Pontificia Universidad
JAVERIANA
Bogotá

STOCHASTIC MODELING OF HYDROMETEOROLOGICAL EXTREMES AND THEIR POSSIBLE RELATION WITH GLOBAL CHANGE

A Master Thesis Presented by

THOMAS ROSMANN

Directed by

EFRAÍN A. DOMÍNGUEZ C.

Submitted to the *Maestría en Hidrosistemas* of the

Pontificia Universidad Javeriana in fulfillment

of the requirements for the degree of

MAGISTER EN HIDROSISTEMAS

June 2014

ACKNOWLEDGEMENTS

At the moment of completion of this document, I would like to take the time to express my gratitude to some of the people who helped me make all this possible.

First and foremost, I want to thank my tutor Efraín Domínguez, whose everlasting patience and compassion were some of the driving forces of this investigation. Not only his motivation, but also his constant availability at any moment made it possible to eliminate the doubts I had by answering all my questions and always giving me new ideas, which is why this work resulted the way it did. Elaborating this thesis with his guidance I learned a lot, not only from the technical side but also human quality.

Also, I would like to thank Andrés Torres for facilitating me to enter the masters program and overcome all the initial problems I had. It was also due to his help and to a big part to Nelson Obregon's generous job offer at the Geophysics Institute that I could continue my studies after the first semester.

I would like to thank all of my colleagues in the masters program, especially Hugo and David for initially including me into the group of students, always being there to help and for their good company. Also, Natalia, Alejandra, Jorge and Felipe for accepting me for who I am and spending a lot of great moments with me. I am lucky to say that I found such good friends and the time I spent in the masters program was one that I will always remember with a smile.

Thanks to all the other faculty staff and professors that shared their knowledge and supported me when I had needed it with an almost completely new topic, a new culture and language, and helped me develop myself thanks to the projects in which I was allowed to participate.

Finally, I would like to thank Sandra for supporting me all this time, for always offering help and good advice to continue working hard. Also her family who opened their doors to me, made me feel welcome and supported me whenever I needed it.

For the final version of the document, I would like to extend the acknowledgements section to express my gratitude to the evaluators of this work, Demetris Koutsoyiannis, Anil Markandya, Marc Neumann, Jorge Escobar and Juan Diego Giraldo, for their meticulous revision and productive comments.

AGRADECIMIENTOS

En el momento de terminar este documento quiero tomar el tiempo para expresar mi gratitud a algunas de las personas que ayudaron durante el camino para llegar a este punto.

Primero de todo, quiero agradecerle a mi director Efraín Domínguez por su infinita paciencia conmigo y su compasión, las cuales fueron unas de las fuerzas más importantes que ayudaron a seguir con el trabajo. Se debe a su motivación y constante disponibilidad en cualquier momento para aclarar todas mis dudas y darme consejos y nuevas ideas que el trabajo resultó de la forma como está. Elaborando éste trabajo de grado con él aprendí mucho, no solamente del lado técnico sino también de la calidad humana.

También quiero decirle gracias al ingeniero Andrés Torres, quien fue la persona que me hizo posible entrar a la Maestría de Hidrosistemas a pesar de todos los problemas que se me presentaron al principio. También se debe a él y por gran parte a la oferta de trabajo generosa del profesor Nelson Obregón en el Instituto Geofísico que pude seguir con la carrera después del primer semestre.

Quiero agradecerles a todos mis compañeros de la maestría, especialmente a Hugo y David por introducirme inicialmente al grupo de estudiantes, por siempre estar a la orden para ayudar en todo cuando lo requería y por su buena compañía. También gracias a Natalia, Alejandra, Jorge y Felipe por aceptarme como soy y pasar tanto tantos buenos momentos juntos. Puedo considerarme afortunado de haber conocido amigos tan buenos que fueron una de las razones principales por las cuales siempre voy a recordarme del tiempo de la maestría con una sonrisa.

Gracias a todos los miembros de la facultad y los profesores que compartieron su conocimiento y me apoyaron en tiempos en que se me hizo difícil entender un tema casi completamente desconocido, una nueva cultura y un idioma diferente, y que ofrecieron a desarrollarme en los proyectos en los cuales tuve la oportunidad de participar.

Finalmente, quiero agradecerle a Sandra por siempre apoyarme en todo y ofrecer su ayuda y buenos consejos para seguir trabajando duro. También a su familia que me abrió las puertas, me hicieron sentir bienvenido y me brindaron su apoyo siempre que lo necesitaba.

Para la versión final de este documento, me gustaría extender la parte de agradecimientos para expresar mi gratitud a los evaluadores de este trabajo, Demetris Koutsoyiannis, Anil Markandya, Marc Neumann, Jorge Escobar y Juan Diego Giraldo por su revisión minuciosa y sus comentarios productivos.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

SECTION 1

INTRODUCTION

In recent years and decades, the topic of global climate change has become one of the most controversial and heavily discussed. By now, scientists have found sufficient proofs that the global climate has experienced abrupt changes in the last decades, since the mid-19th century, but with the strongest increase from the 1950s onwards (IPCC, 2013). The most obvious characteristic is global warming, which most certainly influences on other variables as well, which was proven for some areas in previous studies (Wang et al., 2009; Woo et al., 2008).

The moments when the existence of global change calls the attention of people most drastically is when emergencies occur as a result of it. For a large number of events in the recent past, news about human tragedies as results of climatic phenomena have reached the public, such as for example hurricane Katrina in New Orleans in 2005, typhoon Haiyan on the Philippines in 2013, severe floods in Australia and large parts of central Europe in 2011 and 2013, respectively, an extreme cold wave in North America at the beginning of 2014 or a large number of extensive droughts in India, for example in 2013, and eastern Africa, especially in the years of 2008 to 2009 and 2010 to 2011, just to name a few. In all these cases, a connection to global change was established in the media and discussed in public and politics.

