365. We compare the ACF of this maxima series to the original maxima series (Figure 7.2). The latter has been Box-Cox transformed as well to achieve a comparable situation. The

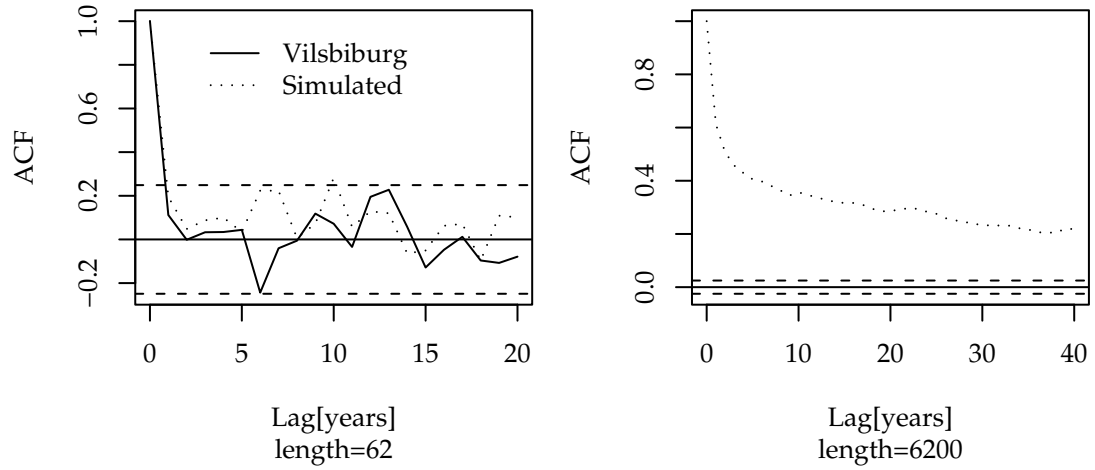

Figure 7.2: Autocorrelation sequence of the empirical maxima series and a section of same length cut out of the simulated series (left). The right panel shows the autocorrelation sequence of the full simulated maxima series. The 95% confidence intervals pertaining to white noise are marked as dashed lines.

autocorrelations of the original series as well as of a section of the simulated record basically fluctuate within the 95% confidence interval for a white noise process. However, the full maxima series from the simulated run of 6 200 data points exhibits significant autocorrelation (Figure 7.2, right). This indicates that, although an existing autocorrelation structure is not necessarily visible in a short sample of a process, it might still be present. Ignoring it, or taking it into account, yields different estimates of confidence intervals. The uncertainty of the ACF estimate is thus an uncertainty of the estimation of confidence intervals.

In the next step, we model the correlation structure of the simulated maxima series with a FARIMA$[p, d, q]$ process with orders $p_{\max} \leq p = 3, q_{\max} \leq q = 1$ [2]. Larger orders are not feasible due to numerical constraints. Among these models, we use the model selection strategies to choose an adequate process to represent the ACF of the simulated maxima series.

The smallest values for the HIC is attained for a FARIMA$[0, d, 0]$ model with $d = 0.398(0.010)$. The goodness-of-fit yields a p-value of $p = 0.319$ indicating a suitable

[2] Here $p_{\max}$ and $q_{\max}$ denote the model orders used for the maxima series.

model. A likelihood-ratio test against FARIMA$[1, d, 0]$ ($p = 0.515$), FARIMA$[1, d, 1]$ ($p = 0.060$) and FARIMA$[2, d, 1]$ ($p = 0.118$) yields that the FARIMA$[0, d, 0]$ is an admissible simplification of the three others and is thus an appropriate model for the ACF of the simulated maxima series. In the following, we assume that this process is also a suitable model for the ACF of the emprirical maxima series.

7.2.3 Confidence Intervals for Return Level Estimates

With the model for the ACF and the distribution of the maxima series, we generate the boostrap ensemble and obtain the 100-year return level estimates as described in Section 7.1. The result is shown in Figure 7.3 (top). From this distribution we obtain an estimate for a one-sided $\alpha 100\%$ confidence limit denoted as r^{α}. It is calculated as follows: sort the R estimates $\widehat{r}_r$, $r = 1, \ldots, R$ in ascending order and take the estimate at position $(N+1)\alpha$. This estimate is denoted as $\widehat{r}_{((N+1)\alpha)}$. With an ensemble size $N_{\text{ens}} = 9\,999$ we ensure $(N+1)\alpha$ being an integer for common choices of α (Davison and Hinkley, 1997).

The 95%-quantiles of the bootstrap ensemble ($\widehat{r}^{0.95}_{\text{boot}} \approx 135\text{m}^3/\text{s}$) and the asymptotic distribution ($r^{0.95}_{\text{asymp}} \approx 110\text{m}^3/\text{s}$) are marked as vertical lines (Figure 7.3). The bootstrap 95% confidence level $\widehat{r}^{0.95}_{\text{boot}}$ clearly exceeds the quantile expected from the asymptotic distribution confirming the substantial increase in uncertainty due to dependence. The difference between the two limits is larger than 20% relative to the return level estimate. Furthermore, the upper tail of the bootstrap ensemble decays slower than the tail of the asymptotic distribution. The interpretation of such a confidence level is the following: In 95% of 100-year return level estimates the expected ("true") 100-year return level will not exceed the 95% confidence limit.

If we use the ARMA$[7, 6]$ process as a model for the daily run-off, instead of the FARIMA$[3, d, 0]$, and repeat the procedure outlined in Section 7.1, we obtain an AR$[1]$ process with $\widehat{a}_1 = 0.267(0.026)$ as model for the ACF of the annual maxima. With this model the semi-parametric bootstrap procedure yields a different result for the distribution of return level estimates as depicted in Figure 7.3 (bottom). The difference between the bootstrap distribution and the asymptotic distribution is here smaller, especially the 95% confidence limits differ only by less than 2% with respect to the return level estimate.

7.3 Summary

We consider the estimation of return levels from annual maxima series using the GEV as a parametric model and maximum likelihood (ML) parameter estimation. If we explicitly account for autocorrelation in the records, a substantial increase in uncertainty of the flood return level estimates is revealed in cases where the underlying process is assumed to be LRD. In the standard uncertainty assessment, i.e. the asymptotic confidence intervals based on the Fisher information matrix or the profile likelihood, autocorrelations are not accounted for. The resulting confidence intervals might be too small to reflect the actual variability of the estimator for correlated records. On the way to fill this gap, we study and compare four bootstrap strategies for the estimation of confidence intervals in the case of correlated data (Appendix C). The semi-parametric bootstrap strategy outperforms the three other approaches and might be of considerable value for flood-risk assessment of water management authorities to avoid floodings on the one hand or misallocation of public funds on the other hand. Furthermore, we expect the approach to be applicable also in other sectors where an extreme value analysis with dependent extremes has to be carried out.

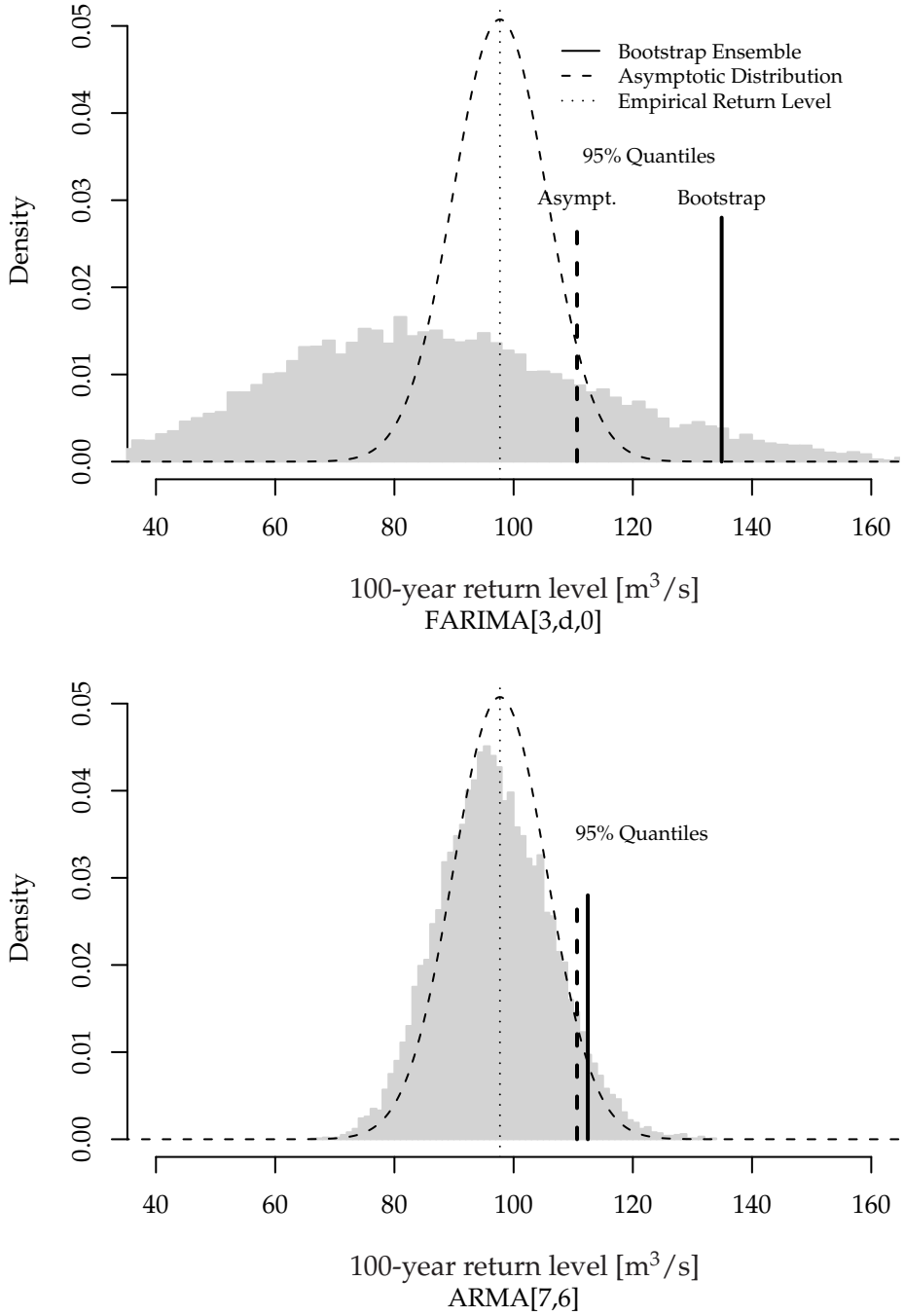

Figure 7.3: Frequency distribution of the 100-year return level estimates from the $bootstrap_{\text{sp}}$ ensemble with 9 999 members based on the FARIMA$[3, d, 0]$ (top) and the ARMA$[7, 6]$ model (bottom) for Vilsbiburg as histogram (grey) compared to the asymptotic distribution of the ML estimator derived from the Fisher information matrix (dashed). The 100-year return level estimate from the empirical maxima series is marked as dotted vertical line. The 95%-quantiles of the asymptotic and bootstrap distributions are indicated as dashed and solid vertical lines, respectively.

The practicability of the approach is illustrated for the gauge Vilsbiburg at the River Vils in the Danube catchment in southern Germany. We derived a 95% confidence limit for the 100-year flood return level on the basis of the most suitable SRD as well as LRD model. For the latter, the confidence limit has been found to be about 20% larger than the one derived from the asymptotic distribution. This is a dimension worth being considered for planing options. This result depends in a delicate way on the model selected for the daily series and accentuates the importance of a reliable model selection strategy. Falsely concluding LRD will result in an overestimation of the uncertainty limits and might lead to dykes and dams being constructed larger than necessary for the desired level of safety. Conversely, ignoring LRD if present, leads to smaller constructions and results in a misleading perception of the level of safety.

Chapter 8

Conclusions and Outlook

8.1 Summary and Conclusions

It is desirable to reduce potential threats or impacts that result from the variability of nature, such as droughts or heat waves that may lead to food shortage, or the other extreme, floods that may lead to severe damage of infrastructure. To prevent such catastrophic events, it is necessary to understand, and to be capable of characterising, nature's variability and the underlying dynamics. A prominent example in this respect is the calculation of design values for flood protection buildings.

Typically one aims to describe the dynamics of geophysical records with physically motivated systems of differential equations. There are, however, situations where this does not support the objectives, or is not feasible. These situations are typically those where little is known about the equations governing the underlying dynamics, or the system is so complex that the model parameters cannot be identified. In such situations it is beneficial to regard certain influences as random, and describe them with stochastic processes. In this thesis, we focus on a description of the autocorrelation function, and thus concentrate on linear stochastic processes to represent the variability of run-off and temperature records. Special emphasis is put on the detection of long-range dependence, which is an algebraic (i.e. slow) decay of the autocorrelation function. It is a priori not evident what process out of the large class of linear stochastic processes is adequate for modelling the record under consideration. Choosing an appropriate model out of a large canon of potentially suitable models is a crucial task. This choice influences subsequent analyses and might thus have far reaching consequences. For example, tests for trends are sensitive to the assumptions made for the underlying process describing the natural variability. The latter is important with respect to temperature series and the detection of climate change. A different example of considerable interest is the variability of river run-off. Here, an inappropriate model choice might imply an underestimation of the uncertainty of flood return level estimates. This leads to a reduction of design values for dykes and consequently to less protection than expected.

Along the guiding questions raised in the introduction, the following conclusions can be drawn:

- **What is an adequate strategy for choosing between different models which may be used to describe the autocorrelation structure? How can long-range dependence be detected in a reliable way?**

  Model selection for nested linear stochastic models of the FARIMA-type is commonly carried out with the likelihood-ratio test or Akaike-type information crite-

ria. If models are non-nested, one can revert to a common "larger" model, in which both alternatives are nested in, and use the likelihood-ratio test to find the suitable simplification. This method of choosing between non-nested models has potentially lower power than a direct comparison and can easily lead to the inability to discriminate between the models. We suggested a simulation-based approach to directly compare non-nested models of the FARIMA-type (Chapter 3). This approach follows the procedure of a statistical test, with the log-likelihood-ratio as the test statistic. Its distribution is obtained via simulations using the two models under consideration. For two simple models and different parameter values, we investigated how reliable estimates of p-value and power are, which are obtained from the simulated distributions of the test statistic. The result turned out to be dependent on the parameter values. In many cases the estimates allow an adequate model selection to be established. Under these circumstances, this approach has potentially more power than the alternative way, which involves a "larger" model. Furthermore, an important feature is that it immediately reveals the ability or inability to discriminate between the two models under consideration. With respect to the estimation of, e.g., flood return levels, this is an important piece of information. Especially if a long-range and a short-range dependent model cannot be distinguished, it is important to recognise this problem and account for it in the uncertainty assessment.

Regarding the detection of long-range dependence (Chapter 4), we suggested to rephrase the problem as a model selection task, i.e. comparing the most suitable long-range dependent and the most suitable short-range dependent model. Approaching the task in this way requires a) a suitable class of long-range and short-range dependent models along with suitable means for parameter estimation and b) a reliable model selection strategy, capable of discriminating also non-nested models. With the flexible FARIMA-class, the Whittle estimator, and the model selection strategies presented and developed in Chapter 3, those requirements are fulfilled. and it is possible to evaluate whether a proposed long-range dependent model is indeed more adequate for the observations than the most suitable short-range dependent model found. This direct confrontation of the two models with characteristics being mutually exclusive presents a new way of looking at the problem of detecting long-range dependence.

The performance of the suggested approach is tested in a simulation study. Hereby, the processes underlying the simulated records are constructed in a way to provide a challenging example with respect to the detection of long-range dependence. Compared to detrended fluctuation analysis (DFA, a heuristic method frequently used to detect long-range dependence), this parametric modelling approach is more specific, i.e. realisations of short-range dependent process are more likely to be detected as such. The imprudent use of DFA is prone to falsely concluding long-range dependence. This, in turn, makes the latter a seemingly ubiquitous phenomenon and bears the risk of trivialising this concept (cf. Appendix A).

- **What is the effect on trend detection caused by autocorrelation?**

  With the analysis of the Prague daily maximum temperature and the northern hemisphere mean temperature, we could demonstrate to what extent the modelling approach can be profitable for trend detection strategies (Chapter 5). In this case, we find a long-range dependent model to be most adequate to represent temperature anomalies. On the basis of this model, we performed a conservative test for a

polynomial trend which accounts for correlation. The approach pursued is based on simulating records without a trend, using the model found for the temperature anomalies. Finding a trend in these artificial series, such as the one observed in the empirical record, turned out to be highly unlikely. This implies that the trend in the Prague record is statistically significant. More refined strategies for trend assessment under correlation involving stochastic models and non-parametric trend estimates are discussed by, e.g., Kallache (2007).

From the modelling perspective, long-range dependence and deterministic trends are competing descriptions of observed departures from the mean value. This is illustrated with an analysis of the northern hemisphere mean temperature record. The temperature anomalies could be described equally well by either a complex short-range dependent (ARMA$[4,0]$) or a potentially long-range dependent process (FARIMA$[0,d,0]$). After subtraction of a linear trend, the short-range dependent process turned out to be the more suitable model, according to the simulation-based model selection. In this context, it is necessary to discuss the adequacy of a linear trend assumption. Given the effects of multiple exogenous forcing factors and a complex non-linear feedback mechanism in the climate system, it is not obvious that a linear, or even low-order polynomial, trend is an adequate hypothesis. However, trend-form misspecification implies a large fraction of variance on large scales that is to be explained by the stochastic process chosen to model the residuals. This favours long-range dependent processes, because of their ability to represent this large-scale variability. This implies that, as soon as more adequate assumptions on the trend form are made, long-range dependent processes might not be required to model the residuals, and conventional short-range dependent models might suffice. This illustrates a quite general problem when discriminating a trend from natural variability: the less assumptions are made for either the trend form or the natural variability, the more difficult their discrimination is.

- **What are the consequences of considering or ignoring autocorrelations for flood risk assessment? How can the uncertainty of flood return level estimates in autocorrelated records be quantified?**

  We demonstrated the effect of autocorrelation on the flood return level uncertainty analysis with a run-off time series recorded at the Gauge Vilsbiburg in southern Germany (Chapter 7). Assuming an underlying short-range dependent process for the observed run-off anomalies, the confidence interval obtained for the 100-year flood return level is approximately consistent with the one obtained from the asymptotic theory developed for uncorrelated observations. If we assume instead an underlying long-range dependent process, the asymptotic confidence interval does not provide a suitable approximation. The upper 95% confidence limit of the 100-year return level increases by 20% relative to the return level estimate. We consider this as a dimension that cannot be disregarded in practical applications. This effect, in turn, has influence on flood risk assessment or flood risk management, because the estimation of return levels and their uncertainty plays an important role in hydrological engineering and decision making: it forms the basis for setting design values for hydraulic structures like bridges, dams and dykes.

  This accentuates the need for a means to reliably quantify the uncertainty in the case of dependence. With the bootstrap-based approach developed in Appendix C, we suggested a methodology to accomplish this task. A crucial prerequisite for this method to yield reliable estimates is the adequate specification of a model for

the autocorrelation function, especially the reliable detection of long-range dependence. With respect to return level estimation, falsely detecting short-range dependence leads to an underestimation of the uncertainty and, thus, to an unjustified belief in safety. Falsely detecting long-range dependence, however, leads to an overestimation of uncertainty bounds and thus to over-dimensioned design values. While this reduces vulnerability and is in line with the precautionary principle, it might entail a misallocation of public funds. Supporting stakeholders in water management authorities in making such decisions was a motivation to develop reliable strategies for model selection and the detection of long-range dependence. The Bavarian Water Management Authority, for example, asked for further cooperation and is interested in making available these concepts for flood risk assessment.

The latter two applications, trend detection in temperature and uncertainty analysis in flood return level estimation, accentuate the importance of having reliable methods at hand for the detection of long-range dependence. In the case of trend detection, falsely concluding long-range dependence implies an underestimation of a trend and possibly leads to a delay of measures needed to take in order to counteract the trend. Ignoring long-range dependence, although present, leads to an underestimation of confidence intervals and thus to an unjustified belief in safety, as it is the case for the return level uncertainty analysis. A reliable detection of long-range dependence is thus highly relevant in practical applications. Other examples are the assessment of the maximum strength of wind gusts, which is important for designing wind turbines, or the exploration of financial market volatility. With rephrasing the detection problem as a model selection task and suggesting refined methods for model comparison, this thesis contributes to the discussion on, and development of, methods for the detection of long-range dependence.

8.2 Outlook

Simulation-Based Model Selection

The way it was used throughout the thesis, the simulation-based approach uses realisation of a FARIMA-type process. These realisations can be obtained from a Gaussian white noise process using a linear filter. Instead of Gaussian white noise, one can use realisations generated by a bootstrap resampling procedure. The latter generates records by resampling with replacement from a set of residuals, which is obtained applying the inverse linear filter to the observations. This variation of the approach might be especially useful in cases where the residuals from inverse filtering do not closely follow a Gaussian distribution. In such cases, it might lead to a significant improvement of the model selection in such situations.

A further potential application of the simulation-based model selection strategy is on the discrimination of stochastic models with trend components. Consider a combined estimation of trend and correlation structure, with the explanatory variables of the trend component formulated either as a polynomial in time or in the form of covariates. Because different explanatory variables do not have to be nested, they cannot be discriminated by standard methods such as the likelihood-ratio test. The proposed simulation-based approach might improve these trend detection strategies.

The algorithms for parameter estimation and parts of the model selection has been made available[1] for the software environment `R` (R Development Core Team, 2004). This

[1] `http://www.pik-potsdam.de/~hrust/tools.html`

package is intended for publication at the corresponding comprehensive archive network[2]. This accessibility to the methods within a consistent and widely used software environment facilitates the use of these concepts for scientists from various disciplines, such as physics, geosciences and economy.

Uncertainty in Return Level Estimates

Quantifying the uncertainty of return level estimates obtained from correlated series with the approach presented here turns out to be a complex procedure as well as a numerically costly undertaking. This might hinder its use on a day-to-day basis by water management authorities. The computer time needed depends on the ensemble size, i.e. the accuracy required, and on the length of the empirical maxima series. For the given example, the time consumed to generate the bootstrap ensemble, once the model for the maxima series has been identified, is of the order of magnitude of several hours. It is desirable to improve the efficiency of this approach or find suitable methods for an approximation. This will allow water management authorities to benefit from this kind of uncertainty analysis and possibly reduce the impact of floods or optimise the allocation of public funds.

Furthermore, the validation study showed that the distribution of return level estimates is close to, but not identical with, the reference distribution. A further investigation of this discrepancy might be useful. A proximate starting point is the way members of the ensemble of simulated maxima series are generated.

Aside from the uncertainty due to dependence in the series, a further important challenge in extreme value statistics is the instationarity of distributions. General methods have been proposed for the block maxima approach and maximum likelihood parameter estimation. Because autocorrelation and trends can be competing descriptions, it would be useful to direct further work in extreme value statistics towards models accounting for instationarities as well as for autocorrelation. Such a comprehensive approach is certainly beneficial for scientific purposes and is also of considerable value to practitioners. Apart from this, such refined approaches to the analysis of extreme events might play an important role in the assessment of the fraction of risk attributable to anthropogenic climate change. A problem which might become important in the near future.

[2]The Comprehensive R Archive Network: `http://cran.r-project.org/`

Appendices

Appendix A

Review of Detrended Fluctuation Analysis

Detrended fluctuation analysis (DFA, Section 2.3.3) has been widely used to infer a LRD process from empirical time series and also to quantify the strength of the dependence, i.e. estimating the Hurst exponent H, or, equivalently, the fractional difference parameter $d = H - 0.5$ (e.g. Bunde and Havlin, 2002; Király and Jánosi, 2002; Bunde et al., 2004; Eichner et al., 2003). Little work has been done on characterising the inferential power of DFA and the properties of the estimator for the Hurst exponent, e.g., quantifying bias and variance. In this chapter, we present a simulation study which provides estimates of bias and variance for power-law noise processes. Furthermore, we critically review the limits of the method to infer LRD, suggest refinements and point to typical pitfalls. Finally, we exemplify the discussion with an analysis of the Prague daily temperature anomalies.

A.1 Bias and Variance for Self-Similar Increment Processes

Analogously to the work of Bardet et al. (2003) and Weron (2002), we use Monte Carlo (MC) simulations to study the bias and variance of the estimator $\widehat{H}_{\mathrm{DFA}n}$ for DFA1 and DFA2 ($n = 1, 2$) introduced in Section 2.3.3. In addition to a white noise process, we consider a process with a spectral density $S(\omega) = |\omega|^{-\beta}$, also referred to as power-law noise. The spectral density of this process is similar to the one of fGn or FD processes; especially the asymptotic behaviour for low frequencies ω is the same. Realisations of power-law noise with a prescribed exponent $\beta = 2H - 1$ can be obtained using the inverse Fourier transform (Section 2.4; Timmer and König, 1995).

Bias

We estimate the bias and variance from MC ensembles of realisations from power-law noise with different length N and prescribed Hurst exponents $0 < H < 1$. The estimate $\widehat{H}_{\mathrm{DFA}n}$ is obtained as the slope of a straight line fit to $\log F(s)$ versus $\log s$ for $s > 10$. We calculate the bias using

$$\mathrm{bias}(\widehat{H}_{\mathrm{DFA}n}) = \overline{\widehat{H}_{\mathrm{DFA}n} - H}, \tag{A.1}$$

with the bar denoting the ensemble mean. For $\widehat{H}_{\mathrm{DFA1}}$ and $\widehat{H}_{\mathrm{DFA2}}$ and an ensemble size of 2 500 runs the bias is shown in Figure A.1. The estimator based on DFA1 exhibits a positive bias for power-law noise with $0 < H < 0.5$, increasing with decreasing H. The bias reduces with an increasing record length. This is in line with an analytical result

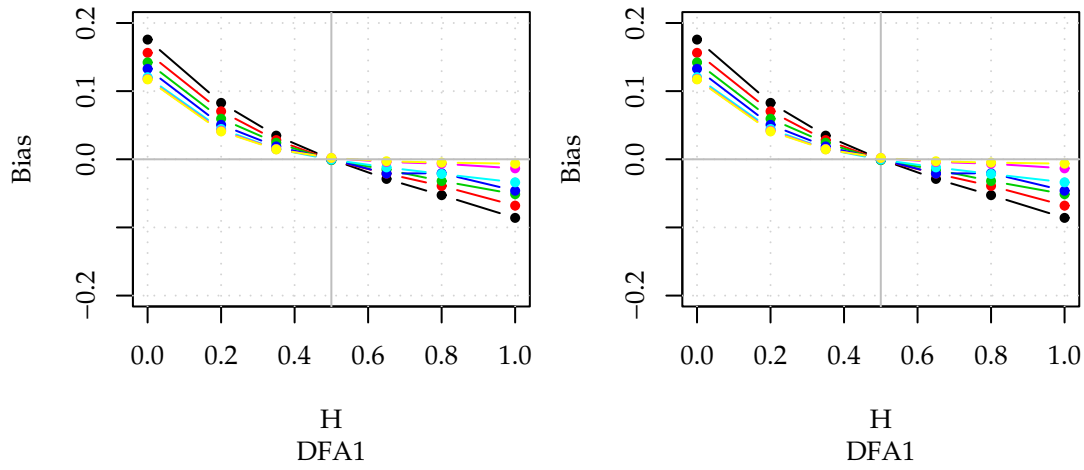

Figure A.1: Bias of $\widehat{H}_{\mathrm{DFA1}}$ (left) and $\widehat{H}_{\mathrm{DFA2}}$ (right). The bias is estimated from a MC ensemble of power-law noise realisations with a prescribed Hurst exponent H and various sample lengths N.

derived by Taqqu et al. (1995). In contrast, we find a negative bias for $0.5 < H < 1$, increasing in magnitude with H. An increasing record length reduces the bias faster than for $0 < H < 0.5$. For the white noise process ($H = 0.5$) no bias is observed for all length of the records. This "neutral point" moves towards $H \approx 0.6$ for the DFA2 based estimator. This is accompanied with an overall increase of the bias.

Variance

The variance of $\widehat{H}_{\mathrm{DFA}n}$ estimated from the same ensemble is shown in Figure A.2. A striking difference to the previous plot is the asymmetry. The variance $\mathrm{var}(\widehat{H}_{\mathrm{DFA1}})$ increases monotonically with $0 < H < 1$. This implies that for a given length N the Hurst exponent can be estimated more precisely for $0 < H < 0.5$ then for a white noise process ($H = 0.5$). Furthermore, Figure A.2 depicts a decreases in variance with the length N of the series.

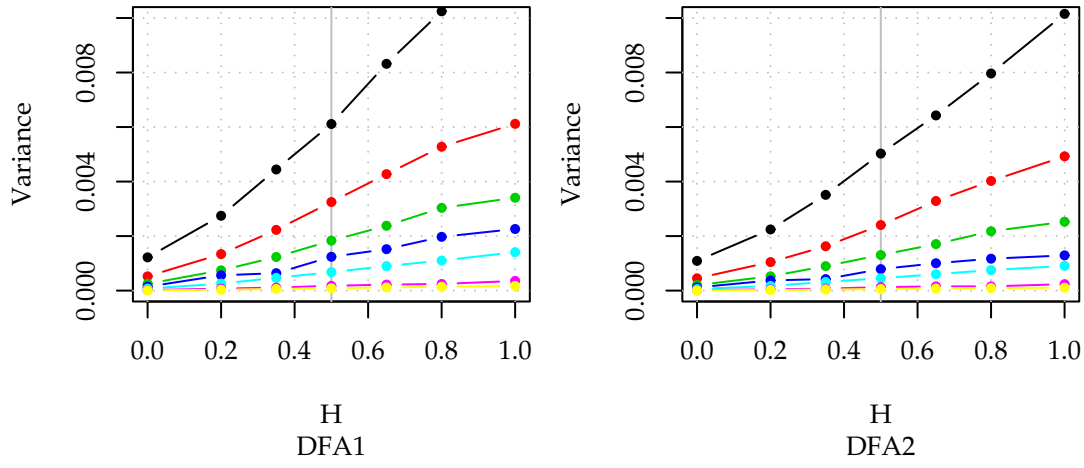

Figure A.2: Variance of $\widehat{H}_{\mathrm{DFA1}}$ (left) and $\widehat{H}_{\mathrm{DFA2}}$ (right). The variance is estimated from a MC ensemble of power-law noise realisations with a prescribed Hurst exponent H. The colour coding of the record length is analogue to Fig. A.1.

Figures A.1 and A.2 allow for a quick and rough estimate of bias and variance for increments of self-similar processes with $0 < H < 1$. For a more precise DFA-bases Hurst-exponent estimation from empirical data, one would have to adjust this simulation studies to the situation at hand, i.e. adjust the record length and the fit interval for the straight line and find a suitable parametric model to represent the data. This leads to a parametric bootstrap approach suitable to estimate confidence intervals. A parametric ansatz using one of the models described in Section 2.2.2 is suitable to accomplish this task. A different approach was suggested by Fraedrich and Blender (2003), they split the series under consideration in m records and performed the analysis m times to obtain an estimate for the variability. This, however, is only feasible if very long records, e.g., from simulation runs, are available.

A far more serious shortcoming of standard DFA analyses, as it is commonly used, is the *a priori* assumption of scaling. Straight lines are frequently fit to $\log F(s)$ versus $\log s$ without evaluating any criteria supporting this simple model (cf. Appendix A.2 and Maraun, Rust, and Timmer, 2004).

A.2 DFA and the Detection of Long-Range Dependence

Detrended fluctuation analysis including a subsequent straight line fit to the logarithmic fluctuation function $\log F(s)$ with $\log s$ as regressor yields an asymptotically unbiased estimator for the Hurst exponent H and thus for the fractional difference parameter d. This has been shown by Taqqu et al. (1995) for processes with the same asymptotic behaviour as fractional Gaussian noise or FARIMA processes. The limiting distribution of this estimator has not been studied so far. This makes statistical inference for the Hurst exponent impossible (cf. Section 2.3.3).

We first investigate the non-asymptotic behaviour of this estimator for power-law noise. Subsequently, we discuss the need to infer that the asymptotic behaviour has been reached for the series under consideration before Hurst exponent estimation is meaningful.

A.2.1 Inference of Scaling

Prior to the estimation of a scaling exponent, the existence of a scaling region has to be inferred from an empirical record (e.g., temperature anomalies, run-off, financial time series, etc.). Besides in Maraun, Rust, and Timmer (2004), it has not been studied, if DFA can be used to infer such a scaling region. It is thus not clear if LRD can be inferred from realisations of a process when it is not *a priori* clear, that this process exhibits scaling. It is therefore still an open question how *sensitive* and *specific*[1] DFA behaves when investigating processes of unknown correlation structure.

A necessary condition for the existence of LRD is the scaling of the fluctuation function $F(s)$ in the asymptotic region according to (2.21). Thus, in the remainder of this section, we address the following questions:

1. How to conclude scaling from the DFA fluctuation function?

[1] A procedure, that detects compatibility with LRD with a high probability, whensoever present, is called sensitive. An algorithm that with a high probability rejects LRD, when not present, is said to be specific. The optimal algorithm would be sensitive and specific. A sensitive but unspecific algorithm, however, would produce many false positive results, i.e. one would frequently detect LRD. This algorithm would not be suitable for a reliable inference. On the other hand, an un-sensitive but specific algorithm would be very conservative and would often reject the existence of LRD.

2. Does a region of scaling necessarily imply LRD?

Two Example Processes

To illustrate the line of argument, we consider a SRD as well as a LRD process and apply DFA to both. As an example for a LRD process we simulate power-law noise according to the method given in Timmer and König (1995) with a Hurst exponent of $H = 0.6$. This process shows power-law scaling in the ACF for a wide range of scales. For the SRD process we choose a superposition of three AR[1]-processes (cf. Section 2.2.1),

$$x(i) = \sum_{j=1}^{3} A_j y_j(i)\,, \quad y_j(i) = a_j y_j(i-1) + \eta_j(i)\,, \tag{A.2}$$

with $\eta_j(i)$ being Gaussian white noise of zero mean and variance $1 - a_j^2$. The latter ensures $\mathrm{var}(y_j) = 1$. We choose $A_1 = 0.913$, $A_2 = 0.396$, $A_3 = 0.098$, $a_1 = 0.717$, $a_2 = 0.953$ and $a_3 = 0.998$. Using $a_j = e^{-1/\tau_j}$ we find the following characteristic time scales for the individual AR[1] processes: $\tau_1 = 3$ days, $\tau_2 = 21$ days and $\tau_3 \approx 1.5$ years. The choice of the model parameters is motivated by the example of the Prague temperature record studied in Section A.3.1 and will become clear during the discussion.

Establish Scaling

Figure A.3 shows the fluctuation functions for a realisation of each of the two example processes with length $N = 70\,492$ corresponding to the Prague temperature record (cf. Section A.3.1). For every order of magnitude, 50 values of equal distance in logarithmic scale are calculated.

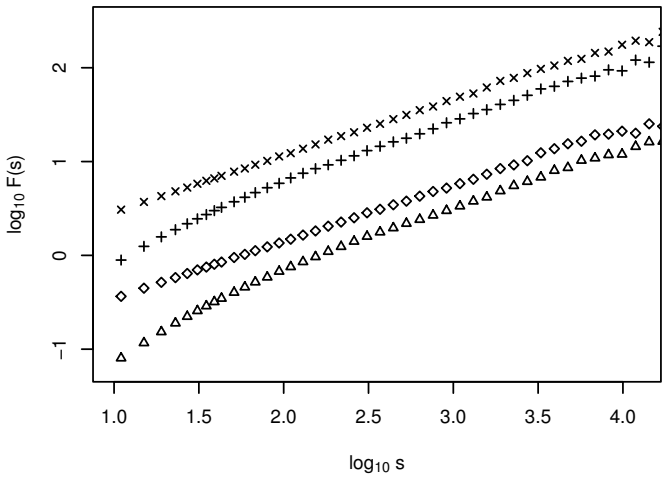

Figure A.3: Fluctuation functions calculated for an artificial LRD process with exponent $H = 0.6$ ($\times$ DFA1, $\diamond$ DFA2) and a superposition of three AR-processes ($+$ DFA1, $\triangle$ DFA2) as defined in Section A.2.1. For every order of magnitude, approx. 50 values are calculated. For clarity, only every third value is plotted.

To reliably infer power-law scaling of the fluctuation function, a straight line in the log-log plot has to be established. Since a straight line is tantamount to a constant slope, local estimates of the slope $\widehat{H}_{\mathrm{DFA}n}$ of $\log F(s)$ versus $\log s$ have to be evaluated for constancy in a sufficient range (e.g., Kantz and Schreiber, 1995; Tsonis and Elsner, 1995;

Timmer et al., 2000). The extend of a sufficient range is still a matter of debate (e.g., Avnir et al. (1998) and references therein). This concept has been introduced in the context of estimating correlation dimensions (Caputo et al., 1986; Tsonis and Elsner, 1995) and, in a different setting, has also been suggested for DFA (Peng et al., 1993).

Calculating Local Slopes

For a finite amount of data the estimation of the local slopes brings along a certain variability. Even for a process like the power-law noise, the local slopes of the empirical fluctuation function show variations around a constant H. This has two consequences for the calculating and interpreting the local slopes:

1. estimating the local slopes by finite differences results in a large variability and
2. this variability has to be quantified to allow for statistical inference.

Regarding the first point, the variability can be reduced fitting a straight line to $\log F(s)$ versus $\log s$ within a small window. The window is then shifted successively over all calculated scales s. Figure A.4 shows the local slopes of a realisation of the SRD model for different window sizes. Choosing the optimal window size, one has to trade bias for variance: for small windows, the bias is small, but the variability renders the interpretation difficult, whereas for large windows, the variance is reduced at the cost of a biased estimate of H. Thus, the extreme case of a single straight line fit to the whole range of scales considered is maximally biased. Since only one value of for the estimate $\widehat{H}_{\mathrm{DFA}n}$ is calculated, this does not allow to evaluate constancy.

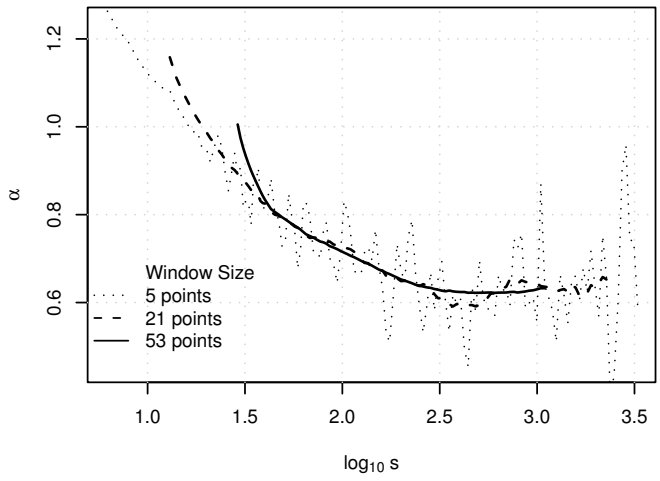

Figure A.4: Local slopes α (or H) of a realisation of the short-memory model for different window sizes. 50 points per order of magnitude are calculated, but for clarity reasons only every second value is plotted for the smallest window size. For small windows, the bias is very low, but the variability renders the interpretation difficult, whereas for large windows, the variance is reduced at the cost of a biased estimator $\widehat{H}_{\mathrm{DFA}1}$.

The second problem of quantifying the variability for a given length of the record is not straightforward. Since vicinal local slopes are not independent, confidence regions cannot be estimated straightforwardly from the procedure described in Section A.2.1 (Denker and Keller, 1986). Instead, we perform Monte Carlo simulations: for the two

example processes, we simulate 1 000 realisations to estimate mean and standard deviation for the estimator $\widehat{H}_{\mathrm{DFA1}}$ for the scales considered. The distribution of the estimates is approximately Gaussian. Thus, we employ the interval $[\overline{\widehat{H}_{\mathrm{DFA1}}} - 1.96\widehat{\sigma}, \overline{\widehat{H}_{\mathrm{DFA1}}} + 1.96\widehat{\sigma}]$ as estimates of the 95% confidence region, with the bar denoting again the ensemble mean. $\widehat{\sigma}^2$ is the variance of $\widehat{H}_{\mathrm{DFA1}}$ estimated from the ensemble.

Inferring Scaling for the Example Records

Figure A.5 displays the local slopes of the DFA1 (a) and DFA2 (b) fluctuation functions, estimated from one realisation of each of the example models. The realisation of the

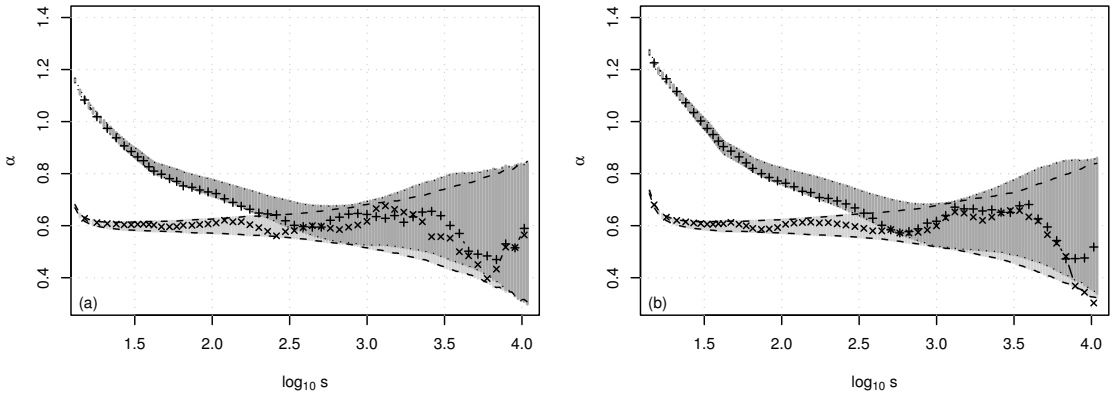

Figure A.5: Local slopes α (or H) of the DFA1 (a) and DFA2 (b) fluctuation function calculated for an artificial LRD process with exponent $H = 0.6$ ($\times$) and a superposition of three AR-processes ($+$) as defined in Section A.2.1 (window size 21 points). The dashed and the dotted lines border the shadowed $1.96\widehat{\sigma}$ intervals obtained from 1 000 realisations of the two processes, respectively.

LRD process shows fluctuations around a constant H within the corresponding $1.96\widehat{\sigma}$ interval, increasing like $\widehat{\sigma} \propto \sqrt{s}$ (Peng et al., 1993). The local slope $\widehat{H}_{\mathrm{DFA1}}(s)$ of the SRD realisation, however, decreases constantly in the beginning and basically follows the local slope of the LRD realisation for scales larger than $\log s \approx 2.5$. Thus, for a certain choice of parameters, a SRD model can mimic scaling in a finite range. Due to the principle of variance superposition for DFA (Hu et al., 2001) a suitable superposition of three AR[1] processes produces this effect in the fluctuation function analogously to the same effect in the spectral domain described in Hausdorff and Peng (1996).

Investigating for LRD one studies primarily the behaviour on large scales s or small frequencies assuming that influences from low frequencies are negligible here and do not bias the estimation of the LRD parameter. In our example, the $1.96\widehat{\sigma}$-cones from the LRD and SRD process are virtually indistinguishable in this range. Thus, based on the given record length and only considering large s, one cannot distinguish the realisations of the two models by means of DFA. For longer time series, the cones would shrink and the region of overlapping would become smaller.

Inference of SRD

However, a general dilemma related to the inference of LRD emerges: For a finite time series, one will always find a SRD model to describe the data (Beran, 1994). Thus, considering the inference of LRD, DFA is sensitive, but not specific. An alternative ansatz is to investigate if the underlying process is SRD. This requires

1. to show compatibility with a SRD model and
2. to exclude possible LRD models.

The first condition is always fulfilled, since one will always find a SRD model to describe a finite data set. The second condition is not fulfilled for the given example, because the record length is not sufficient to detect the SRD character $H = 0.5$ for large s of the AR-model by means of DFA. For longer time series as shown in Figure A.6, when a plateau of $H = 0.5$ is identifiable, LRD can be excluded and the specificity of DFA to infer SRD increases.

A.2.2 Pitfalls

We discuss typical difficulties for DFA and similar heuristic approaches. Problems arise for example when investigating the results only in a double logarithmic plot, or, if the observed behaviour is not consistent with the asymptotic behaviour. Also the transfer from a scaling exponent of the fluctuation function to the ACF is only possible in the asymptotic limit.

The Double Logarithmic Plot

Investigating only the double logarithmic plot of the fluctuation function, one is tempted to rashly conclude for LRD. Due to properties of the logarithm, fluctuations are suppressed in a double logarithmic plot and the deviation from a straight line is not easily visible (Tsonis and Elsner, 1995). Also, restricting the analysis to a straight line in the log-log plot forces $F(s)$ in the *procrustean bed* of power-laws. It will always yield some value for the slope but the suitability of the linear description is not evaluated. For the inference of LRD, this procedure would be sensitive but not specific. This results in attributing LRD to all processes with an estimate $\widehat{H}_{\mathrm{DFA}n} > 0.5$ for the largest scales observed. Such a result would trivialise the concept of LRD and provide no insight into the process. Thus, to reliably infer a power-law, a straight line may not be assumed *a priori* but has to be established, as discussed in Section A.2.1. Even if scaling is present, it is difficult to determine the beginning and ending of the scaling region in the log-log plot. However, the estimate $\widehat{H}_{\mathrm{DFA}n}$ derived from a straight line fit strongly depends on the fit boundaries if the realisation does not stem from a scale free process.

Finite scaling of short-memory processes

According to Section 2.1.4, the autocorrelations of SRD processes decay exponentially for large s and are negligible on scales large compared to the decorrelation time

$$\begin{aligned} \tau_D &= 1 + 2\sum_{s=1}^{\infty} \rho(s) = 1 + 2\sum_{s=1}^{\infty} e^{-s/\tau} \\ &\approx 2\tau \text{ for } \tau \gg 1. \end{aligned} \tag{A.3}$$

Consequently, for scales large enough, the slope of the fluctuation function of such a process converges to $H = 0.5$. However, for a finite set of data one cannot be *a priori* sure that the series is long enough to observe this. For a record of the SRD model defined in Section A.2.1 (length 70 492 points) the local slopes of the fluctuation function for the largest observed scales are compatible with power-law scaling. A plateau about $H = 0.5$ is not observed (Figure A.5). Thus, one might be tempted to conclude an underlying LRD

process. However, analysing a much longer record (1 000 000 points) of the same model yields such a plateau about $H = 0.5$ for large s (Figure A.6). Therefore, for a process with unknown correlation structure it is misleading to use solely the estimate $\widehat{H}_{\mathrm{DFA}n} > 0.5$ as argument for LRD. It might very well be that the record is too short to observe the plateau about $H = 0.5$ characteristic for SRD. This stresses the need for confidence intervals for

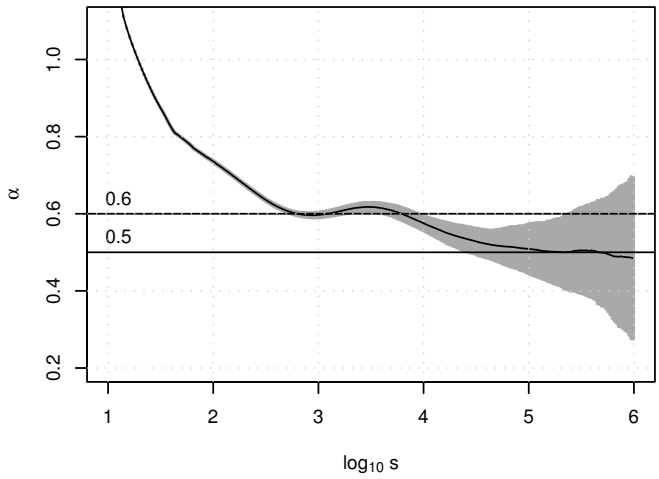

Figure A.6: Local slopes α (or H) from the empirical fluctuation function of the short-memory model (A.2), estimated from 200 realisations of length N=1 000 000 (solid line). A region of approximatively constant slope occurs between $\log s \approx 2.8$ ($s \approx 600$) and $\log \approx 3.8$ ($s \approx 6\,000$, corresponds to approximately 16 years). On larger scales, the slope reduces to $\widehat{H}_{\mathrm{DFA1}} = 0.5$, revealing the SRD nature of the model.

the estimate and for means to reliably discriminate between SRD and LRD processes.

DFA to ACF transfer of a finite "scaling" region

As shown in Sect. A.2.1, under certain conditions also SRD processes can mimic a finite "scaling" region. Thus, the question arises, if such a scaling region derived from the fluctuation function corresponds to a power-law like behaviour in the same region in the ACF. Such an assumption was used by, e.g., Koscielny-Bunde et al. (1996). To challenge this hypothesis, we compare the properties of the fluctuation function (Figure A.6) with the analytical ACF (Figure A.7). The dashed lines depict the ACF of the single AR[1] processes with the largest and the smallest time scale. The ACF of the superposition of the three AR[1] processes is given by the dashed-dotted line. The solid line represents a power-law with exponent $\gamma = 0.8$. This value can be expected when applying $H = 1 - \gamma/2$ (2.68) to the estimate $\widehat{H}_{\mathrm{DFA1}} \sim 0.6$ derived from the fluctuation function. The range of scales where the power-law-like behaviour holds for the fluctuation function is $\log s \gtrsim 2.5$ (Figure A.5). We find that the region of almost constant slope of the ACF is located on smaller scales between $s \approx 1$ and maximally $s \approx 1\,000$ ($\simeq$ 3 years). Thus, based on a finite scaling region found in the fluctuation function of a SRD process, it is in general not valid to conclude that an equal scaling region exists also for the ACF.

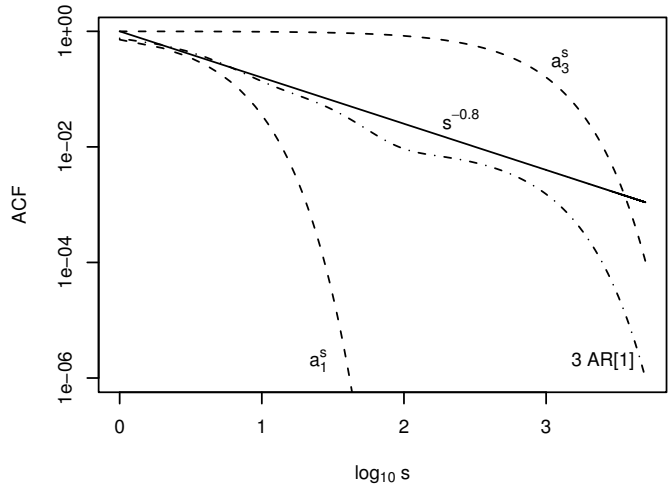

Figure A.7: Analytical ACF of the SRD model (A.2) (dashed dotted line). The dashed lines depict the ACF of the single AR[1] processes with shortest and longest time constant. In a range of scales from $\log s \approx 0$ ($s \approx 1$) to maximally $\log s \approx 3$ ($s \approx 1\,000$) the ACF approximately follows the power law with $\gamma = 0.8$ (solid line). For larger scales, it turns into an exponential decay determined by the AR[1] process with the largest time constant $\tau \approx 1.5$years.

A.3 Investigating the Prague Temperature Anomalies

In the following, we analyse the temperature anomalies from Prague and compare DFA and the full parametric approach applied to the same series in Section 5.1. To challenge both strategies, we furthermore impose the task of discriminating the Prague record from an artificial SRD time series. The latter is a realisation of a superposition of three AR[1] processes. The superposition was constructed such that its realisations mimic the Prague temperature residuals as closely as possible. The structure of this process is given in (A.2), the corresponding parameters in Section A.2.1. By definition this process is SRD.

We start with a DFA for the Prague normalised temperature anomalies and for a realisation of the AR-superposition. This is followed by the full-parametric modelling approach applied only to the AR-superposition. The most suitable parametric model for the temperature series itself has been determined already in Section 5.1.

A.3.1 Detecting Long-Range Dependence using DFA

DFA is reported to eliminate the influences of a possible trend in the mean on the analysis (e.g., Kantelhardt et al., 2001), we thus use the normalised temperature anomalies and study DFA1 and DFA2. An investigation of higher orders of DFA does not significantly affect the discussion presented while for DFA1 the effect of a trend might be suspected. The fluctuation function $F(s)$ (2.67) is calculated for approximately 50 points per order of magnitude upto $s_{\max} = N/4$ and is shown in Figure A.8 in double logarithmic representation. The behaviour is qualitatively different from white noise. To depict the variability of this estimate of the fluctuation function, we performed simulations with the AR model (A.2) and estimate its fluctuation function from 1 000 realisations with the same length as the observed record. We approximate a 95% confidence region by plotting $1.96\widehat{\sigma}$ intervals as grey shadows. As expected, the variability of the fluctuation function estimate increases wit the time scale s. Because the observed fluctuation functions for DFA1 and

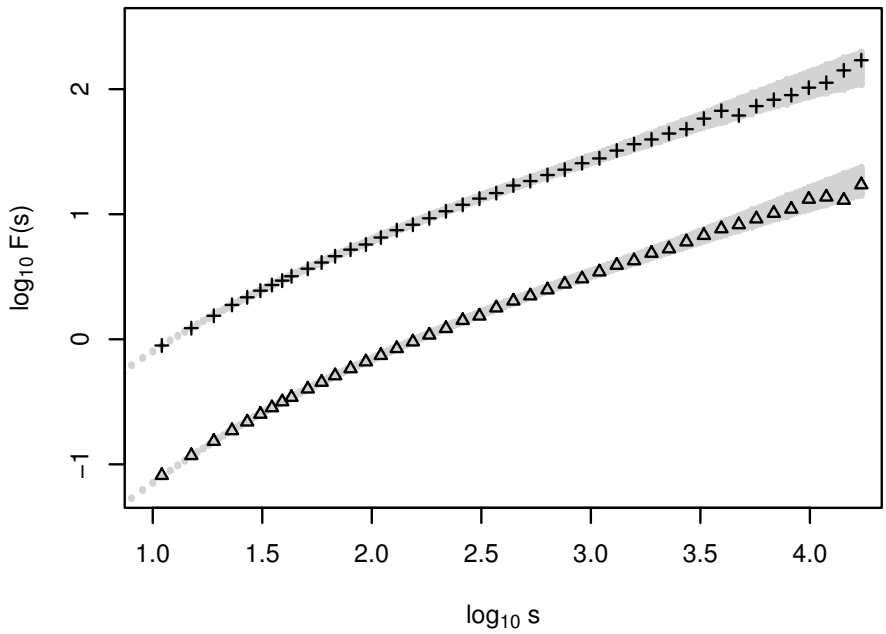

Figure A.8: DFA1 (+) and DFA2 (△) fluctuation function of Prague daily temperature data calculated for approximately 50 points per order of magnitude. Only every 4th point is shown to enhanced clarity. The shadows mark $1.96\widehat{\sigma}$ confidence regions derived from 1 000 runs of the AR model.

DFA2 are both well inside the 95% confidence interval, we consider the Prague temperature anomalies compatible with the AR model from the viewpoint of DFA.

Following the discussion in Section A.2.1, we estimate the local slopes to investigate for power-law scaling. From the fluctuation function of the Prague record, we estimate the local slopes using a straight line fit in a small window of 21 points. Figure A.9 shows the result and compares it to 1 000 runs from the AR model and from power-law noise with exponent $H = 0.6$. Accounting for the variability, we find the observed slopes consistent with a constant slope as expected from the power-law noise process (light grey) for $\log s \gtrsim 2.5$. In the same range of scales, however, the observed slopes also agree with the AR model. Since we are looking for a scaling region, we consider at the most $\log s \gtrsim 2.5$. Thus, from the given data, one cannot decide whether the Prague temperature time series is a realization of a SRD (dark grey) or a LRD process (light grey). This plot shows furthermore, that considering the scales $\log s \lesssim 2.5$, the power-law noise is not a sufficient model for the observed series.

A.3.2 Detecting Long-Range Dependence using Parametric Models

In Section 5.1.1 we found the FARIMA$[2, d, 2]$ being the most suitable model for the Prague maximum temperature residuals with parameters given in Table 5.2. The fractional difference parameter was estimated to be $\widehat{d} = 0.109(0.017)$ for this model. Employing the Gaussian limiting distribution, we find a 95% confidence interval for the difference parameter of $[0.076, 0.142]$ and thus $d = 0$ for SRD processes is not included. We thus infer the underlying process to be LRD.

It remains to investigate whether we can correctly identify the realisation of the AR-superposition as coming from an SRD process. This is pursued in the following.

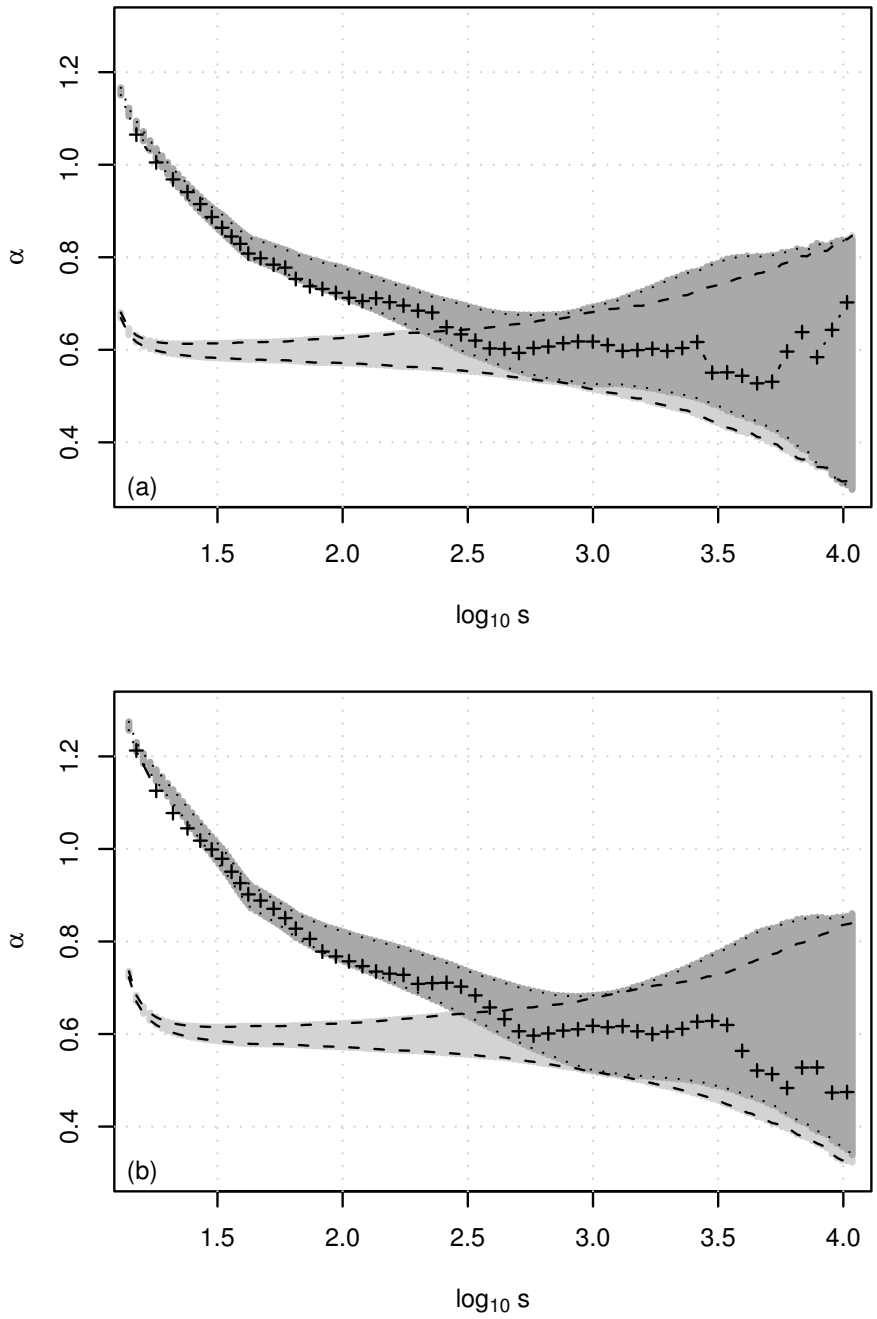

Figure A.9: Local slopes α (or H) of the fluctuation functions plotted in Figure A.8 for DFA1 (a) and DFA2 (b) of the Prague daily temperature data. The dotted lines border the $1.96\widehat{\sigma}$ confidence regions of the short-range correlated model (A.2) (dark shadow), the dashed lines those of the long-memory model with $H = 0.6$ (light shadow).

Stochastic Modelling of the AR[1] Superposition

The AR model (A.2) is constructed such that realisations generated from it mimic the characteristics of the Prague temperature anomalies. We thus do not have to account for deterministic components, such as the seasonal cycle or a trend. Starting with the set of FARIMA$[p,d,q]$ and ARMA$[p,q]$ models compiled in Section 5.1.1 we reduce these with the model selection strategies analogously to the procedure for the temperature anomalies (Section 5.1): the goodness-of-fit test, the HIC-based model selection and, finally, the likelihood-ratio-based model selection.

Goodness-of-fit On the basis of the goodness-of-fit test (3.4) applied to the before mentioned set of models we reject only the FD (FARIMA$[0,d,0]$) model on a 5%-level of significance.

HIC Model Selection Figure A.10 depicts the HIC values for the models considered. It is striking that for the most cases the ARMA$[p,q]$ model has a smaller HIC than the

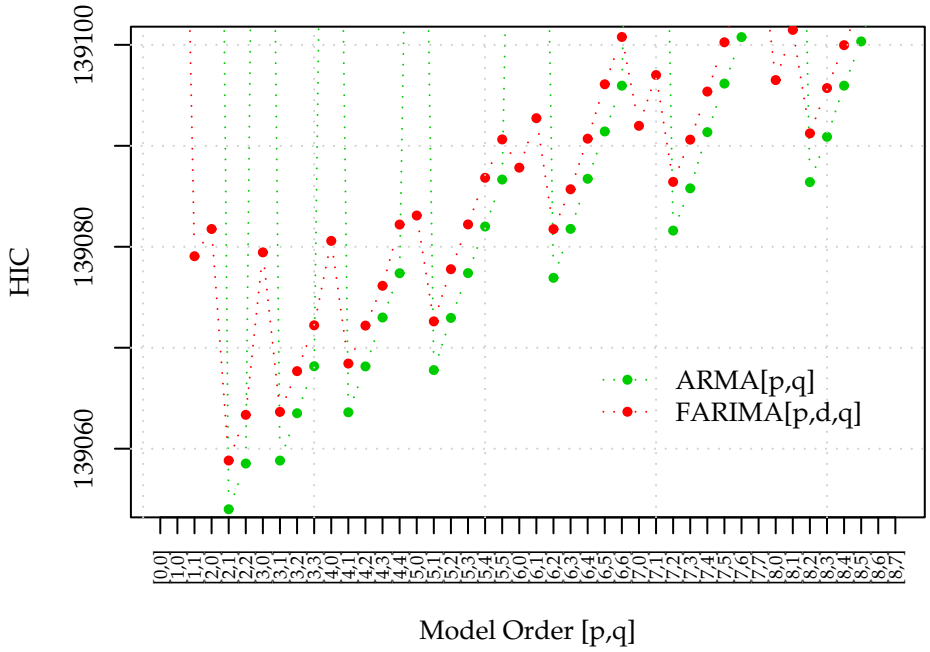

Figure A.10: HIC for various ARMA$[p,q]$ (green) and FARIMA$[p,d,q]$ (red) models fitted to realisation of the AR model (A.2) The model orders $[p,q]$ are plotted along the abscissa.

corresponding FARIMA$[p,d,q]$ model. Investigating the fractional difference parameter for the latter models (Table A.1) reveals that besides FARIMA$[1,d,0]$, FARIMA$[1,d,1]$ and FARIMA$[2,d,0]$ all estimates are compatible with the $d = 0$ within one (or, in one case, 1.96) standard deviation. We retain these models, as well as ARMA$[p,q]$ models with orders $(p,q) \in \{(2,1),(2,2),(3,1),(3,2)\}$. Keeping other FARIMA$[p,d,q]$ models is not meaningful, because their fractional difference parameter is compatible with zero and thus they are equivalent to the corresponding ARMA$[p,q]$ processes.

Likelihood-Ratio Test With a likelihood-ratio test, we investigate whether the three models with a non-trivial fractional difference parameter are admissible simplifications of

Model	$\widehat{d}(\sigma_{\widehat{d}})$
$[1,d,0]$	$0.148(0.010)$
$[1,d,1]$	$0.158(0.011)$
$[2,d,0]$	$0.160(0.011)$
$[2,d,1]$	$0.000(0.015)$
$[2,d,2]$	$0.000(0.037)$
$[3,d,1]$	$0.000(0.037)$
$[3,d,2]$	$0.012(0.017)$
$[4,d,1]$	$0.000(0.040)$
$[3,d,3]$	$0.028(0.038)$
$[4,d,2]$	$0.027(0.038)$
$[5,d,1]$	$0.000(0.042)$
$[4,d,3]$	$0.023(0.022)$
$[5,d,2]$	$0.000(0.041)$

Table A.1: Estimates and asymptotic standard deviation of the fractional difference parameter obtained for the FARIMA$[p,d,q]$ models with smallest HIC.

ARMA$[p,q]$

	Model f	Model g	p-val
1	$[3,2]$	$[2,2]$	1
2	$[3,2]$	$[3,1]$	0.713
3	$[2,2]$	$[2,1]$	0.584
4	$[3,1]$	$[2,1]$	0.939

FARIMA$[p,d,q]$

	Model f	Model g	p-val
1	$[2,d,1]$	$[2,d,0]$	<0.001
2	$[2,d,1]$	$[1,d,1]$	<0.001
3	$[2,d,1]$	$[1,d,0]$	<0.001

Table A.2: p-values for a likelihood-ratio test of model g being an admissible simplification of model f for ARMA$[p,q]$ and FARIMA$[p,d,q]$ models of the realisation stemming from the AR[1]-superposition (A.2).

the FARIMA$[2,d,1]$, the simplest models with trivial difference parameter. The p-values in Table A.2 (right) reveal that we can reject this hypothesis for all three cases on a 5%-level (or even 1%-level) of significance. This implies, that there is no evidence for a LRD process underlying this data series.

Among the ARMA$[p,q]$ models, we find the ARMA$[2,2]$ and ARMA$[3,1]$ as admissible simplifications of the ARMA$[3,2]$ (Table A.2, left, lines 1 and 2). Both models can be further simplified by ARMA$[2,1]$ (lines 3 and 4). The latter is the most suitable model out of the canon we started with for the realisation of the AR-superposition (A.2).

Using the full-parametric modelling approach we are thus able to correctly identify the realisation of a superposition of three AR[1] processes as coming from an SRD process. This was not possible by means of DFA. The original model, a superposition of AR[1] processes (A.2), was not included in the model canon we used to describe the example series with. However, such a superposition can be well approximated with ARMA$[p,q]$ processes.

A.4 Summary

In a simulation study, we investigated bias and variance for a DFA-based estimator for the Hurst exponent using realisations of power-law noise. These properties of the estimator

are not only influenced by the time series length but also by the Hurst exponent of the underlying process. Without knowledge about bias and variance, inference about the Hurst exponent cannot be made.

We further considered the inference of LRD by means of DFA with respect to the notions of sensitivity and specificity. The inference of a LRD process underlying a given time series requires not only to show compatibility of the data with a LRD process in a certain range of scales. Furthermore, other possible correlation structures, especially SRD, have to be excluded.

Power-law like behaviour in some range of scales of the DFA fluctuation function alone is frequently taken as evidence for LRD. To reliably infer power-law scaling, it must not be assumed but has to be established. This can be done by estimating local slopes and investigating them for constancy in a sufficient range. However, finite data sets bring along natural variability. To decide, if a fluctuating estimation of the slope has to be considered as being constant, we calculated empirical confidence intervals for a LRD and a simple SRD model.

Discussing typical difficulties of interpreting DFA results, we note that scaling cannot be concluded from a straight line fit to the fluctuation function in a log-log representation. Additionally, we show that a local slope estimate $\widehat{H}_{\mathrm{DFA}n} > 0.5$ for large scales does not necessarily imply long-memory. If the length of the time series is not sufficiently large compared to the time scales involved, also for SRD processes $\widehat{H}_{\mathrm{DFA}n} = 0.5$ may not be reached. Finally, we demonstrated, that it is not valid to conclude from a finite scaling region of the fluctuation function to an equivalent scaling region of the ACF.

With the Prague temperature anomalies and a corresponding artificial series from a SRD process, we exemplify the problems and pitfalls discussed. Using DFA we could not discriminate the Prague record from the artificial series. Furthermore, by means of a log-log plot, we classify the underlying process of both series as LRD, leading to a false positive result for the SRD process. A reliable identification was possible with the full-parametric ansatz (Section 4.4).

Because it is always possible to find and SRD process to describe a finite set of data, some criterion is needed to evaluate the performance of the description taking the complexity of the model into account. We thus approach the problem of distinguishing SRD and LRD processes on the basis of a realisation of a parametric modelling point of view developed in Chapter 4.

Appendix B

Long-Range Dependence – Effects, Methods, Mechanisms

A simple example to illustrate the effect of long-range dependence on the estimation of statistical quantities is illustrated in the first section. This is followed by some physical explanations of the phenomenon. Further, we describe the log-periodogram regression and the rescaled-range statistic as methods to quantify LRD. This is followed by introducing two LRD specific statistics which can be used as further criteria for the simulation-based model selection described in Section 3.3.2. These statistics are based on the log-periodogram regression and on DFA. Subsequently, we describe the generation of an artificial series, which is used as a challenging example to illustrate the LRD detection strategies described in Chapter 4.

B.1 Effects of Long-Range Dependence

We illustrate the effect of LRD on statistical quantities estimated from a realisation of a stochastic process with a simple example: consider the arithmetic mean $\bar{X} = N^{-1} \sum_{t=1}^{N} X_t$ as an estimator for the the expectation value μ, i.e. $\widehat{\mu} = \bar{X}$. A basic result in statistics is the decrease in variance of the arithmetic mean with an increasing sample size N. For random variables X_t, identically distributed with expectation value μ and variance σ^2, the variance of the arithmetic mean $\bar{X}$ can be calculated as (Beran, 1994)

$$\mathrm{var}(\bar{X}) = \frac{\sigma^2}{N^2} \sum_{s,t=1}^{N} \mathrm{cor}(X_s, X_t). \tag{B.1}$$

For independent X_t with $\mathrm{cor}(X_s, X_t) = 0$ for $s \neq t$ this reduces to the well known result $\mathrm{var}(\bar{X}) = \sigma^2/N$. In case of a dependent process X_t, the calculation of $\mathrm{var}(\bar{X})$ requires to consider the ACF $\rho(\tau)$. For a weakly stationary process, we find

$$\mathrm{var}(\bar{X}) = \frac{\sigma^2}{N} \left[1 + 2 \sum_{\tau=1}^{N-1} \left(1 - \frac{\tau}{N} \right) \rho(\tau) \right]. \tag{B.2}$$

In case of a more specific description of the stationary process equation (B.2) can be further simplified. It is instructive to consider two examples: an autoregressive process (AR[1], cf. Section 2.2.1), as a prominent representative of a SRD processes, and a LRD process, specified only by the limiting behaviour (2.21).

Example 1: Short-Range Dependence

With the autocorrelation function for an AR[1] process $\rho(\tau) = a^\tau$ with $|a| < 1$ as described in Section 2.2.1, equation (B.2) reads

$$\mathrm{var}(\bar{X}) = \frac{\sigma^2}{N}\left[1 + 2\sum_{\tau=1}^{N-1}\left(1 - \frac{k}{N}\right)a^\tau\right], \tag{B.3}$$

which in the limit of large N reduces to

$$\mathrm{var}(\bar{X}) \approx \frac{\sigma^2}{N}\left[1 + \frac{2a}{1-a}\right] = \frac{\sigma^2}{N}c_a, \tag{B.4}$$

with $c_a = 1 + \frac{2a}{1-a}$ being a constant depending on the AR coefficient. In this case the variance of the mean decreases at the same rate as for uncorrelated data, namely $\mathrm{var}(\bar{X}) \propto N^{-1}$, but with a different constant.

Example 2: Long-Range Dependence

For a process with an algebraically decaying ACF, $\rho(\tau) \propto \tau^{-\gamma}$ with $\gamma \in (0,1)$, for large lags the sum in (2.20) does not converge to a constant value. It behaves instead as

$$\sum_{\tau=-(N-1)}^{N-1} \rho(\tau) \propto N^{1-\gamma}, \tag{B.5}$$

which in turn leads to

$$\mathrm{var}(\bar{X}) \propto \frac{\sigma^2}{N^\gamma}. \tag{B.6}$$

Because $\gamma < 1$, this algebraic decay of the variance of the mean is slower than the exponential decay for uncorrelated or short-range correlated processes (Beran, 1994).

These examples point towards the pivotal difference in records from LRD and SRD processes: given a sample of length N, for SRD processes the mean value can be estimated with the variance decreasing as $1/N$, while for a LRD process the variance of this estimator decreases only with $1/N^\gamma$ and is thus in general larger. Consequently, the uncertainty of statistical measures inferred from records of LRD processes is larger than in the case of SRD. Statistical measures which are affected are, e.g., regression coefficients (Sibbertsen, 1999), semi-parametric trend estimates (Kallache et al., 2005; Craigmile et al., 2004) or quantiles estimated from the frequency distribution (Koutsoyiannis, 2003). The latter effect is of tremendous importance in extreme value statistics, e.g., when setting design values for hydraulic constructions as dams, bridges or dykes (cf. Section 7; Rust et al., 2006). The effect on regression coefficients is relevant with respect to trend detection, a highly debated topic in the early years of climate change research: assuming a LRD process underlying temperature records allows small trends to "hide" in the low frequency variability and renders detection more difficult (cf. Sections 5.1.3, 5.2; Kallache, 2007). It is thus of considerable importance to reliably detect an underlying LRD process. Having discovered that a frequently used heuristic approach is prone to falsely detect LRD (Section A.2; Maraun, Rust, and Timmer, 2004), the development of a different approach to the detection of LRD based on parametric modelling and model selection is suggested in Chapter 4.

B.2 Physical Explanations of the Phenomenon

In the time after Hurst (1951), various explanations or descriptions of the Hurst phenomenon emerged. In the 1960s Mandelbrot suggested increments of self-similar processes to describe this effect (cf. Section 2.2.2). The aggregation of processes with different time scales was suggested by Hipel and McLeod (1978). Certain types of instationarities being held responsible for the phenomenon were discussed by Klemes (1974). The latter two approaches are discussed in brevity in the following. An extensive overview about the Hurst phenomenon from a hydrological perspective has been given by Koutsoyiannis (2002, 2003).

Aggregation of Processes with Different Characteristic Time Scales

Hipel and McLeod (1978) showed that it is possible to preserve the Hurst phenomenon (in particular the rescaled range statistic, Section B.3.1) for a finite time series with conventional linear stochastic models (cf. Section 2.2.1). Granger (1980) showed that a superposition of n first order autoregressive processes (AR[1], Section 2.2.1) lead to a long-range dependent process in the limit of large n. The parameters a and σ_η of these processes (2.23) have to be independent random variables following a beta distribution. Physically, the parameter a is related to the relaxation time of the process. This implies that a suitable superposition of relaxations with different characteristic time scales can evoke the Hurst phenomenon.

Instationarities

Instationarities in the empirical series, such as a trend in the mean, have been held responsible for the Hurst phenomenon (e.g., Klemes, 1974). The latter motivated the development of methods supposed to eliminate the influence of trends, like the detrended fluctuation analysis (DFA) described in Section 2.3.3.

Other Explanations

Further variants of models showing the Hurst phenomenon are proposed for example by Davidsen and Schuster (2002); Koutsoyiannis (2002) and Lillo et al. (2002). Especially interesting is the idea of Maheu (2005). He suggests that heteroscedasticity, i.e. a non-homogeneous variance, might also account for the phenomenon. Heteroscedasticity can also be found in run-off records (Elek and Márkus, 2004).

B.3 Further Heuristic and Semi-Parametric Methods to Quantify LRD

B.3.1 Rescaled Adjusted Range

Hurst (1951) studied the flow of the Nile using the *rescaled adjusted range statistic* (R/S). Consider the inflow X_t at time t, the cumulative inflow $Y_t = \sum_{i=1}^{t} X_i$ and the mean $\bar{X}_{t,s} = 1/s \sum_{i=t+1}^{t+s} X_t$. The rescaled adjusted range is then

$$R/S = \mathrm{E}\left[\frac{R(t,s)}{S(t,s)}\right]_t, \tag{B.7}$$

with

$$R(t,s) = \max_{0\le i\le s}[Y_{t+i} - Y_t - \frac{i}{s}(Y_{t+s} - Y_t)] - \min_{0\le i\le s}[Y_{t+i} - Y_t - \frac{i}{s}(Y_{t+s} - Y_t)],$$

and

$$S(t,s) = \sqrt{s^{-1}\sum_{i=t+1}^{t+s}(X_i - \bar{X}_{t,s})^2}.$$

For large s, a plot of $\log R/S$ versus $\log s$ is approximately scattered around a straight line with slope H (Beran, 1994).

The R/S statistic was refined by Mandelbrot and others (e.g., Mandelbrot and Wallis, 1968b; Mandelbrot, 1972, 1975). Later Mandelbrot (1975) derived the asymptotic behaviour for R/S for short-range and long-range dependent processes. The latter can be described following Beran (1994):

For X_t such that X_t^2 is ergodic and $t^{-\frac{1}{2}}\sum_{k=1}^{t} X_k$ converges weakly to *Brownian motion* as $t \to \infty$ then for $s \to \infty$ and $Q(t,s) = R(t,s)/S(t,s)$

$$s^{-\frac{1}{2}}Q(t,s) \xrightarrow{d} \xi, \tag{B.8}$$

with ξ being a nondegenerate random variable.

It remains open for which range of s the described asymptotic behaviour starts and what the confidence interval of the estimate is. For a finite sample the distribution of H is neither normal nor symmetric. All this makes inference about the value of H difficult.

In fact, in the presence of unanticipated high frequency behaviour, Lo (1991) noticed the invalidity of a test for $H = 0.5$, i.e. uncorrelated series. Based on the asymptotic theory for the R/S statistic by Mandelbrot (1975) this null hypothesis is rejected too often. Lo (1991) developed a corrected statistic accounting for a wide range of high frequency components.

The R/S statistic was not designed to discriminate between short-range and long-range dependent processes, it is rather a tool to measure the Hurst exponent H in situations where the region of asymptotic behaviour is known. This is a severe shortcoming encountered also in the description of further heuristic approach such as DFA (cf. Appendix A). In many fields where long-range dependence is to be taken into account, such as geophysics or econometrics, also high frequency components are present. They need to be accounted for in order to obtain reliable results.

B.3.2 Log-Periodogram Regression

In case a full-parametric description of the spectral density (Chapter 4) cannot be achieve or is not pursued. The fractional difference parameter can be estimated using a semi-parametric approach in the spectral domain. Referring to a definition of LRD (2.22), the spectral density $S(\omega)$ close to the origin behaves as

$$S(\omega) \propto c_f|\omega|^{-\beta},\ |\omega| \to 0. \tag{2.22}$$

The exponent β can be related to the fractional difference parameter d by $\beta = 2d$. Taking the logarithm on both sides, the equation motivates a regression approach to estimate β or, equivalently, the fractional difference parameter d. Consider

$$\log I(\omega_j) \approx \log c_f - \beta \log|\omega_j| + u_j, \tag{B.9}$$

with a positive constant c_f and $j = l, \ldots, m$.

The GPH Estimator

Geweke and Porter-Hudak (1983) (GPH) were the first to suggest the estimation of a parameter quantifying the long-range dependence (either β or d) using this kind of regression. However, instead of $\log \omega_j$ they considered $\log|1 - e^{i\omega_j}|$ as regressor motivated by the exact form of the spectral density for fractional ARIMA processes (2.50). Furthermore, they included frequencies ω_j in the regression starting with $j = 0$ to obtain the estimate $\widehat{d}_{\text{lp}} = \widehat{\beta}/2$. Geweke and Porter-Hudak argued for the following Gaussian limiting distribution of this estimator:

$$m^{1/2}(\widehat{d}_{\text{lp}} - d) \xrightarrow{d} N\left(0, \frac{\pi^2}{24}\right). \tag{B.10}$$

With this result, they provided a simple way to approximate confidence intervals and carry out hypothesis testing (cf. Section 3.1.1).

However, Robinson (1995a) showed that their arguments do not hold. Mainly because the residuals u_j cannot be considered as asymptotically independent or homoscedastic. Despite this, the desired limiting distribution (B.10) could be established by Robinson for Gaussian processes using $\log \omega_j$ as regressor as suggested in (B.9) and by trimming out the smallest frequency ω_j (Robinson, 1995a). This result was generalised to linear processes by Velasco (2000).

Semi-Parametric Whittle-Estimation

An even more efficient estimator was suggested by Künsch (1986) with the limiting distribution provided by Robinson (1995b). It is essentially based on Whittle estimation of the model $S(\omega) = C\omega^{-2d}$ for the first m frequencies:

$$\widehat{d}_K = \arg\min_d \left\{ \log\left[\frac{1}{m}\sum_{j=1}^{m} \frac{I(\omega_j)}{\omega_j^{-2d}}\right] - \frac{2d}{m}\sum_{j=1}^{m} \log \omega_j \right\}. \tag{B.11}$$

Under milder regularity conditions than those required for (B.9), the limiting distribution is given by

$$m^{1/2}(\widehat{d}_k - d) \xrightarrow{d} N\left(0, \frac{1}{4}\right), \tag{B.12}$$

having a smaller variance and leading thus to a more efficient estimator. Different from (B.9) no closed form for $\widehat{d}_k$ can be given. The estimate has to be obtained numerically in the same way as described for the full-parametric Whittle estimator in Section 2.3.2.

Bandwidth Choice and Related Approaches

For $\widehat{d}_K$, as well as for $\widehat{d}_{\text{lp}}$ a bandwidth m has to be chosen for the estimation. A brief review on various approaches on the choice of an optimal bandwidth is given in Robinson (2003). Henry and Robinson (1996) discuss an optimal choice of the bandwidth in the sense of a bias-variance trade-off.

The log-periodogram regression for the non-stationary domain $d > 0.5$ has been discussed by Hurvich and Ray (1995). A similar approach to estimate the long-range dependence parameter using wavelet analysis is described by Percival and Walden (2000) and Wornell (1993).

Combining Log-Periodogram Regression and ARMA$[p,q]$ Parameter Estimation

It is conceivable to estimate the (fractional) difference parameter d and the autoregressive moving average parameters a_i and b_j separately. In the early days of ARIMA$[p,d,q]$ processes (with $d \in \mathbb{N}$), the estimation for integer differences has been carried out by heuristic methods and for the following inference about the AR and MA parameters d was assumed to be known (Box and Jenkins, 1976). For fractional differences $d \in \mathbb{R}$, the first step can be carried out using, e.g., the log-periodogram regression; the inference about ARMA parameters, for example, by maximum likelihood. For the log-periodogram regression it is possible to specify a confidence interval for the estimate of d in contrast to the heuristic methods of estimating the integer difference paramters. However, the influence of estimating d on the estimation of the autoregressive and moving average parameters is not accounted for. In a combined approach, i.e. simultaneous estimation of the parameters d, a_i and b_j, this influence can be quantified via the covariance matrix $\mathbf{V}$ of the parameter vector $\boldsymbol{\theta}$ (Section 2.3.1).

B.4 Specific Test Statistics for Model Selection

In this section, we evaluate two statistics which can be used in the simulation-based model selection described in Section 3.3.2. Being especially interested in the low frequency behaviour, we investigate the performance of the log-periodogram regression (Section B.3.2) and the DFA (Section 2.3.3) as further model selection criteria. For both approaches one has to choose an appropriate bandwidth for the straight line fit. We use this free parameter to optimise the power (Section 3.3.3), i.e. the separation of the two distributions resulting from the simulations. For both criteria, we calculate p-values and power and try to identify an optimal bandwidth. It turned out to be also instructive to examine the dependence of the estimates on the bandwidth.

As an example, we consider the setting introduced in Section 4: a realisation of an ARMA$[3,2]$ process (Section 4.1; Appendix B.5) and models FAR$[1]$ and ARMA$[3,2]$. We estimate the model parameters from the realisation and obtain thus two fully specified models. These two models are used to generate two ensembles of realisations. For all ensemble members the estimates $\widehat{d}_{\mathrm{lp}}$ (or $\widehat{d}_{\mathrm{DFA}}$, Section 2.3.3) are calculated. The estimate from the observed series is then compared to the distribution obtained from the ensembles. The p-values and power can be obtained in the same way as described for the likelihood-ratio (cf. Section 3.3.2). For better readability, we denote the ARMA$[3,2]$ model as f and the corresponding hypothesis as H_f, analogously g represents the FAR$[1]$ model.

B.4.1 Log-Periodogram Regression

Bandwidth-Dependence of the Power

The power for a critical nominal level of $\alpha = 0.05$ as a function of the bandwidth of the linear fit is shown in Figure B.1 (left) for testing H_f against H_g (green) and vice versa (red). In both cases a maximum power is achieved for a bandwidths $0 < \omega < \omega_1 = 0.0021$: $\widehat{\mathrm{pow}}_f(g, 0.05) = 0.851$ and $\widehat{\mathrm{pow}}_g(f, 0.05) = 0.836$. Thus this test is not as powerful as the one based on the likelihood-ratio which exhibits a power $\widehat{\mathrm{pow}} > 0.9$ for both test situations (Section 4.4.1). The distribution functions for the ensembles of fractional difference parameter estimates for the two hypothesis are shown in Figure B.1 (right) for the bandwidth with maximal power. The estimate obtained for the "observed" series is

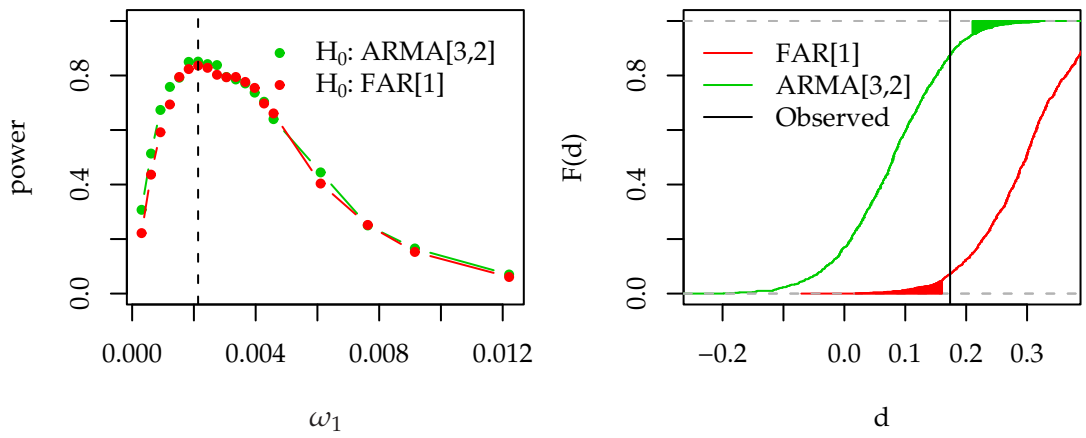

Figure B.1: Power of testing H_f against H_g (green) and vice versa (red) at level $\alpha = 0.05$ for various bandwidths $0 < \omega < \omega_1$ in the spectral domain (left). The lower bound ω_1 where maximal power is achieved is marked with a dashed vertical line. Distributions of the estimated fractional difference parameters obtained from the bootstrap ensembles using the log-periodogram regression and the bandwidths of maximal power (right).

depicted as a vertical bar and is located in the gap between the two critical regions. Thus, neither hypothesis can be rejected in this example.

Calculating p-Values

We further calculate p-values analogously to Section 3.3.2. For the bandwidth of maximal power, testing H_f yields a p-value of $\widehat{p}_f = 0.126$. Testing H_g, we find $\widehat{p}_g = 0.071$. Figure B.2 shows the p-values for different bandwidths together with the power for the two tests. The left plot refers to testing H_f, the right plot H_g. In regions of high power,

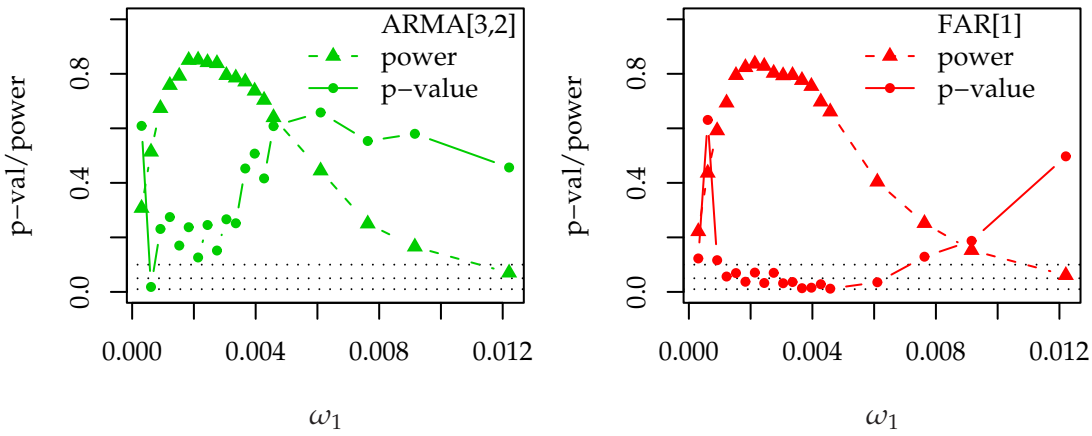

Figure B.2: p-values for different bandwidths together with the power for testing H_f (left) and H_g (right) based on the log-periodogram regression. The dotted horizontal lines mark the 1%, 5% and 10%-level (bottom to top).

we find p-values for H_f fluctuating above any reasonably level of significance in the left plot. Whereas testing H_g p-values close to, or even below the 5%-level of significance can

be observed in this region. This can be taken as indications for rejecting the FAR[1] in favour of the ARMA[3, 2].

Bandwidth-Dependence of the Estimate

We confront the estimates $\widehat{d}_{\text{lp}}$ for different bandwidth with the corresponding mean values and standard deviations of the ensembles. This might provide further indications for the model selection (Figure B.3). For a small bandwidth the fluctuation of the estimate $\widehat{d}_{\text{lp}}$

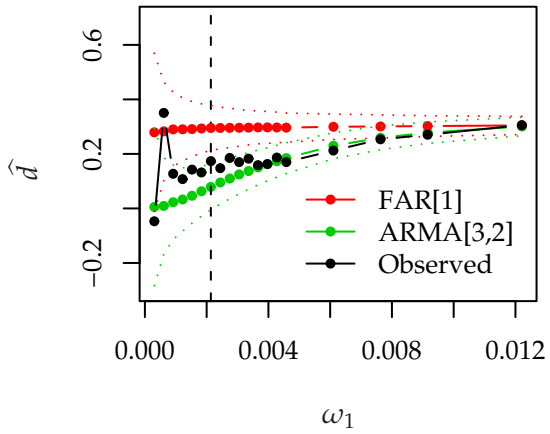

Figure B.3: Dependence of the mean and standard deviation of the distributions and the estimate on the bandwidth. The vertical line marks the bandwidth where maximal power is attained.

obtained from the observed series is large. It falters between the mean values of both hypothesis. With an increasing bandwidth this estimate approaches the course of the mean of H_f and stays close to it while the two distributions merge. This may be taken as further indication for the ARMA[3, 2] hypothesis.

B.4.2 Detrended Fluctuation Analysis

Bandwidth-Dependence of the Power

We calculate the power as a function of the range of scales (or bandwidth) for the straight line fit to the log-fluctuation function obtained with DFA (Figure B.4, left). For testing H_f, we find a maximum power at the standard 5%-level of $\widehat{\text{pow}}_f(g, 0.05) = 0.861$ for a range of scales $\log s > \log s_1 = 2.6$. The maximum power $\widehat{\text{pow}}_g(g, 0.05) = 0.851$ for H_g is attained for a range of scales $\log s > \log s_1 = 2.7$. Also using this statistic does not yield a test as powerful as the one based on the likelihood ratio (Section 4.4.1). The two distribution functions for the fractional difference parameter estimates are shown in Figure B.4 (right) together with the estimate of the "observed" series. Similar to the log-periodogram regression, we find the observed value $\text{lr}_{\text{obs}} = 0.190$ in a gap between the two critical regions. We can thus not reject neither hypothesis.

Calculating p-Values

We calculate the p-values for the range of scales with maximum power. This yields $\widehat{p}_f = 0.132$ for a test of compatibility with H_f and $\widehat{p}_g = 0.057$ for H_g. Figure B.5 shows

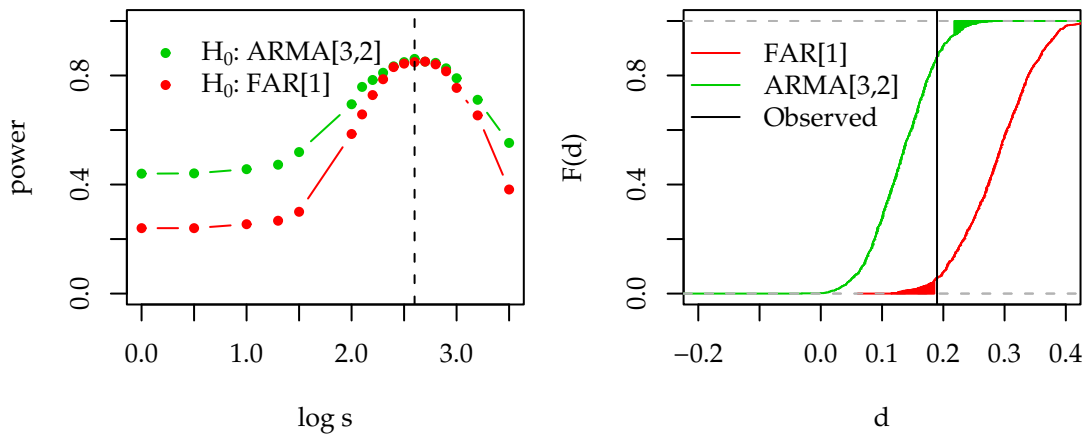

Figure B.4: Power of testing H_f hypothesis against H_g (green) and vice versa (red) for various ranges of scales $\log s > \log s_1$ in the time domain (left). Distributions of the estimated fractional difference parameters (right) obtained from the bootstrap ensembles using DFA in the time domain and the range of scales where maximal power for testing hypothesis f is achieved.

the p-values together with the respective power for a variety of ranges of scales used for the straight line fit. Here, we also find the p-values in regions of high power predominantly above a 10%-level for H_f and frequently below the 5%-level for H_g. We note that the p-values do not fluctuate as much as they do in the spectral domain. However, the difference to the 10%-level in the case of model f and the difference to the 5%-level in case of model g are also less pronounced.

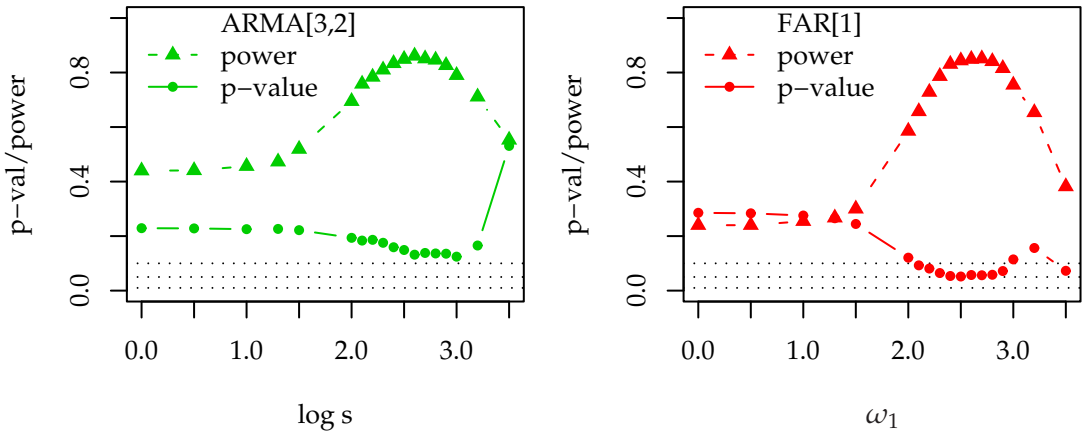

Figure B.5: Power and p-value for the testing H_f (left) and H_g (right) based on the DFA. The horizontal lines mark the 1%, 5% and 10%-levels (bottom to top).

Bandwidth-Dependence of the Estimate

The p-values changing with the range of scales used for the fit (Figure B.6) indicate that the estimates $\widehat{d}_{\mathrm{DFA}}$ rather follows the mean of H_f. The distinction is not as clear as it is in the case of the log-periodogram regression in Figure B.3. This, as well as the p-values obtained in the previous paragraph, may be taken as a hint to favour the ARMA$[3,2]$.

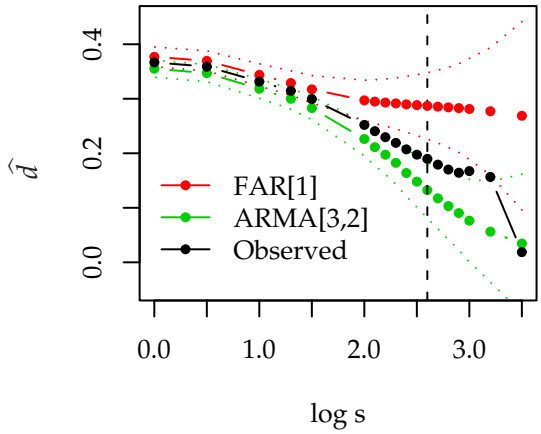

Figure B.6: Development of the mean and standard deviation of the hypothesis with varying range of scales for DFA. The vertical line marks the range where maximal power is attained.

B.5 Constructing the Example Process – Detailed Description

The aim of this section is to generate challenging example series for testing the strategies to detect LRD (Chapter 4). We start with a realisation $\{x_i\}_{i=1,\ldots,N}$ of an FARIMA$[1, d, 0]$ (or, for short, FAR[1]) process with parameters $a_1 = 0.7$, $d = 0.3$ and length $N = 2^{15} = 32\,768$. The length and the parameters are chosen such that they are plausible for river run-off, the context which we are going to apply this detection strategy in (Chapter 6). A realisation of this process is shown in the time and spectral domain in Figure B.7. Now,

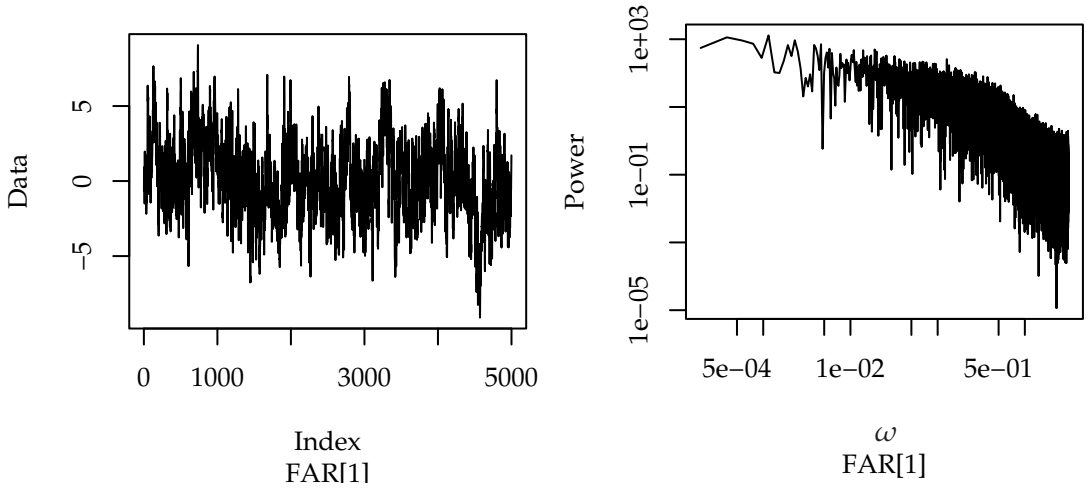

Figure B.7: Realisation of a FAR[1] process with parameters $a_1 = 0.7$, $d = 0.3$ and length $N = 32\,768$, in the time domain (left, only the first 5 000 data points) and in the frequency domain (right).

we aim to find a SRD model which describes this realisation well. We expect such a SRD model to produce realisations which will be difficult to discriminate from realisations of the original LRD process.

Selecting an ARMA$[p,q]$ model for the FAR$[1]$ Realisation

We fit various ARMA$[p,q]$ models with $p < 4$ and $q < \min(p,2)$ and compare the values for the Hannan-Quinn criterion (HIC) (cf. Section 3.2.2) in Figure B.8. The smallest value

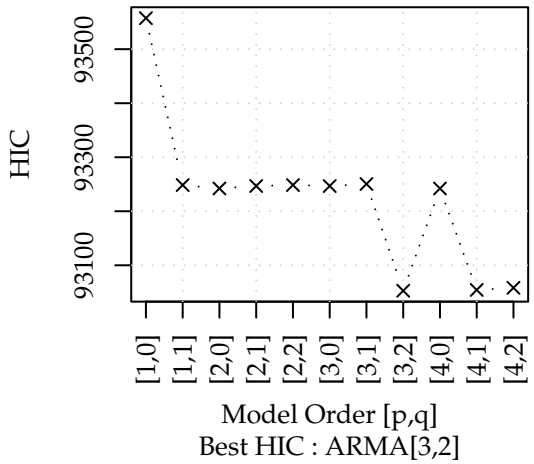

Figure B.8: Multiple ARMA$[p,q]$ fits to the realisation of the FAR$[1]$ process shown in Figure B.7. The plot shows the values for the Hannan-Quinn criterion (HIC) for models with the order $1 < p < 4$ and $0 < q < \min(p,2)$.

is attained for the ARMA$[3,2]$ process. Very close to this value, separated by less than 2 units, we find the ARMA$[4,1]$ model. About 4 units more yields the HIC for ARMA$[4,2]$. Note, that at this stage we do not necessarily intent to find one "best model". We are rather looking for some ARMA$[p,q]$ model representing the given FAR$[1]$ realisation well enough, such that realisations can be easily mistaken for coming from an LRD process. There is no need to discriminate between these three similar performing models. For our upcoming example, we choose the ARMA$[3,2]$ process with parameters $a_1 = 2.178, a_2 = -1.448, a_3 = 0.309, b_1 = -1.184$ and $b_2 = 0.218$. The model is completely specified and we can now generate a realisation, which are used to study the detection strategies for LRD in Chapter 4.

Appendix C

Bootstrap Methods for Confidence Interval Estimation

We develop and discuss non-asymptotic strategies to assess the variability of return level estimates from correlated data. These strategies are based on the bootstrap. With bootstrap we denote here the generation of an ensemble of artificial maxima series simulated from a model which was motivated by an empirical record (Davison and Hinkley, 1997). In the following, we discuss and compare four different approaches by means of a simulation study: given a known stochastic process, we generate an ensemble of realisations which is used to assess the "true" variability of a 100-year return level estimate. On the basis of one such realisation which is considered as an "observed" record we aim to quantify the estimators variability using the four bootstrap approaches. The results of the various approaches are compared to the "true" variability.

Prior to the discussion of the bootstrap approaches, we introduce the basic concepts of extreme value statistics and return level estimation which are required here.

C.1 Extreme Value Statistics and the General Extreme Value Distribution

We introduce the Fisher-Tippett theorem, one of the fundamentals in extreme value statistics and corresponding parameter estimation strategies for models of extremes. Furthermore, a generalisation of this theorem to stationary processes is discussed.

C.1.1 The Fisher-Tippett Theorem

The pivotal element in extreme value statistics is the three-types theorem, discovered by Fisher and Tippett (1928) and later formulated in full generality by Gnedenko (1943). It motivates a family of probability distributions, namely the general extreme value distributions (GEV), as models for block maxima from an observed record, e.g., annual maximum discharge. Consider the maximum

$$M_n = \max\{X_1, \ldots, X_n\} \tag{C.1}$$

of a sequence of n IID random variables $X_1, \ldots, X_n$ with common distribution function. This can be, for example, daily measured run-off at a gauge. M_n represents the maximum over n daily measurements. In hydrological applications the block size n defaults to $n = 365$, i.e. annual maxima. This choice eliminates the influence of seasonality. Larger

values for n are rarely possible because increasing n reduces the length of the maxima series.

The three types theorem states that the probability distribution of the maxima series converges to a specific distribution $G(z)$:

$$\Pr\{(M_n - b_n)/a_n \leq z\} \rightarrow G(z), \text{ as } n \rightarrow \infty, \tag{C.2}$$

with a_n and b_n being normalisation constants and $G(z)$ a non-degenerate distribution function known as the general extreme value distribution (GEV)

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}. \tag{C.3}$$

z is defined on $\{z \in \mathbb{R} | 1 + \xi(z-\mu)/\sigma > 0\}$. The model has a location parameter μ, a scale parameter σ and a form parameter ξ. The latter decides whether the distribution is of type II (Fréchet, $\xi > 0$) or of type III (Weibull, $\xi < 0$). The type I or Gumbel family

$$G(z) = \exp\left[-\exp\left\{-\left(\frac{z-\mu}{\sigma}\right)\right\}\right], \{z \in \mathbb{R} | -\infty < z < \infty\} \tag{C.4}$$

is obtained in the limit $\xi \rightarrow 0$ (Coles, 2001).

It is convenient to transform (C.2) into

$$\Pr\{M_n \leq z\} \approx G((z - b_n)/a_n) = G^*(z). \tag{C.5}$$

The resulting distribution $G^*(z)$ is also a member of the GEV family and allows the normalisation constants and the location, scale and shape parameter to be estimated simultaneously.

C.1.2 The Fisher-Tippett Theorem for Dependent Series

We consider an autocorrelated stationary series $\{X_1, X_2, \ldots\}$ and define a condition of near-independence: for all $i_1 < \ldots < i_p < j_1 < \ldots < j_q$ with $j_1 - i_p > l$,

$$|\Pr\{X_{i_1} \leq u_n, \ldots, X_{i_p} \leq u_n, X_{j_1} \leq u_n, \ldots, X_{j_q} \leq u_n\} - \tag{C.6}$$

$$\Pr\{X_{i_1} \leq u_n, \ldots, X_{i_p} \leq u_n\}\Pr\{X_{j_1} \leq u_n, \ldots, X_{j_q} \leq u_n\}| \leq \alpha(n, l), \tag{C.7}$$

where $\alpha(n, l_n) \rightarrow 0$ for some sequence l_n, with $l_n/n \rightarrow 0$ as $n \rightarrow \infty$. It can be shown that the three types theorem holds also for correlated processes satisfying this condition of near-independence (Leadbetter et al., 1983; Coles, 2001). This remarkable result implies that the limiting distribution of the maxima of uncorrelated and (a wide class) of correlated series belongs to the GEV family.

C.1.3 Parameter Estimation for the General Extreme Value Distribution

To fully specify the model, we estimate the GEV parameters for location (μ), scale (σ) and shape (ξ) from the data. Estimates can be obtained in several ways: probability weighted moments (Hosking et al., 1985; Hosking, 1990), maximum likelihood (Smith, 1985; Coles, 2001) or Bayesian methods (Smith, 1987; Coles and Tawn, 1996; Coles, 2001). These different approaches have advantages and also drawbacks which are discussed in, e.g., Hosking et al. (1985); Coles and Dixon (1999) and Smith (2001). In the following we focus on ML estimation as the most general method. Within this framework models can be easily extended, for example to non-stationary distributions (Coles, 2001) as used in (Kallache, 2007; Bárdossy and Pakosch, 2005).

Maximum-Likelihood Estimation

Let $\{M_{n,1}, M_{n,2}, \ldots, M_{n,m}\}$ be a series of independent block maxima observations, where n denotes the block size and m the number of blocks available for estimation. We denote $M_{n,i}$ as z_i. The likelihood function now reads

$$L(\mu, \sigma, \xi) = \prod_{i=1}^{m} g(z_i; \mu, \sigma, \xi), \tag{C.8}$$

where $g(z) = dG(z)/dz$ is the probability density function of the GEV. In the following, we consider the negative log-likelihood function $l(\mu, \sigma, \xi | z_i) = -\log L(\mu, \sigma, \xi | z_i)$. For the GEV, we have

$$l(\mu, \sigma, \xi | z_i) = -m \log \sigma - (1 + 1/\xi) \sum_{i=1}^{m} \log \left[1 + \xi \left(\frac{z_i - \mu}{\sigma} \right) \right] - \sum_{i=1}^{m} \left[1 + \xi \left(\frac{z_i - \mu}{\sigma} \right) \right]^{-1/\xi}. \tag{C.9}$$

with $1 + \xi \left(\frac{z_i - \mu}{\sigma} \right) > 0$, for $i = 1, \ldots, m$.

For the Gumbel distribution ($\xi \to 0$) extra treatment is required and the log-likelihood reads

$$l(\mu, \sigma, \xi | z_i) = -m \log \sigma - \sum_{i=1}^{m} \left(\frac{z_i - \mu}{\sigma} \right) - \sum_{i=1}^{m} \exp \left\{ - \left(\frac{z_i - \mu}{\sigma} \right) \right\}. \tag{C.10}$$

Minimising the log-likelihood with respect to $\boldsymbol{\theta} = (\mu, \sigma, \xi)^\dagger$ leads to the ML estimate $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}, \widehat{\sigma}, \widehat{\xi})^\dagger$ for the GEV. Under suitable regularity conditions – among them independent observations z_i – and in the limit of large block sizes ($n \to \infty$) $\widehat{\boldsymbol{\theta}}$ is multivariate normal distributed:

$$\widehat{\boldsymbol{\theta}} \sim \mathcal{N}_{\mathrm{MV}}(\boldsymbol{\theta}_0, I_E(\boldsymbol{\theta}_0)^{-1}) \tag{C.11}$$

with $I_E(\boldsymbol{\theta})$ being the expected information matrix (or Fisher information matrix) measuring the curvature of the log-likelihood. Denoting the elements of the inverse of I_E evaluated at $\widehat{\boldsymbol{\theta}}$ as $\beta_{j,k}$ we can approximate an $(1 - \alpha)100\%$-confidence interval for each component j of $\widehat{\boldsymbol{\theta}}$ by

$$\widehat{\theta}_j \pm z_{\frac{\alpha}{2}} \sqrt{\beta_{j,j}}, \tag{C.12}$$

with $z_{\frac{\alpha}{2}}$ being the $(1 - \alpha/2)$ quantile of the standard normal distribution (Coles, 2001).

For maximum-likelihood parameter estimation for the GEV, we use the package `evd` (Stephenson and Ferro, 2004) written for the open source statistical language environment R (R Development Core Team, 2004)[1].

Calculating an m-Year Return Level

The m-year return level can be calculated straightforwardly once the location, scale and shape parameter are estimated. In case of the Gumbel distribution ($\xi = 0$) the equation reads

$$\widehat{r}_m = \widehat{\mu} - \widehat{\sigma} \log(y), \tag{C.13}$$

with $y = -\log(1 - \frac{1}{m})$. For a non-zero shape parameter ξ we have

$$\widehat{r}_m = \widehat{\mu} - \frac{\widehat{\sigma}}{\widehat{\xi}} \left(1 - y^{-\widehat{\xi}} \right), \tag{C.14}$$

[1] Both are freely available from `http://cran.r-project.org`.

An approximated confidence interval for $\widehat{r}_m$ can be obtained under the hypothesis of a normally distributed estimator making use of the standard deviation $\sigma_{\widehat{r}_m}$. The latter can be calculated from the information matrix using the delta method (Coles, 2001). For the Gumbel distribution we obtain

$$\sigma^2_{\widehat{r}_m} = \beta_{11} - (\beta_{22} + \beta_{21}) \log\left(-\log\left(1 - \frac{1}{m}\right)\right) + \beta_{22}\left(\log\left(-\log\left(1 - \frac{1}{m}\right)\right)\right)^2. \quad \text{(C.15)}$$

C.2 Bootstrapping the Estimators Variance

We discuss four different strategies to generate a bootstrap ensemble. Three of them explicitely account for correlations in the series.

bootstrap$_{\mathrm{cl}}$ The first ansatz is a classical bootstrap resampling of the maxima (Efron, 1979; Davison and Hinkley, 1997), denoted in the following as *bootstrap*$_{\mathrm{cl}}$: one ensemble member is generated by sampling with replacement from the annual maxima series. Autocorrelation is not taken into account here.

iaaft$_{\mathrm{d}}$ We denote the second strategy as *iaaft*$_{\mathrm{d}}$, it makes use of the daily observations. Ensemble members are obtained using the iterative amplitude adjusted Fourier transform (IAAFT) – a surrogate method described in Section C.2.4. The IAAFT generates artificial series (so-called surrogates) preserving the distribution and the correlation structure of the observed daily record. Subsequently, we extract the maxima series to obtain an ensemble member. Linear correlation is thus accounted for in this case.

bootstrap$_{\mathrm{fp}}$ The third strategy is a full-parametric bootstrap approach, which is denoted as *bootstrap*$_{\mathrm{fp}}$. It is based on a parametric model for the distribution function and one for the autocorrelation function of the annual maxima. Here, the Fisher-Tippett theorem motivates a parametric model for the maxima distribution.

bootstrap$_{\mathrm{sp}}$ Based on the latter, we propose a semi-parametric approach, denoted as *bootstrap*$_{\mathrm{sp}}$. It similarly uses a parametric model for the ACF of the maxima series. Instead of the GEV, we choose a non-parametric model for the distribution function.

While the first two strategies are common tools in time series analysis and are well described elsewhere (Davison and Hinkley, 1997; Schreiber and Schmitz, 2000), we focus on describing the full-parametric and semi-parametric bootstrap strategy. Before going into detail, we discuss the main idea, which is used to "emulate" sampling from a dependent process.

C.2.1 Emulating Dependence

A return level estimate is derived from an empirical maxima series which we consider as a realisation of a stochastic process. The aleatoric uncertainty of such an estimate is basically a consequence of the variability between different realisations of this process. The latter is in general larger for samples from dependent processes. A simple example illustrates this effect in the following.

As a measure for the variability between the empirical distributions of different realisations/samples, we consider the deviation of a realisation's empirical distribution

function from the true distribution function. It is more likely to have obtained a good representative for the true maxima distribution from one sample if this variability is low. Consider two long realisations ($N = 10\,000$), one from an LRD (FARIMA$[1, d, 0]$, Section 2.2.2) and one from an uncorrelated process. We compare the difference between the empirical cumulative distribution functions $\widehat{F}_s(x)$ of short sections and the empirical cumulative distribution $\widehat{F}_0(x)$ of the entire realisation by means of the Kolmogorov-Smirnov distance $D = \max_x |\widehat{F}_s(x) - \widehat{F}_0(x)|$ (DeGroot, 1975). Smaller distances D indicate a larger similarity between $\widehat{F}_s$ and $\widehat{F}_0$. Figure C.1 shows the empirical cumulative distribution function of differences D for various sections. For the correlated process, we find the dis-

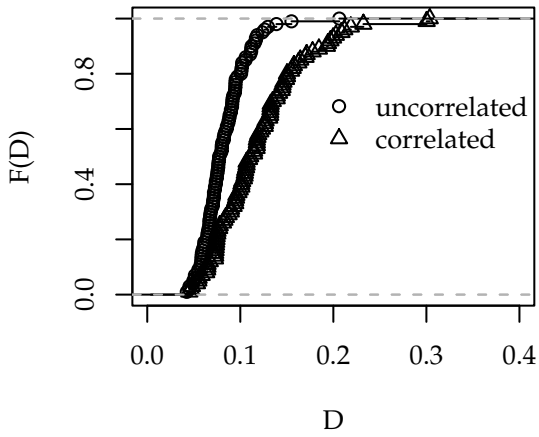

Figure C.1: Empirical cumulative distribution function $\widehat{F}(D)$ of the Kolmogorov-Smirnov distances D between sections of 100 points and the entire series of 10 000 points. For the uncorrelated record (circles) distances D are located at smaller values than for the correlated record (triangles).

tribution of distances D located at larger values. This implies a larger diversity among the distributions than for the uncorrelated process. Realisations of correlated processes are therefore not as likely to yield as representative results for the underlying distribution as a comparable sample of an uncorrelated process.

Because the variability of the return level estimator is a consequence of the corresponding maxima distribution, we employ this illustrative example for our purposes and derive a strategy to emulate the sampling from a dependent series. Instead of sampling from a dependent process with the desired marginal distribution, we use the following procedure: 1) generate a long sample from an uncorrelated process with the desired marginal distribution, 2) rearrange the values such that they satisfy a desired autocorrelation structure, and 3) divide the long record into sections of the original length and estimate the desired quantities from them. Due to the rearrangement, similar values are nearby (in time) and samples, i.e. short sections, get less informative regarding the marginal distribuion function. Ideally, the marginal distribution function and the ACF used to generate this long record should be close to the underlying properties of the process under consideration. This means to provide a satisfying model for the distribution of the maxima series and for its ACF. Realisations of these two models are then combined to obtain one sample satisfying both characteristics, the desired distribution function and the ACF.

In the following, we describe the way these two models are obtained and how their realisations are combined.

C.2.2 Modelling the Distribution

The aim of modelling the distribution is to provide means to generate realisations from this model used in a later step of the bootstrap procedure. For the semi-parametric approach, the distribution of the maxima is modelled by the empirical cumulative distribution function from the observed series. Realisations can be obtained simply by sampling with replacement from the observed maxima series (Davison and Hinkley, 1997). The full-parametric ansatz instead exploits the Fisher-Tippett theorem for extreme values (Section C.1). It uses the parametric GEV family as a model for the maxima. Realisations are then obtained directly by sampling from this model.

C.2.3 Modelling the ACF using FARIMA$[p, d, q]$ Processes

The autocorrelation function is modelled using FARIMA$[p, d, q]$ processes (Section 2.2.2), for parameter estimation and model selection we use the Whittle estimator (Section 2.3.2) and the model selection strategies described in Chapter 3.

There is, however, a further difficulty to overcome: the modelling of the ACF of the maxima series is not straightforward. The reason is that the maxima series are in practice relatively short and, analysed with standard methods, appear to be weakly correlated or even uncorrelated, as exemplified with the ACF estimate in Figure 7.2 (left, solid line). Direct modelling using the empirical maxima series is therefore frequently not feasible. To supermount this difficulty, we first model the ACF of the daily run-off records with a FARIMA$[p, d, q]$ process and assume this to be a good representation of the true ACF. We use this model to generate one or more series sufficiently long to model the ACF of their maxima series with a FARIMA$[p, d, q]$ process. The latter is then assumed to be a reasonable model for the empirical maxima series as well. However, generating time series longer than the length of the empirical one means to interpret the ACF (or the spectrum) of the model on time scales larger than the length of the empirical series.

C.2.4 Combining Distribution and Autocorrelation

Having a model for the distribution function and for the ACF, we can generate realisations, i.e. a sample $\{X_i\}_{i=1,\ldots,N}$ from the distribution model and a series $\{Y_i\}_{i=1,\ldots,N}$ from the FARIMA$[p, d, q]$ model. To obtain a time series $\{Z_i\}_{i=1,\ldots,N}$ with distribution equal to the one of $\{X_i\}$ and ACF comparable to that of $\{Y_i\}$, we employ the iterative amplitude adjusted Fourier transform (IAAFT), described in the following.

The IAAFT was developed by Schreiber and Schmitz (1996) to generate surrogate time series used in tests for nonlinearity (Theiler et al., 1992). The surrogates are generated such that they retain the linear part of the dynamics of the original time series including a possible non-linear static transfer function. This implies that the power spectrum (or ACF, equivalently) and the frequency distribution of values are conserved. The algorithm basically changes the order of the elements of a record in a way that the periodogram stays close to a desired one (Schreiber and Schmitz, 2000).

Besides using the IAAFT on the daily series to generate an ensemble of surrogates (denoted as $iaaft_{\mathrm{d}}$), we employ this algorithm also to create records $\{Z_i\}$ with a periodogram prescribed by a series $\{Y_i\}$ and a frequency distribution coming from $\{X_i\}$.

C.2.5 Generating Bootstrap Ensembles

With the preceedingly described three building blocks we are now able to obtain realisations with distribution and ACF of comparable to the empirical maxima series. This provides the basis to generate the full-parametric and semi-parametric ensembles $bootstrap_{\mathrm{fp}}$ and $bootstrap_{\mathrm{sp}}$, respectively. The strategies can be outlined as follows:

0. If necessary, transform the daily run-off data (length N_{data}) such that this the distribution follows approximately a Gaussian distribution. In many practically relevant cases, this can be achieved using a logarithmic or the more general Box-Cox transformation (Section D.2.2; Hipel and McLeod, 1994; Box and Cox, 1964).

1. Model the correlation structure of the empirical maxima:
   (a) Remove periodic cycles (e.g., annual, weekly).
   (b) Model the correlation structure of the resulting anomaly series (e.g., daily or monthly) with FARIMA$[p, d, q]$ processes (Section 2.2.2) and select the most suitable model (Chapter 3).
   (c) Generate one or more long series from this model ($N_{\mathrm{long}} \gtrsim 100 N_{\mathrm{data}}$).
   (d) Add the periodic cycles from step 1a. The result is a long series sharing the spectral characteristics, especially the seasonality with the empirical record.
   (e) Extract the annual maxima series.
   (f) Model the ACF of this series using a FARIMA$[p_{\mathrm{max}}, d_{\mathrm{max}}, q_{\mathrm{max}}]$[2] process, select an appropriate model. In case several realisations are obtained in step 1c, the parameter estimates can be averaged.

2. Model the distribution of the empirical maxima series (length N_{max}) according to the approach used:

   bootstrap$_{\mathrm{fp}}$**:** Estimate the parameters of a GEV model from the empirical maxima series using, for example, maximum-likelihood estimation (Section C.1.3).

   bootstrap$_{\mathrm{sp}}$**:** Use the empirical maxima distribution as model.

3. Generate an ensemble of size N_{ens} of maxima series (length N_{max}) with correlation structure and value distribution from the models found in 1 and 2:
   (a) Generate a series $\{Y_i\}_{i=1,\ldots,N_{\mathrm{ens}}N_{\mathrm{max}}}$ with the FARIMA$[p_{\mathrm{max}}, d_{\mathrm{max}}, q_{\mathrm{max}}]$ model from step 1f.
   (b) Generate a sample $\{X_i\}_{i=1,\ldots,N_{\mathrm{ens}}N_{\mathrm{max}}}$ according to the approach used:
   bootstrap$_{\mathrm{fp}}$**:** from the GEV model specified in step 2a.
   bootstrap$_{\mathrm{sp}}$**:** from sampling with replacement from the empirical maxima series.
   (c) By means of IAAFT $\{X_i\}$ is reordered such that its correlation structure is similar to that of $\{Y_i\}$. This yields $\{Z_i\}_{i=1,\ldots,N_{\mathrm{ens}}N_{\mathrm{max}}}$.
   (d) Splitting $\{Z_i\}$ into blocks of size N_{max} yields the desired ensemble of maxima series.

4. Back-transform the ensemble members applying the inverse transformation from step (0).

[2] Here, the index max indicates that these parameters or model orders pertain to the maxima series

Finally, estimate the desired return level from each ensemble member (Section C.1.3). This yields a frequency distribution of return level estimates which can be used to assess the variability of this estimator.

C.3 Comparison of the Bootstrap Approaches

On the basis of a simulation study, we carry out a comparison of the four different bootstrap approaches *bootstrap*$_{\text{cl}}$, *iaaft*$_{\text{d}}$, *bootstrap*$_{\text{fp}}$, and *bootstrap*$_{\text{sp}}$. The stochastic process we use here is chosen such that its correlation structure as well as its value distribution are plausible in the context of river run-off. This is a FARIMA$[p, d, q]$ process similar to those used by Montanari et al. (1997) to model run-off. Since the marginal distribution of this process as described in Section 2.2.2 is Gaussian, we subsequently carry out an exponential transformation to obtain a log-normal distribution. This process (including the transformation) is used to generate a Monte Carlo ensemble of (simulated) daily run-off series. For each ensemble member we extract the maxima series and estimate a 100-year return level. The resulting distribution of 100-year return level estimates represents the estimator's variability for this process. In the following, this distribution is used as a reference to measure the performance of the various bootstrap approaches. In order to call a bootstrap approach useful, we require it to reproduce this reference distribution of return level estimates reasonably well on the basis of one ensemble member. In practical applications, this one ensemble member is replaced by an observed series.

We now take a representative realisation out of this ensemble and consider it as a record, we possibly could have observed. On the basis of this "observed" series, we generate the four bootstrap ensembles according to the approaches presented in Section C.2.5. The resulting four frequency distributions of the 100-year return level estimates are then compared to the distribution of the reference ensemble and to the asymptotic distribution of the ML-estimator for IID observations.

C.3.1 Monte Carlo Reference Ensemble

We simulate the series for the reference ensemble with a FARIMA$[1, d, 0]$ process with parameters $d = 0.25$ (or $H = 0.75$), $\phi_1 = 0.9$, variance $\sigma^2 \approx 1.35$, zero-mean Gaussian driving noise $\eta \sim \mathcal{WN}(0, \sigma_\eta)$, and length $N = 36\,500$ (100 years of daily observations). The skewness typically found for river run-off is achieved by subsequently transforming the records to a log-normal distribution. To resemble the procedure of estimating a 100-year return level we extract the annual maxima ($n = 365$) and estimate a 0.99 quantile using a Gumbel distribution as parametric model. From an ensemble of 100 000 runs, we obtain a distribution of 100-year return levels (0.99 quantiles) serving as a reference for the bootstrap procedures.

C.3.2 The Bootstrap Ensembles

On the basis of the "observed" series we generate the four bootstrap ensembles *iaaft*$_{\text{d}}$, *bootstrap*$_{\text{cl}}$, *bootstrap*$_{\text{fp}}$, and *bootstrap*$_{\text{sp}}$. Following the outline in Section C.2.5, we start with a log-transformation as a static transformation (step 0). Because there are no periodic cycles in this example, we skip step 1a and start modelling the correlation structure of the "observed" daily series using FARIMA$[p, d, q]$ processes (Section C.2.3) and the proposed model selection strategies (Chapter 3). The model selected is a FARIMA$[1, d, 0]$ with parameters and asymptotic standard deviation: $d = 0.250(0.008), \phi = 0.900(0.005)$

and $\sigma_\eta^2 = 0.0462$ (step 1b). This model is used to generate an artificial series longer than the original one ($N_{\text{long}} = 100 N_{\text{data}}$, step 1c). The extracted annual maxima series (step 1e) contains $N_{\text{max}} = 10\,000$ data points. Modelling this artificial maxima series results in a FARIMA$[0, d, 0]$ (step 1f) with $d = 0.205(0.008)$ and $\sigma_\eta^2 = 0.535$.

The next step (2) consists of building a model for the distribution of the maxima series. For the full-parametric bootstrap ensemble $bootstrap_{\text{fp}}$, we get back to the "observed" series and model the annual maxima ($N_{\text{max}} = 100$) with a parametric Gumbel distribution (step 2a). This results in ML-estimates for the location and scale parameters: $\mu = 10.86, \sigma = 8.35$. The semi-parametric approach $bootstrap_{\text{sp}}$ does not need a model but uses a classical bootstrap resampling from the empirical annual maxima series. We can now generate the desired bootstrap ensembles $bootstrap_{\text{fp}}$ and $bootstrap_{\text{sp}}$ both with 1 000 ensemble members ($N_{\text{ens}} = 1\,000$, step 3). Finally, the ensemble is back-transformed using the exponential function (step 4).

Figure C.2 compares the frequency distributions of estimated return levels from the four bootstrap ensembles to the reference distribution (grey filled) and to the asymptotic distribution of the ML-estimator (dotted). The left plot shows the result of the $bootstrap_{\text{cl}}$ (solid) and the $iaaft_{\text{d}}$ (dashed) ensembles. While $bootstrap_{\text{cl}}$ accounts for more variability than the asymptotic distribution, $iaaft_{\text{d}}$ exhibits less variability, although it takes autocorrelation of the daily data into account. This might be due to the fact that the records in the $iaaft_{\text{d}}$ ensemble consists of exactly the same daily run-off values arranged in a different order. While this allows for some variability on the daily scale, an annual maxima series extracted from such a record is limited to a much smaller set of possible values. Because the temporal order of the maxima series does not influence the return level estimation, the variability of the estimates is reduced.

The right panel in Figure C.2 shows the result from the $bootstrap_{\text{sp}}$ (solid) and the $bootstrap_{\text{fp}}$ (dashed) ensembles. The latter strategy is slightly better than $bootstrap_{\text{cl}}$ but still yields a too narrow distribution. In contrast, the result from the $bootstrap_{\text{sp}}$ ensemble gets very close to the reference ensemble. Thus, this approach is a promising strategy to improve the uncertainty analysis of return level estimates and is studied in more detail in the following section.

C.3.3 Ensemble Variability and Dependence on Ensemble Size

The dependence on the ensemble size can be studied performing an extensive simulation study. We generate different sets of $bootstrap_{\text{sp}}$ ensembles, each set containing 100 ensembles of a fixed size. We are interested in the variability of the ensemble runs within one set of fixed size and in the effect of the ensemble size. The ensemble size varies between $N_{\text{ens}} = 50$ and $N_{\text{ens}} = 6\,000$. Figure C.3 shows these quantiles estimated from these ensembles for different ensemble sizes.

The variability of the quantile estimates within a set of ensembles of the same size decreases with increasing ensemble size, as indicated by the converging clouds of grey dots. Consequently, the ensemble size should be chosen according to the accuracy needed. For small sizes, the means of the selected quantiles are close to the values from the reference ensemble, especially for the three upper quantiles. With an increasing size, difference to the reference ensemble increases for the extreme quantiles until they stagnate for ensembles with more than about 2 000 members. In this range, the the difference between the 95% quantile of the bootstrap and the reference ensemble is less than 4% of the return levels estimate. A reason for this mismatch might be the dependence of ensemble members in one ensemble. One might thus hypothesise, that also the quantiles estimated from

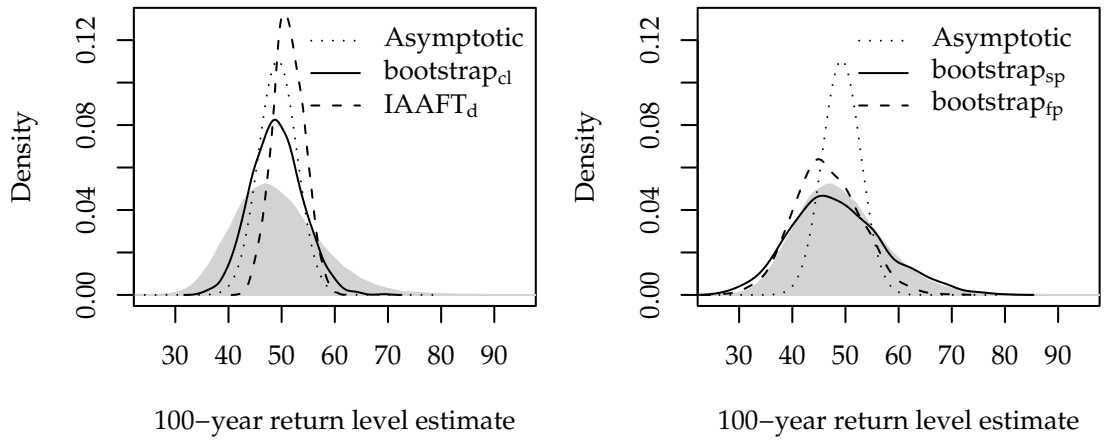

Figure C.2: Comparison of the result for different bootstrap ensembles to the MC reference ensemble (grey area) and the asymptotic distribution of the ML-estimator (dotted). The bootstrap ensembles consist each of 1 000 members. The left plot shows the results of the non-parametric bootstrap resampling and the daily IAAFT surrogates. The full-parametric and semi-parametric bootstrap strategies are shown in the right plot.

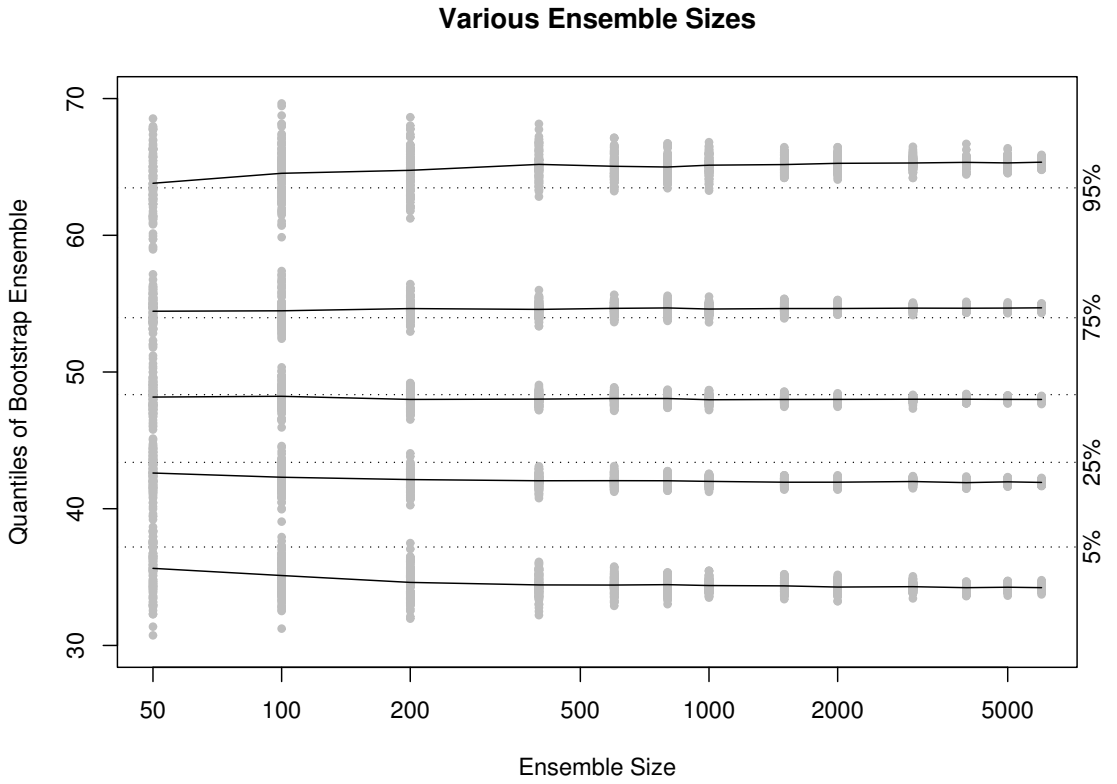

Figure C.3: The quantiles of the semi-parametric bootstrap ensembles of different size. 100 ensembles of the same size are grouped in a set. The 5%, 25% 50% 75% and 95%-quantiles of each ensemble in a set is marked with a grey dot. This results in grey areas representing the variability within a set of ensembles. The solid lines connect the mean values of the set for each quantile. The quantiles from the reference ensemble are represented as a dotted line.

the ensemble members are dependent. It is conceivable to generate independent ensemble members, simply by cutting one member out of the long series, instead of N_{ens}, and repeat this N_{ens} times. This is, however, numerically more costly.

C.4 Summary

We develop and compare four methods to generate bootstrap ensembles of maxima series which are used to quantify the aleatoric uncertainty of a return level estimator in the presence of correlation. With a simulation study, we could verify that one of the four strategies, the semi-parametric bootstrap $bootstrap_{\text{sp}}$, works reasonably well and outperform the other three strategies. This approach is studied in more detail with respect to inter-ensemble variability and a variable ensemble size. This variability among ensembles decreases with an increasing ensemble size. For large ensemble sizes, the mean values for the quantiles obtained from one set of ensembles does not coincide with the quantiles of the reference ensemble. The difference is small compared to the quantiles obtained on the basis of the IID assumption. A reason for this mismatch might be the dependence of the ensemble members within one ensemble. This can be a starting point for further investigation.

The approach is presented in the framework of GEV modelling of annual maxima using maximum likelihood. The concept can also be applied in the context of other models for the maxima distribution (e.g., log-normal, Pearson type III, etc.) or different parameter estimation strategies (e.g., probability weighted moments). Furthermore, it is conceivable to extend the class of models describing the dependence to FARIMA$[p, d, q]$ models with conditional driving noise (FARIMA-GARCH (Elek and Márkus, 2004)) or seasonal models (Montanari et al., 2000; Lohre et al., 2003).

The properties of the modelling approach using FARIMA$[p, d, q]$ and a subsequent adjustment of the values has been investigated in more detail using simulation studies by Venema, Bachner, Rust, and Simmer (2006). It was demonstrated that the combination of FARIMA$[p, d, q]$ models and the IAAFT is able to reproduce also other characteristics of time series then the distribution and power spectrum. Also the increment distribution and structure functions for river run-off are reasonably well recovered.

In the approach described, we obtain a model for the ACF of the maxima series only with the help of a model of the daily series. The longer this daily series is the more reliable the model will be. It is, however, also possible to include available annual maxima in the procedure for periods where daily series have not been recorded

Appendix D

Data Sources and Preprocessing

This part of the appendix gives the sources of the data which was analysed in this thesis. Furthermore, the methods used to preprocess these records are briefly explained and, in addition, the preprocessing itself is presented in detail for the Prague daily mean temperature studied in Section 5.1, and the three run-off records analysed in Chapter 6.

D.1 Data Sources

Temperature

The record of the maximum temperature measured at the Clementinum in Prague (Czech Republic) was obtained from the `xDAT` data based hosted at the Potsdam Institute for Climate Impact Research (Potsdam, Germany) and maintained by the *Scientific Data Management Group*: M. Flechsig, A. Glauer, C. Rachimow, M. Wrobel. The data set was downloaded 22 April 2003 and was provided by the Czech Hydro-meteorological Institute.

Northern hemisphere mean temperatures from 1856 to 2005 at a monthly resolution was obtained from the Climate Research Unit (University of East Anglia). The data set used here can be found at `http://www.cru.uea.ac.uk/ftpdata/tavenh2v.dat`. A detailed description on how the data set was obtained is given by Jones et al. (1999).

Run-Off

The run-off data from the Danube river at Achleiten and the Große Vils at Vilsbiburg were kindly provided by the Bayerisches Landesamt für Umwelt (München, Germany) in person of Dr. J. Neummann within the framework of the project "Skalenbezogene Analyse hydrologischer Zeitreihen: Bestimmung von Trends, Charakterisierung der Fluktuationen, stochastische Modellierung und Extremwertstatistik" financed by the German Federal Ministry of Education and Research.

Vistula run-off data measured at the gauge Tczew was obtained from the Global Runoff Data Centre at the Federal Institute of Hydrology (Koblenz, Germany). The data file was generated at 20 September 2006. The GRDC internal numbering is 6458010.

D.2 Preprocessing

As preprocessing, we understand a) a static transformation function applied to the observed values in order to symmetrise the frequency distribution and to obtain a more

Gaussian like shape, and b) the estimation and subtraction of periodic cycles and polynomial trends. A periodic cycle is, for example, an annual cycle in the case of temperature records and, additionally, a potential weekly cycle for the run-off records.

The static transformation function as well as the estimation and subtraction of the periodic cycles are explained in the following. This is followed by a detailed description of the preprocessing of the individual data sets.

D.2.1 Estimation of Periodic Cycles

A description of a periodic component can be obtained by estimating the mean and the variance for days with a specified position s in the cycle, i.e. for an annual cycle: 1 January, 2 January, and so on. For a record with n cycles of length S, we have (Hipel and McLeod, 1994)

$$\widehat{\mu}_s = \frac{1}{n}\sum_{j=1}^{n} x_{s,j}, \quad s = 1, 2, \ldots, S, \tag{D.1}$$

where s denotes the location in the cycle and j indexes the successive cycles. For daily data and an annual cycle $S = 365$. Analogously, one can estimate the variance as

$$\widehat{\sigma}_s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_{s,j} - \widehat{\mu}_s)^2, \quad s = 1, 2, \ldots, S. \tag{D.2}$$

We can now calculate a series of normalised anomalies by

$$\tilde{y}_{s,j} = \frac{x_{s,j} - \widehat{\mu}_s}{\widehat{\sigma}_s}, \tag{D.3}$$

or, even simpler, if there is no need to account for seasonality in the variance

$$\tilde{y}_{s,j} = x_{s,j} - \widehat{\mu}_s. \tag{D.4}$$

The latter expression can be derived within the framework of linear filtering as shown, e.g., by Hipel and McLeod (1994).

D.2.2 Box-Cox Transformation

With a non-linear static transformation function asymmetries and discrepancies to a Gaussian distributions can be reduced. Such a transformation is the Box-Cox transformation (Box and Cox, 1964). For records $\{x_i\}_{i=1,\ldots,N}$ with $x_i > 0$ for all $i \in \{1, \ldots, N\}$, it is defined as

$$y = \begin{cases} \frac{(x^\lambda - 1)}{\lambda}, & \lambda \neq 0 \\ \log(x), & \lambda = 0 \end{cases}.$$

The parameter λ is chosen such that the unconditional Gaussian likelihood is maximised. In the context of run-off records, the use of the Box-Cox transformation is discussed by, e.g., Hipel and McLeod (1994).

D.3 Preprocessing of Run-Off and Temperature Records

D.3.1 Prague Daily Maximum Temperatures

Annual Cycle

The most dominant feature in the Prague temperature record is the annual seasonal cycle which we interpret as a deterministic time dependent variation of the mean and possibly the variance with a fixed period. An estimate of this cycle is shown in Figure D.1 (left). The temperature maximum is approximately at the beginning of August. A minimum variance can be observed in October. We can now calculate a series of normalised temperature anomalies (Figure D.1, right) using (D.3).

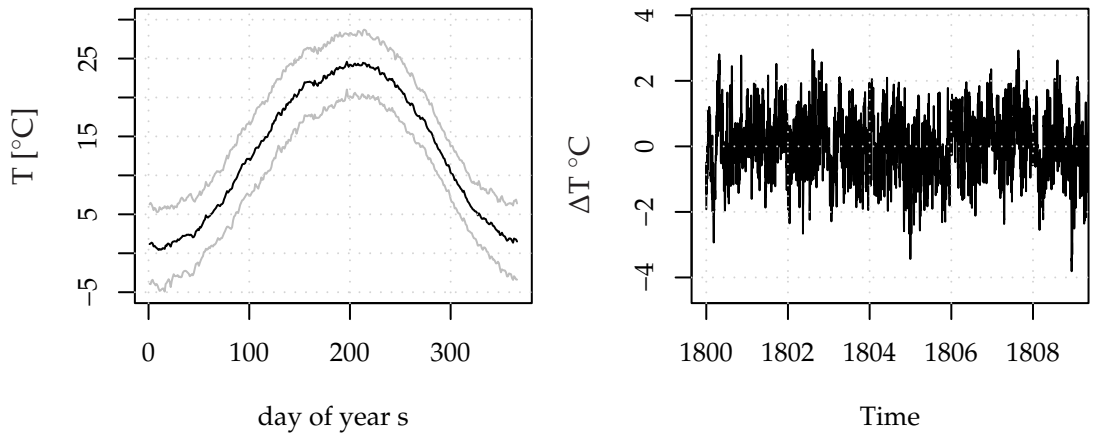

Figure D.1: Mean annual cycle for Prague maximal temperature series (left). The lines mark the mean (black) plus/minus one standard deviation (grey). The right plot shows the first ten years of the normalised temperature anomalies ΔT.

Box-Cox Transformation

We check the compatibility of the deseasonalised record with a Gaussian distribution using a quantile-quantile plot (QQ-plot, Figure D.2, left). Besides the tails the agreement is already quite good. A Box-Cox transformation can be used to transform the record such that its distribution is closer to a Gaussian. The improvement is small in this case. The likelihood for the transformation attains its maximum setting $\widehat{\lambda} = 1.07$. Although the improvement is minor, we consider in the following the Box-Cox transformed record.

Trend in the Mean

To investigate for a potential instationarity in the mean, we plot the Box-Cox transformed temperature anomalies together with a ten year running mean (red, Figure D.2). The plot suggests a small decrease in the mean value during the first half of the 19th century and an increase during the 20th century. Because the trend is not monotonic, we consider a polynomial of second order,

$$y = a + bt + ct^2, \tag{D.5}$$

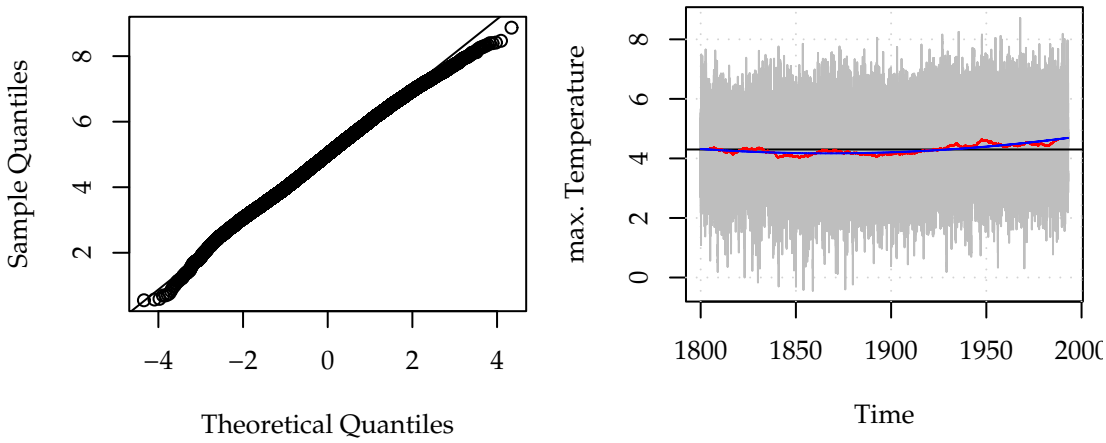

Figure D.2: QQ-plot against a Gaussian for the normalised temperature anomalies (left). Deseasonalised and Box-Cox transformed anomalies (right, grey) with the overall mean (black) a 10-year running mean (red) and a polynomial fit of second order (blue).

to parametrically describe the trend. The regression parameters are estimated using least-squares and the result is shown as a green line (Figure D.2, right). The record with this trend component subtracted is referred to as residuals in the following.

It is not straightforward to decide whether the second order polynomial is to be preferred to, e.g., a straight line or – even more fundamental – if such a deterministic trend component is significant in a statistical sense at all. The reliability of such a statistical test depends on the assumption of the underlying stochastic process (cf. Cohn and Lins, 2005). Standard approaches like the F-test, likelihood-ratio test or t-test assume a white noise process, i.e. independent observations. Because we have a strongly correlated record, as shown by the autocorrelation series (Figure 5.2), this assumption is not adequate. More refined strategies are thus needed. We suggest a conservative test accounting for correlation in Section 5.1.3. More advanced, non-parametric trend detection strategies for correlated series are discussed by, e.g., Kallache, Rust, and Kropp (2005), Craigmile et al. (2004) and Kallache (2007).

D.3.2 Danube Daily Mean Run-Off at Achleiten

Box-Cox Transformation

The histogram of the Danube run-off record (Figure 6.2, right) reveals the asymmetric shape of the observed run-off frequency distribution. We use a Box-Cox transformation (Section D.2.2) with a ML estimate of the parameter $\widehat{\lambda} = -0.098$ to obtain a more Gaussian-like shape. A comparison of the transformed record to a Gaussian distribution is carried out after subtraction of the periodic cycles and the trend (Figure D.5, right).

Periodic Cycles

For the transformed series, we estimate an annual periodic component for the mean and variance (Appendix D.2.1). Figure D.3 (left) depicts one cycle of this component. The discharge peaks at about the middle of the year. Having removed the annual cycle, peaks in the periodogram of the resulting anomalies Q^* at $f = 1/7$, $f = 2/7$ and $f = 3/7$

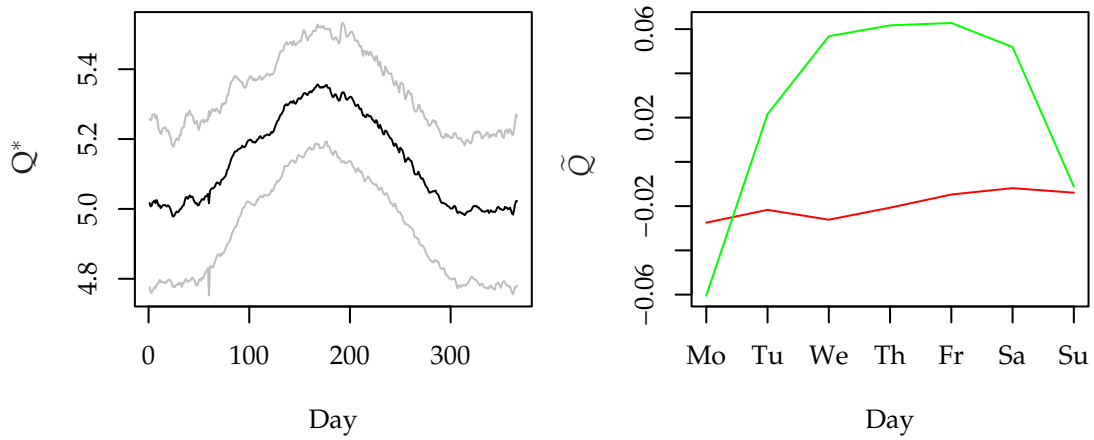

Figure D.3: Mean annual cycle for the Achleiten run-off record (left), mean (black) plus/minus one standard deviation (grey). The right plot shows the mean weekly cycle for the period 1901-1959 (red) and 1960-2003 (green). To facilitate the presentation, the standard deviation is not shown here. Q^* denotes the Box-Cox transformed run-off and $\tilde{Q}$ is a dimensionless quantity.

(corresponding to periods of 7 days and higher harmonics) [1] give evidence for an additional weekly component (Figure D.4, left). In a first attempt, we estimate and remove a weekly cycle analogously to the annual one. This reduces but not eliminates the peaks in the periodogram at the considered frequencies. Contrary to the annual cycle, the weekly periodicity is of anthropogenic nature, e.g. due to flow regulations or industrial use. This cycle is thus not necessarily constant within the time period considered. An analysis based on the windowed Fourier transform suggests a major change in the amplitude of the weekly cycle at about 1960. Therefore, instead of assuming a constant cycle over the full observation period, we allow two different cycles: one for the periods 1901 to 1959 and a second for 1960 to 2003 (Figure D.3, right). The σ intervals are large compared to the variability of the mean values and we refrain from plotting them to obtain a clearer picture of the difference in the cycles' amplitudes. Compared to the pronounced annual seasonality, these cycles are rather small in magnitude. They are, nevertheless, subtracted, since they hinder the further analysis. Figure D.4 (right) reveals a substantial reduction in the peaks of the periodogram, corresponding to the weekly cycle.

Trend in the Mean

This preprocessing leaves us with the anomaly series depicted in Figure D.5 (left). A visual inspection of the ten-year running mean (red) suggests no noticeable tendency. The straight line fit (green) to the dimensionless values yields an increase of $3.14 \cdot 10^{-3}$ per year. If we assume uncorrelated observations, a t-test for the slope being compatible with zero results in a p-value of $p = 0.07$ and is thus not significant on a 5%-level. A significance test as well as confidence intervals for a correlated series is not straightforward (Appendix D.3.1). In this case, however, the trend will as well not be significant (on the 5%-level) if we account for the autocorrelation (cf. Cohn and Lins, 2005).

The QQ-plot of the anomalies versus a Gaussian distribution shows compatibility for the central part of the distribution but reveals minor deviations in the tails (Figure D.5,

[1] If the weekly cycle is not well approximated by a sinusoidal function at the respective frequency ($f = 1/7$), higher harmonic components at frequencies $f \in \{2/7, 3/7\}$ can be used to ameliorate the description.

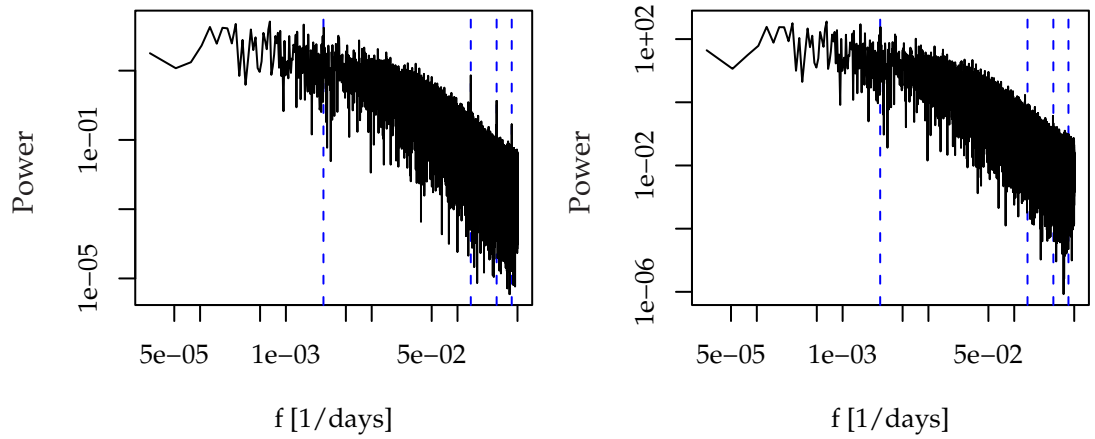

Figure D.4: Periodogram of the Danube run-off anomalies at Achleiten with the annual cycle in mean and variance removed (left). The same record with additionally the weekly periodic component removed (right). The dashed blue lines mark the frequencies corresponding to the annual ($f = 1/365$) and weekly periodicities ($f = 1/7$, $f = 2/7$ and $f = 3/7$).

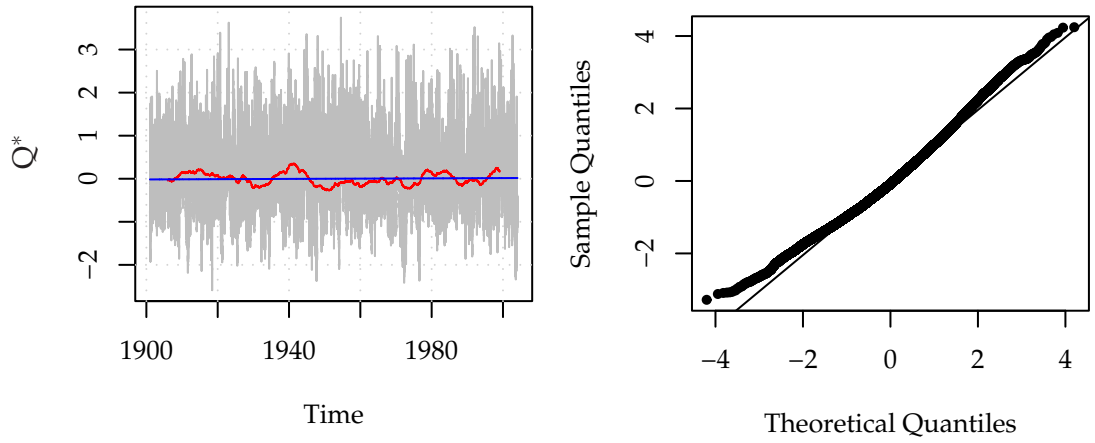

Figure D.5: Danube run-off anomalies as time series (grey, left) with a 10-year running mean (red) and a straight-line fit (blue). The right panel compares the residuals to a Gaussian in a QQ-plot.

right).

D.3.3 Große Vils Daily Mean Run-Off at Vilsbiburg

Box-Cox Transformation

The histogram in Figure 6.7 (right) reveals the asymmetric frequency distribution of the observed run-off values. We thus initially use a Box-Cox transformation (Section D.2.2) with a maximum-likelihood estimate of the exponent $\widehat{\lambda} = -0.588$ to obtain a more Gaussian-like shape. The comparison of the transformed record to a Gaussian is carried out after the periodic cycles and the trend have been subtracted (Figure D.8, right).

Periodic Cycles

For the transformed series, we estimate an annual periodic component for the mean and variance analogously to Section D.3.1. Figure D.6 (left) depicts the annual cycle in mean and standard deviation. The period of increased discharge is winter. After subtracting

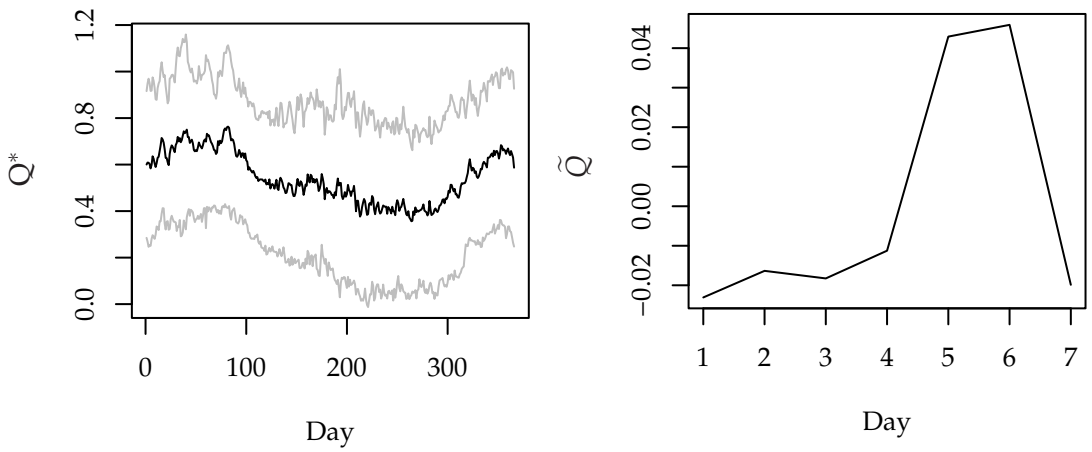

Figure D.6: Annual cycle of the mean (black) and the mean plus/minus one standard deviation (grey) for the Vilsbiburg run-off data (left). The mean weekly cycle is shown in the right plot. To facilitate the presentation, the standard deviation is not shown. Q^* denotes the Box-Cox transformed run-off and $\tilde{Q}$ is a dimensionless quantity.

the annual cycle, peaks in the periodogram of the run-off anomalies (Figure D.7, left) at frequencies $f = 1/7$ and $f = 2/7$ give evidence for a weekly component. Thus, we analogously estimate the mean and variance for this component (Figure D.6, right). Compared to the annual cycle the weekly cycle is rather small. Nevertheless, we subtract it for further analysis. Figure D.4 (right) shows that this substantially reduces the peaks in the periodogram.

Trend in the Mean

Subtraction of the weekly cycle leaves us with the series depicted in Figure D.8 (left). A ten-year running mean (red) suggests a slightly increasing tendency which we model with a straight line (green). Due to the autocorrelation in the series it is not straightforward to decide whether this deterministic trend component is significant. We subtract this trend and refer to the discussion on significance in Section D.3.1.

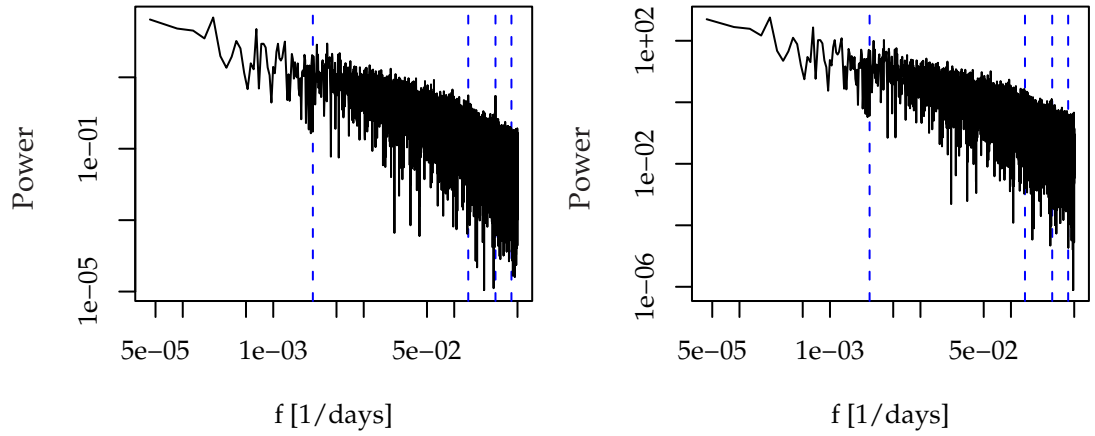

Figure D.7: Periodogram of the run-off anomalies (left) and the same record with the weekly periodic component removed (right). The dashed vertical lines mark the frequencies corresponding to an annual ($f = 1/365$) and a weekly periodicity ($f = 1/7, f = 3/7$).

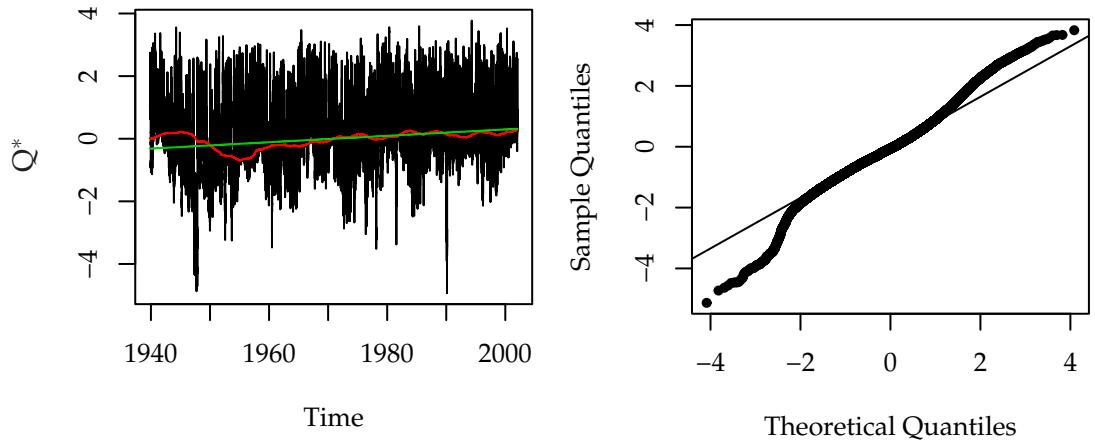

Figure D.8: Transformed run-off series for the gauge Vilsbiburg with the annual and weekly periodic components removed as time series (left) and compared to a Gaussian in a QQ-plot (right).

A QQ-plot of the residuals from the straight-line fit versus a Gaussian distribution reveals that the similarity of the frequency distribution to a Gaussian is large close to the centre but not in the tails, Figure D.8 (right).

D.3.4 Wisla Monthly Mean Run-Off at Tczew

Box-Cox Transformation

Figure 6.13 (right) reveals the asymmetric shape of the observed run-off distribution. We thus first use a Box-Cox transformation (Section D.2.2) with a ML estimate $\widehat{\lambda} = -0.180$ to obtain a more Gaussian-like shape which facilitates the further analysis.

Annual Cycle

For the transformed series, we estimate an annual periodic component for the mean and variance on the basis of the monthly data analogously to Section D.3.1. Figure D.9 (left) depicts one cycle of this component. We find the peak discharge in spring and low water

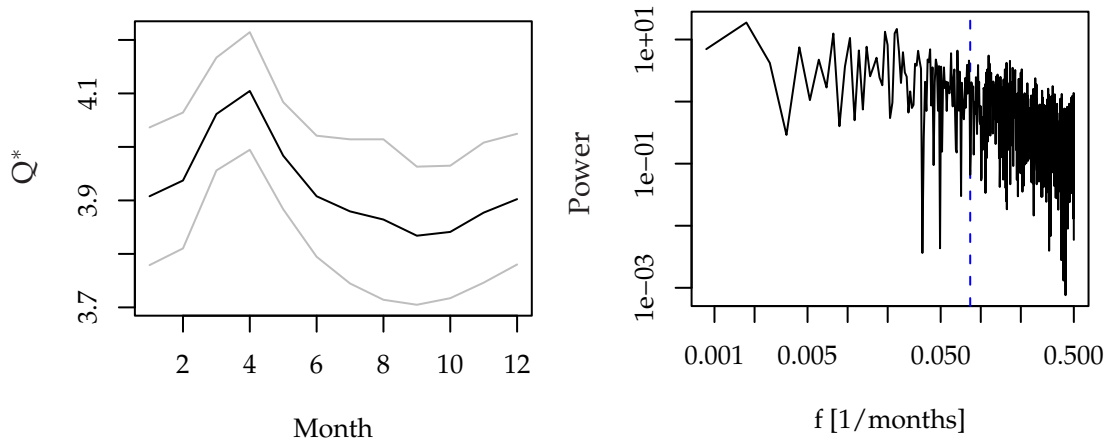

Figure D.9: Mean annual cycle for the Wisla run-off record (left), mean (black) plus/minus one standard deviation (grey). The right plot shows the periodogram of the Box-Cox transformed run-off record with the annual cycle in mean and variance removed The dotted vertical lines mark the frequency corresponding to an annual periodicity ($f = 1/365$). Q^* denotes the Box-Cox transformed run-off.

levels in autumn. The periodogram of the record with the annual cycle subtracted is depicted in Figure D.9 (right).

Trend in the Mean

This preprocessing leaves us with the series depicted in Figure D.10 (left). A ten-year running mean (red) visually suggests a slight increasing tendency which we model with a straight line (green). Least-squares estimation of the slope on the basis of the dimensionless run-off anomalies yields an increase of $2.35 \cdot 10^{-3}$ per year. In the same way as discussed for the previous analyses a significance test and confidence intervals for this trend is not straightforward as the series shows prominent autocorrelation. Under the assumption of uncorrelated observations, a t-test for the slope yields a p-value of $p = 0.031$ leading to significance on the 5%-level. With reference to discussion on significance in

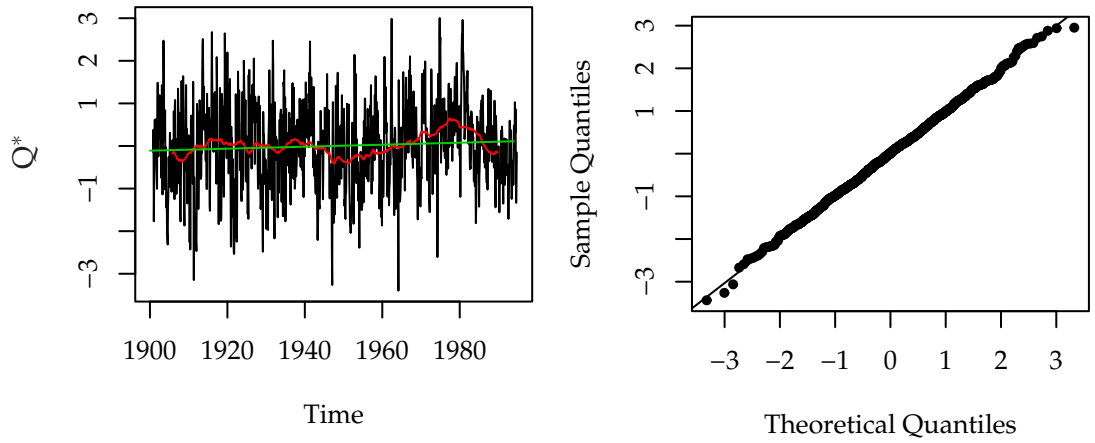

Figure D.10: Wisla run-off anomalies at gauge Tczew with the annual periodic component removed as time series (left) and compared to a Gaussian in a QQ-plot (right).

Section D.3.1, we subtract this trend. In the following, we denote the obtained transformed series with the seasonal and trend component removed as residual series. The QQ-plot of these residuals versus a Gaussian distribution shows that the transformed series is close to normal (Figure D.10, right).

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov and F. Csaki (Eds.), *2nd International Symposium on Information Theory*, Budapest, pp. 267–281. Akademie Kiade.

Apel, H., A. H. Thieken, B. Merz, and G. Blöschl (2004). Flood risk assessment and associated uncertainty. *Nat. Haz. Earth Syst. Sci. 4*, 295–308.

Avnir, D., O. Birham, D. Lindar, and O. Malcai (1998). Is the geometry of nature fractal? *Science 279*, 39.

Bardet, J.-M., G. Lang, G. Oppenheim, A. Philippe, and M. S. Taqqu (2003). Generators of long-range dependent processes: A survey. In P. Doukhan, G. Oppenheim, and M. S. Taqqu (Eds.), *Theory and Applications of Long-Range Dependence*, pp. 579–623. Boston: Birkhäuser.

Bárdossy, A. and S. Pakosch (2005). Wahrscheinlichkeiten extremer Hochwasser unter sich ändernden Klimaverhältnissen. *Wasserwirtschaft 7-8*, 58–62.

Becker, A. and U. Grünewald (2003). Flood risk in central Europe. *Science 300*(5622), 1099.

Beran, J. (1992). A goodness-of-fit test for time series with long-range dependence. *J. R. Statist. Soc. 54*(3), 749–760.

Beran, J. (1994). *Statistics for Long-Memory Processes*. Monographs on Statistics and Applied Probability. Chapman & Hall.

Beran, J. (1995). Maximum likelihood estimation of the differencing parameter for invertible short and long memory autoregressive integrated moving average models. *J. R. Statist. Soc. 57*(4), 659–672.

Beran, J., R. J. Bhansali, and D. Ocker (1998). On unified model selection for stationary and nonstationary short- and long-memory autoregressive processes. *Biometrika 85*(4), 921–934.

Beven, K. (Ed.) (1998). *Distributed Hydrological Modelling*. Wiley John & Sons.

Billingsley, P. (1995). *Probability and Measures*. Wiley Series in Probability and Statistics. New York: Wiley.

Bisaglia, L. (2002). Model selection for long-memory models. *Quaderni di Statistica 4*, 33–49.

Blender, R. and K. Fraederich (2003). Long time memory in global warming simulations. *Geophys. Res. Lett. 30*(14), 1769.

Bloomfield, P. (1992). Trends in global temperature. *Climatic Change*, 1–16.

Box, G. E. P. and D. R. Cox (1964). An analysis of transformations. *J. R. Statist. Soc. 62*(2), 211–252.

Box, G. E. P. and G. M. Jenkins (1976). *Time Series Analysis: forecasting and control*. New Jersey: Prentice Hall.

Brockwell, P. J. and R. A. Davis (1991). *Time series: Theory and Methods*. Springer Series in Statistics. Berlin: Springer.

Bunde, A., J. F. Eichner, S. Havlin, E. Koscielny-Bunde, H. J. Schellnhuber, and D. Vyushin (2004). Comment on "Scaling of atmosphere and ocean temperature correlations in observations and climate models". *Phys. Rev. Lett. 92*(3), 039801.

Bunde, A., J. F. Eichner, J. W. Kantelhard, and S. Havlin (2005). Long-term memory: A natural mechanism for the clustering of extreme events and anomalous residual times in climate records. *Phys. Rev. Lett. 94*(048701).

Bunde, A. and S. Havlin (Eds.) (1994). *Fractals in Science*. Berlin: Springer.

Bunde, A. and S. Havlin (2002). Power-law persistence in the atmosphere and in the oceans. *Physica A 314*(1-4), 15–24.

Bunde, A., S. Havlin, E. Koscielny-Bunde, and H.-J. Schellnhuber (2001). Long term persistence in the atmosphere: global laws and tests of climate models. *Physica A 302*, 255–267.

Byrd, R., P. Lu, J. Nocedal, and C. Zhu (1995). A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comp. 16*, 1190–1208.

Caballero, R., S. Jewson, and A. Brix (2002). Long memory in surface air temperature:detection, modeling, and application to weather derivative valuation. *Clim. Res. 21*, 127–140.

Cahalana, R. F., W. Ridgwaya, W. J. Wiscombea, T. L. Bella, and J. B. Snider (1994). The albedo of fractal stratocumulus clouds. *J. Atmos. Sci. 51*(16), 2434–2455.

Caputo, J. G., B. Malraison, and P. Atten (1986). Determination of attractor dimension and entropy for various flows: An experimentalist's viewpoint. In G. Mayer-Kress (Ed.), *Dimensions and Entropies in Chaotic Systems*, Volume 32 of *Springer Series in Synergetics*. Berlin: Springer Verlag.

Chen, Z., P. C. Ivanov, K. Hu, and H. E. Stanley (2002). Effects of nonstationarities on detrended fluctuation analysis. *Phys. Rev. E 65*, 041107.

Cohn, T. and H. Lins (2005). Nature's style: Naturally trendy. *Geophys. Res. Lett. 32*(L23402).

Coles, S. G. (2001). *An Introduction to Statistical Modelling of Extreme Values*. London: Springer.

Coles, S. G. and M. J. Dixon (1999). Likelihood-based inference for extreme value models. *Extremes 2*(1), 5–23.

Coles, S. G. and J. A. Tawn (1996). A Bayesian analysis of extreme rainfall data. *Appl. Stat. 45*, 463–478.

Coles, S. G. and J. A. Tawn (1999). Statistical methods for extreme values. Internet.

Cox, D. R. (1961). Tests of separate families of hypotheses. In *Proc. Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 105–123.

Cox, D. R. and D. V. Hinkley (1994). *Theoretical Statistics*. London: Chapman & Hall.

Craigmile, P. F., P. Guttorp, and D. B. Percival (2000). Wavelet-based parameter estimation for trend contaminated fractionally differenced processes. *J. Time Ser. Anal.*.

Craigmile, P. F., P. Guttorp, and D. B. Percival (2004). Trend assessment in a long memory dependence model using the discrete wavelet transform. *Environmetrics 15*, 313–335.

Dahlhaus, R. (1989). Efficient parameter estimation for self-similar processes. *Ann. Statist. 17*, 1749–1766.

Davidsen, J. and H. G. Schuster (2002). Simple model for $1/f^{\alpha}$ noise. *Phys. Rev. E 65*(026120).

Davison, A. C. and D. B. Hinkley (1997). *Bootstrap Methods and their Application*. Cambridge series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

DeGroot, M. H. (1975). *Probability and Statistics* (Second ed.). Series in Statistics. Reading: Addison-Wesley.

Denker, M. and G. Keller (1986). Rigorous statistical procedures for data from dynamical systems. *J. Stat. Phys 44*, 67–93.

Deo, R. and C. Hurvich (1998). Linear trend with fractionally integrated errors. *J. Time Ser. Anal. 19*, 379–397.

Doukhan, P., G. Oppenheim, and M. S. Taqqu (Eds.) (2003). *Theory and Application of Long-Range Dependence*. Boston: Birkhäuser.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Stat. 7*(1), 1–26.

Eichner, J. F., E. Koscielny-Bunde, A. Bunde, and H.-J. S. S. Havlin (2003). Power law persistence in the atmosphere: A detailed study of long temperature records. *Phys. Rev. E 68*.

Elek, P. and L. Márkus (2004). A long range dependent model with nonlinear innovations for simulating daily river flows. *Nat. Hazards Earth Syst. Sci. 4*, 227–283.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications* (third ed.). New York: John Wiley & Sons, Inc.

Fisher, R. A. and L. H. C. Tippett (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proc. Cambridge Phil. Soc. 24*, 180–190.

Fraedrich, K. and R. Blender (2003). Scaling of atmosphere and ocean temperature correlations in observations and climate models. *Phys. Rev. Lett. 90*, 108501.

Fraedrich, K. and R. Blender (2004). Reply to comment on "scaling of atmosphere and ocean temperature correlations in observations and climate models". *Phys. Rev. Lett. 92*(3), 039802–1.

Frisch, U. (1995). *Turbulence : The Legacy of A. N. Kolmogorov*. Cambridge University Press.

Geweke, J. and S. Porter-Hudak (1983). The estimation and application of long memory time series models. *J. Time Ser. Anal. 4*, 221–237.

Gil-Alana, L. (2005). Statistical modeling of the temperatures in the northern hemisphere using fractional integration techniques. *J. Clim. 18*, 5357–5369.

Giratis, L., P. Kokoszka, and R. Leipus (2001). Testing for long memory in the presence of a general trend. *J. Appl. Prob. 38*, 1033–1054.

Gnedenko, B. V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Ann. Math. (2) 44*, 423–453.

Govindan, R. B. and H. Kantz (2004). Long-term correlations and multifractality in surface wind speed. *Europhys. Lett. 68*(2), 184–190.

Govindan, R. B., D. Vyushin, A. Bunde, S. Brenner, S. Havlin, and H.-J. Schellnhuber (2002). Global climate models violate scaling of the observed atmospheric variability. *Phys. Rev. Lett. 89*, 028501.

Granger, C. W. and R. Joyeux (1980). An introduction to long-memory time series models and fractional differences. *J. Time Ser. Anal. 1*(1), 15–29.

Granger, C. W. J. (1980). Long memory relationships and the aggregation of dynamic models. *J. Econometrics 14*, 227–238.

Hannan, E. J. (1980). The estimation of the order of an ARMA process. *Ann. Statist. 8*(5), 1071–1081.

Hannan, E. J. and B. G. Quinn (1979). The determination of the order of an autoregression. *J. R. Statist. Soc. 41*, 190–195.

Hasselmann, K. (1997). Multi-pattern fingerprint method for detection and attribution of climate change. *Clim. Dyn. 13*, 601–611.

Hausdorff, J. M. and C.-K. Peng (1996). Multiscaled randomness: A possible source of $1/f$ noise in biology. *Phys. Rev. E 54*, 2154–2157.

Held, H. and T. Kleinen (2004). Detection of climate system bifurcations by degenerate fingerprinting. *Geophys. Res. Lett. 31*(L23207).

Henry, M. and P. M. Robinson (1996). Bandwidth choice in Gaussian semiparametric estimation of long-range dependence. In P. M. Robinson and M. Rosenblatt (Eds.), *Athens Conference on Applied Probability and Time Series Analysis*, Volume II: Time Series Analysis, New York, pp. 220–232. Springer.

Henry, M. and P. Zaffaroni (2003). The long-range dependence paradigm for macroeconomics and finance. In P. Doukhan, M. S. Taqqu, and G. Oppenheim (Eds.), *Theory and Applications of Long-Range Dependence*, pp. 417–438. Boston: Birkhäuser.

Hinde, J. (1992). Choosing between non-nested models: a simulation approach. In L. Fahrmeir, B. Fracis, R. Gilchrist, and G. Tutz (Eds.), *Advances in GLIM and Statistical Modelling*, Number 78 in Lecture Notes in Statistics. New York: Springer.

Hipel, K. and A. McLeod (1994). *Time Series Modelling of Water Resources and Environmental Systems*. Developement in Water Science. Amsterdam: Elsevier.

Hipel, K. W. and A. I. McLeod (1978). Preservation of the rescales adjusted range 2. Simulation studies using Box-Jenkins models. *Water Resour. Res. 14*(3), 509–516.

Hosking, J. R. M. (1981). Fractional differencing. *Biometrika 68*, 165–176.

Hosking, J. R. M. (1990). L-moments: analysis and estimation of distributions using linear combinations of order statistics. *J. R. Statist. Soc. B*, 105–124.

Hosking, J. R. M., J. R. Wallis, and E. F. Wood (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics 27*(3), 251–261.

Hu, K., P. C. Ivanov, Z. Chen, P. Carpena, and H. E. Stanley (2001). Effects of trends on detrended fluctuation analysis. *Phys. Rev. E 64*, 011114.

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Trans. Am. Soc. Civil Eng. 116*, 770–799.

Hurvich, C. M. and B. K. Ray (1995). Estimation of the memory parameter for nonstationary or nonivertible fractionally integrated processes. *J. Time Ser. Anal. 16*(17-41).

IPCC (2001). Climate change 2001: The scientific basis, contribution of working group I to the Third Assessment Report. Intergovernmental Panel on Climate Change (IPCC).

Jones, P. D. and A. Moberg (2003). Hemispheric and large-scale surface air temperature variations: An extensive revision and an update to 2001. *J. Clim. 16*(1), 206–223.

Jones, P. D., M. New, D. E. Parker, S. Martin, and I. G. Rigor (1999). Surface air temperature and its changes over the past 150 years. *Rew. Geophys. 37*(2).

Kallache, M. (2007). *Trends in the Mean and Extreme Values of Hydro-Meteorological Time Series*. Ph. D. thesis, University of Bayreuth, Bayreuth.

Kallache, M., H. W. Rust, and J. Kropp (2005). Trend assessment: Applications for hydrology and climate. *Nonlin. Proc. Geoph. 2*, 201–210.

Kantelhardt, J. W., E. Koscielny-Bunde, H. H. A. Rego, S. Havlin, and A. Bunde (2001). Detecting long-range correlations with detrended fluctuation analysis. *Physica A 295*, 441–454.

Kantz, H. and T. Schreiber (1995). Dimension estimates and physiological data. *Chaos 5*, 143–153.

Kaufmann, R. K. and D. I. Stern (1997). Evidences for human influence on climate from hemispheric temperature relations. *Nature 388*, 39–44.

Király, A. and I. Jánosi (2002). Stochastic modeling of daily temperature fluctuations. *Phys.Rev.E 65*(051102).

Klemes, V. (1974). The Hurst phenomenon: A puzzle? *Water Resour. Res. 10*(4), 675–688.

Kolmogorov, A. N. (1941). Local structures of turbulence in fluid for very large Reynolds numbers. In S. K. Friedländer and L. Topper (Eds.), *Transl. in Turbulence*, pp. 151–155. New York: Interscience Publisher. 1961.

Koscielny-Bunde, E., A. Bunde, S. Havlin, and Y. Goldreich (1996). Analysis of daily temperature fluctuations. *Physica A 231*, 393–396.

Koscielny-Bunde, E., A. Bunde, S. Havlin, H. E. Roman, Y. Goldreich, and H.-J. Schellnhuber (1998). Indication of a universal persistence law governing atmospheric variability. *Phys. Rev. Lett. 81*, 729–732.

Koscielny-Bunde, E., J. W. Kantelhardt, P. Braun, A. Bunde, and S. Havlin (2006). Long-term persistence and multifractality of river runoff records: Detrended fluctuation studies. *J. Hydrol. 322*, 120–137.

Koutsoyiannis, D. (2002). The Hurst phenomen and fractional Gaussian noise made easy. *Hydrol. Sci. J. 47*(4), 573–595.

Koutsoyiannis, D. (2003). Climate change, the Hurst phenomenon, and hydrological statistics. *Hydrol. Sci. J. 48*(1).

Krämer, W. and P. Sibbertsen (2002). Testing for sturctural changes in the presence of long memory. *Int. J. Buis. Econ. 1*(3), 235–242.

Kron, W. and T. Tumerer (2002). Water related disasters: Loss trends and possible counter measures from a (re-)insurance viewpoint. In *EU Concerted Action MITCH (Mitigation of climate-induced natural hazards), Workshop III, Floods, Droughts and Landslides – Who plans, who pays?*, Potsdam, pp. 549–558.

Kundzewicz, Z. and H.-J. Schellnhuber (2004). Floods in the IPCC TAR perspective. *Natural Hazards 31*, 111–128.

Künsch, H. R. (1986). Statistical aspects of self-similar processes. In *Proceedings of the First World Congress of the Bernoulli Society*, pp. 67–74. VNU Science Press.

Latif, M., D. Anderson, T. Barnett, M. Kane, R. Kleemann, A. Leetmaa, J. O'Brien, A. Rosati, and E. Schneider (1998). A review of the predictability and prediction of enso. *J. Geophys. Res. 103*(C7), 14375–14393.

Leadbetter, M. R., G. Lindgren, and H. Rootzén (1983). *Extremes and related properties of random sequences and processes.* Springer Series in Statistics. New York: Springer.

Li, W. K. (2004). *Diagnostic Checks in Time Series*, Volume 102 of *Monographs on Statistics and Applied Probability*. Boca Raton: Chapman & Hall/CRC.

Lillo, F., S. Micciche', and R. N. Mantegna (2002). Long-range correlated stationary Markovian processes. *Phys. Rev. Lett.*. submitted, also cond-mat/0203442.

Lo, A. W. (1991). Long-term memory in stock market prices. *Econometrica 59*, 1279–1313.

Lohre, M., P. Sibbertsen, and T. Könnig (2003). Modeling water flow of the Rhine River using seasonal long memory. *Water Resour. Res. 39*(5), 1132.

Maheu, J. M. (2005). Can GARCH models capture the long-range dependence in financial market volatility? *Studies Nonlin. Dyn. and Econometrics 9*(4).

Mallows, C. (1973). Some comments on c_p. *Technometrics 15*, 661–675.

Mandelbrot, B. (1972). Statistical methodology for non-periodic cycles: from the covariance to R/S analysis. *Ann. Econ. Soc. Meas. 1*(3), 259–290.

Mandelbrot, B. (1975). Limit theorems on the self-normalized range for weakly and strongly dependent processes. *Z. Wahr. verw. Geb. 31*, 271–285.

Mandelbrot, B. B. and J. W. van Ness (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Rev. 10*, 422–437.

Mandelbrot, B. B. and J. R. Wallis (1968a). Noah, Joseph and operational hydrology. *Water Resour. Res. 4*(5), 909–918.

Mandelbrot, B. B. and J. R. Wallis (1968b). Robustness of the rescaled range R/S and measurement of non-cyclic long-run statistical dependence. *Water Resour. Res. 5*, 967–988.

Maraun, D., H. W. Rust, and J. Timmer (2004). Tempting long-memory - on the interpretation of DFA results. *Nonlin. Proc. Geophys. 11*, 495–503.

Merriam-Webster (2005). Merriam-webster online dictionary. `http://www.m-w.com/dictionary`.

Merz, B. (2006). *Hochwasserrisiken*. Stuttgart: E. Schweizerbart'sche Verlagsbuchhandlung.

Merz, B. and A. Thieken (2005). Separating natural and epistemic uncertainty in flood frequency analysis. *J. Hyd. 309*, 114–132.

Metzler, R. (2003). Comment on "power-law corelations in the southern-oscillation-index fluctuations characterizing El Niño". *Phys. Rev. E 67*(018201).

Milhoj, A. (1981). A test of fit in time series series models. *Biometrika 68*, 177–188.

Monetti, R. A., S. Havlin, and A. Bunde (2001). Long term persistence in the sea surface temperature fluctuations. *Physica A 320*, 581–589.

Montanari, A., R. Rosso, and M. S. Taqqu (1997). Fractionally differenced ARIMA models applied to hydrologic time series: Identification, estimation and simulation. *Water Resour. Res. 33*(5), 1035.

Montanari, A., R. Rosso, and M. S. Taqqu (2000). A seasonal fractional ARIMA model applied to the nile river monthly flows. *Water Resour. Res. 36*(5), 1249–1259.

Münchner Rück (2003). *Jahresrückblick Naturkatastrophen* (10. Jahrgang ed.). topics. München: Münchner Rückversicherungsgesellschaft.

Nelder, J. A. and R. Mead (1965). A simplex algorithm for function minimization. *Computer J. 7*, 308–313.

Oke, T. R. (1973). City size and the urban heat island. *Atmos. Env. 7*, 769 – 779.

Okubo, P. G. and K. Aki (1987). Fractal geometry in the San Andreas fault system. *J. Geophys. Res. 92*(B1), 345–356.

Orrel, D., L. Smith, J. Barkmeijer, and T. Palmer (2001). Model error in wheather forecasting. *Nonlin. Proc. Geophys. 8*, 357–371.

Peng, C.-K., S. V. Buldyrev, A. L. Goldberger, S. Havlin, M. Simons, and H. E. Stanley (1993). Finite-size effects on long-range correlations: Implications for analyzing DNA sequences. *Phys. Rev. E 47*(5), 3730–3733.

Peng, C.-K., S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger (1994). Mosaic organization of DNA nucleotides. *Phys. Rev. E 49*(2), 1685.

Peng, C.-K., S. Havlin, H. E. Stanley, and A. L. Goldberger (1995). Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos 5*(1), 82–87.

Percival, D. B. and A. T. Walden (2000). *Wavelet Methods for Time Series Analysis*. Camebridge University Press.

Press, W. H., S. A. Teukolsky, W. T. Vetter, and B. P. Flannery (1992). *Numerical Recipes in C* (second ed.). Cambridge: Cambridge University Press.

Priestley, M. B. (1992). *Spectral Analysis and Time Series*. London: Academic Press.

R Development Core Team (2004). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rahmstorf, S. and H.-J. Schellnhuber (2006). *Der Klimawandel*. München: C. H. Beck.

Rebetez, M., H. Meyer, O. Dupont, D. Schindler, K. G. J. Kropp, and A. Menzel (2006). Heat and drought 2003 in Europe: a climate synthesis. *Ann. Forest Sci. 63*, 569–577.

Renner, C., J. Peinke, R. Friedrich, O. Chanal, and B. Chabau (2002). Universality of small scale turbulence. *Phys. Rev. Lett. 89*(124502).

Ripley, B. D. (2004). *Methods and Models in Statistics*, Chapter Selecting Amongst Large Classes of Models, pp. 155–170. London: Imperial College Press.

Robinson, P. (2005). Efficiency improvements in inference on stationary and nonstationary fractional time series. *Ann. Statist. 3*(4), 1800–1842.

Robinson, P. M. (1995a). Gaussian semiparametric estimation of long-range dependece. *Ann. Statist. 23*(5), 1630–1661.

Robinson, P. M. (1995b). Log-periodogram regression of time-series with long-range dependence. *Ann. Statist. 23*(3), 1048–1072.

Robinson, P. M. (Ed.) (2003). *Time Series with Long Memory*. Advanced Texts in Econometrics. Oxford University Press.

Romano, M., M. Thiel, U. Schwarz, J. Kurths, H. Lange, and M. Haus (2001). Conceptual model of runoff from a forested catchment. *Int. J. Bif. Chaos 11*(10), 2567.

Rust, H. W., M. Kallache, H.-J. Schellnhuber, and J. Kropp (2006). Confidence intervals for flood return level estimates using a bootstrap approach. *Preprint*.

Rybski, D., A. Bunde, S. Havlin, and H. von Storch (2006). Long-term persistence in the climate and the detection problem. *Geophy. Res. Lett. 33*(L06718).

Salas, J. D. and R. A. Smith (1981). Physical basis of stochastic models of annual flow. *Water Res. Research. 17*(2), 428–430.

Santer, B. D., W. Brüggemann, U. Cubasch, K. Hasselmann, H. Höck, E. Maier-Reimer, and U. Mikolajewicz (1994). Signal-to-noise analysis of time-dependent greenhouse warming experiments. *Clim. Dyn. 9*, 267–285.

Schreiber, T. and A. Schmitz (1996). Improved surrogate data for nonlinearity tests. *Phys. Rev. Lett. 77*, 635.

Schreiber, T. and A. Schmitz (2000). Surrogate time series. *Physica D 142*, 346–382.

Schumway, R. H. and D. S. Stoffer (2006). *Time Series Analysis and its Application* (second ed.). New York: Springer.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist. 6*, 461–464.

Sibbertsen, P. (1999). *Robuste Parameterschätzung im linearen Regressionsmodell bei Fehlertermen mit langem Gedächtnis*. Ph. D. thesis, Universtity of Dortmund.

Sibbertsen, P. (2001). Long-memory versus structural breaks: An overview. SFB 475 28, University of Dortmund, Dortmund.

Smith, H. F. (1938). An empirical law describing heterogeneity in the yields of acricultural crops. *J. Agric. Sci. 28*, 1–23.

Smith, J., N. Taylor, and S. Yadav (1997). Comparing the bias and misspecification in ARFIMA models. *J. Time Ser. Analysis 18*(5), 507–527.

Smith, R. (2003). Statistics of extremes,with applications in environment, insurance and finance.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika 72*(1), 67–90.

Smith, R. L. (1987). Comparison of maximum likelihood and Bayesian estimators for the three-parameter Weibull distribution. *Appl.Stat. 36*(3), 358–369.

Smith, R. L. (1993). Long-range dependence and global warming. *Stat. Env.*.

Smith, R. L. (2001). Environmental statistics. Internet. Version 5.0.

Smith, R. L., T. M. L. Wigley, and B. D. Santer (2003). A bivariate time series approach to anthropogenic trend detection in hemispheric mean temperatures. *J. Clim. 16*, 1228–1240.

Stephenson, A. and C. Ferro (2004). *evd: Functions for extreme value distributions*. R package version 2.1-6.

Talkner, P. and R. O. Weber (2000). Power spectrum and detrended fluctuation analysis: Application to daily temperatures. *Phys. Rev. E 62*(1), 150–160.

Taqqu, M. S. and V. Teverovsky (1997). Robustness of whittle-type estimators for time series with long-range dependence. *Stoch. Mod. 13*, 723–757.

Taqqu, M. S., V. Teverovsky, and W. Willinger (1995). Estimators for long-range dependence: An empirical study. *Fractals 3*(4), 785–798.

Teverovsky, V. and M. S. Taqqu (1997). Testing for long-range dependence in the presence of shifting means or a slowly declining trend. *J. Time Ser. Anal. 18*, 279–304.

The, Y.-K., J. Fernandez, O. Popa, H. Lerche, and J. Timmer (2005). Estimating rate constants from single ion channel currents when the initial distribution is known. *Eur. Biophys. J. 34*, 306–313.

Theiler, J., S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer (1992). Testing for nonlinearity in time series: The method of surrogate data. *Physica D 58*(77).

Timmer, J. (1998). Modeling noisy time series: Physiological tremor. *Int. J. Bif. Chaos 8*(7).

Timmer, J., S. Haeussler, M. Lauk, and C. H. Luecking (2000). Pathological tremors: Deterministic chaos or nonlinear stochastic oscillators? *Chaos 10*, 278–288.

Timmer, J. and M. König (1995). On generating power law noise. *A&A 300*(3), 707–710.

Timmer, J., T. G. Müller, I. Swameye, O. Sandra, and U. Klingmüller (2004). Modelling the nonlinear dynamics of cellular signal transduction. *Int. J. Bif. Chaos 14*(6), 2069–2079.

Timmer, J., U. Schwarz, H. Voss, I. Wardinski, T. Belloni, G. Hasinger, M. van der Klis, and J. Kurths (2000). Linear and nonlinear time series analysis of the black hole candidate cygnus x-1. *Phys. Rev. E 61*(2).

Tsonis, A. A. and J. B. Elsner (1995). Testing for scaling in natural forms and observables. *J. Stat. Phys. 81*(5/6), 869–880.

UNDP (2004). Reducing disaster risk: A challenge for development. UNDP (United Nations Developments Programme) Bureau for Crisis Prevention and Recovery (www.undp.org/bcpr/disred/rdr.htm).

Velasco, C. (2000). Non-Gaussian log-periodogram regression. *Econometric Theory 16*, 44–79.

Velasco, V. and P. Robinson (2000). Whittle pseudo-maximum likelihood estimation for nonstationary time series. *J. Am. Stat. Assoc. 95*, 1229–1243.

Venema, V., S. Bachner, H. W. Rust, and C. Simmer (2006). Statistical characteristics of surrogate data based on geophysical measurements. *Nonlin. Proc. Geophys.*, 449–466.

Vyushin, D., I. Zhidkov, S. Havlin, A. Bunde, and S. Brenner (2004). Volcanic forcing improves atmosphere-ocean coupled general circulation model scaling performance. *Geophys. Res. Lett. 31*(L10206).

Vyushin, D. I., V. E. Fioletov, and T. G. Shepherd (2007). Impact of long-range correlations on trend detection in total ozone. *J. Geophys. Res.*. submitted.

Weron, R. (2002). Estimating long-range dependence: finite sample properties and confidence intervals. *Physica A 312*, 285–299.

Whittle, P. (1953). Estimation and information in stationary time series. *Ark. Mat. 2*(423-434).

Wigley, T. M. L. and P. D. Jones (1981). Detecting CO2-induced climatic change. *Nature 292*, 205 – 208.

Williams, D. A. (1970). Discriminating between regression models to determine the pattern of enzyme synthesis in synchronous cell cultures. *Biometrika 28*, 23–32.

WMO/UNEP (2002). *IPCC Workshop on Changes in Extreme Weather and Climate Events*, Beijing, China. WMO/UNEP.

Woodward, W. A. and H. L. Gray (1995). Selecting a model for detecting the presence of a trend. *J. Clim. 8*, 1929–1937.

Wornell, G. W. (1993). Wavelet-based representations for the $1/f$ family of fractal processes. *Proceedings of IEEE 81*(10).

Yajima, Y. (1985). On estimation of long-memory time series models. *Austr. J. Statist. 27*, 303–320.