In its recently released Fifth Assessment Report (IPCC, 2013), the Intergovernmental Panel on Climate Change (IPCC) mentions that it can confidently be stated that temperature has been rising steadily since the 1950s and that the change in precipitation over the global land mass is characterized as being of medium strength. However, it is only worded as *likely* that changes in extreme events have occurred since the beginning of observation in about 1950. Among these extreme events are increases of heat waves on a global scale and heavy rainfall events in North America and Europe.

Another recently published report, the United Nations' World Water Development Report 2014 (UNESCO, 2014), predicts an increase in worldwide water consumption of about 55% until 2050, with which energy and alimentation demand are closely related. Just this one fact states the importance to study profoundly the state of water resources and any impacts that influence on them. One of these influences is the occurrence of extreme events, which can cause a variety of problems in water supply, including shortages, contamination or damages to infrastructure, among others.

The behavior of extreme events in river discharge series has been studied in a number of investigations using various of different methods, among which trend analysis is the most frequent one. Both linear and nonlinear trends have been applied, as well as flood frequency analysis. Results of these studies in many cases indicate intensifications of the extreme events, but also state that the period that reliable hydrometeorological data is too short to prove the existence of changes (Bordi et al., 2009). A more detailed study of previous works that investigated the topic of trends in hydrometeorological time series will be given in section 3. Other studies treated with the topic of occurrence of extreme events, such as floods, which seemed to have increased over the last years (Kundzewicz et al., 2013), or relate extreme events or trends in hydrometeorological time series to the level of CO_2 emissions or microclimatological indices (Hirsch and Ryberg, 2012; Moreno, 2011), but do not always succeed to prove the impact of these indicators.

In order to prevent harm to persons and impacts on structures and ecosystems, various measures on the administrative level have been taken. One of many examples is the Floods Directive of the European Parliament (European Parliament, 2007), which has been applied to national laws of the member states of the European Union and triggered a large number of research projects related to the topic.

But not only on the administrative level, also in many other fields of economy or research, the topic of extreme events causes an increased interest. Needless to say, the impact of these events are crucial for the work of insurance companies (Spekkers et al., 2013), but also in many other diverse fields research has been conducted, for example for the design of urban drainage systems (Smith et al., 2002), the risk of extinction of species due to extreme events (Colomer et al., 2014) or the evaluation of irrigation pricing during drought periods (Nikouei and Ward, 2013), just to name a few.

One of the crucial questions that is asked in many of the studies is if the behavior of extreme events changes and how humans can adapt to that change. To answer this question, it is not enough to state that events are changing, but also how they are working and which influences cause the change. A more profound understanding of the statistical characteristics of the time series under investigation can be achieved by studying stochastically the random processes that define them (Maldonado, 2009). Stochastic models have been applied successfully in hydrological applications, which were applied to different fields of hydrological studies other than the study of extreme events. Frolov (2006) constructed various dynamic stochastic models to describe the long-term variations in the mean annual discharges of the Volga River as do Dolgonosov and Korchagin (2007) to describe the runoff dynamics in the Moskva and Volga River basins. Domínguez and Rivera (2010) propose a stochastic model using the Fokker-Planck-Kolmogorov equation for predicting the monthly effluent to the Betania hydropower plant in the upper Magdalena basin and Moreno (2011) also uses stochastic model to evaluate the hydrological forcing in Colombia. Naidenov proposed a physical explanation of probabilistic characteristics of hydrological processes (Naidenov and Podsechin, 1992; Naidenov and Shveikina, 2005, 2002), as well as Koutsoyiannis et al. (2008) in the Nile River basin, while Kovalenko developed a new modeling framework heavily based on the concepts of the theory of stochastic processes for the simulation and forecasting of complex systems (Kovalenko, 2012, 1986; Kovalenko et al., 1993). All of these previous works show the ability of stochastic models to relate the probabilistic characteristics of hydrological processes with the system input signal and physio-geographic and other characteristics of the watersheds, in which they are located.

According to the above mentioned, it is necessary to demonstrate if the frequency of hydrometeorologically extreme phenomena has intensified or changed otherwise in the last years and model stochastically, in particular on a daily level, the precipitation-runoff relation with the purpose of understanding the possible mechanisms of alteration of the structure of stochastic processes that lead to an intensification of the hydrometeorological extreme events with special attention to the consequences that global change could have on the local hydrological processes. This understanding might lead to new practices in watershed management, prevention and mitigation of extreme events adequate to the process of the global change that is happening to the planet.

This study can be seen as a first step in the process towards a profound understanding of the stochastic characteristics describing the regimes of extreme events in discharge time series. It is divided into 4 sections: Section 2 describes the concepts that will be used, as well as the data and computational tools used and created for the purpose of this investigation and that resulted from the study of literature. Section 3 describes the methodology and results of a global study of trends in hydrometeorological time series to corroborate the existence of changes in the global climate system and try to answer the question if changes in hydrometeorological extremes can be sufficiently be described by them, or if a more profound modeling method is needed. Section 4 thoroughly explains the proposed methodology of a stochastic model describing the evolution of extreme discharges in time with an inverse modeling approach. From the results of the model's application in 4 test basins presented in section 5, the proposed deterministic kernel of the processes is validated and its parameters related to the physical properties of the watershed, as well as external influences. Finally, the model is applied to evaluate if alterations in the parameters change the probabilistic regime of the process and its results will be presented and discussed.

The schematic outline of the document is presented in Figure 1, where the processes 2 to 5 will be explained later in the document in detail. In these sections, further schematic descriptions will be provided for the respective topics.

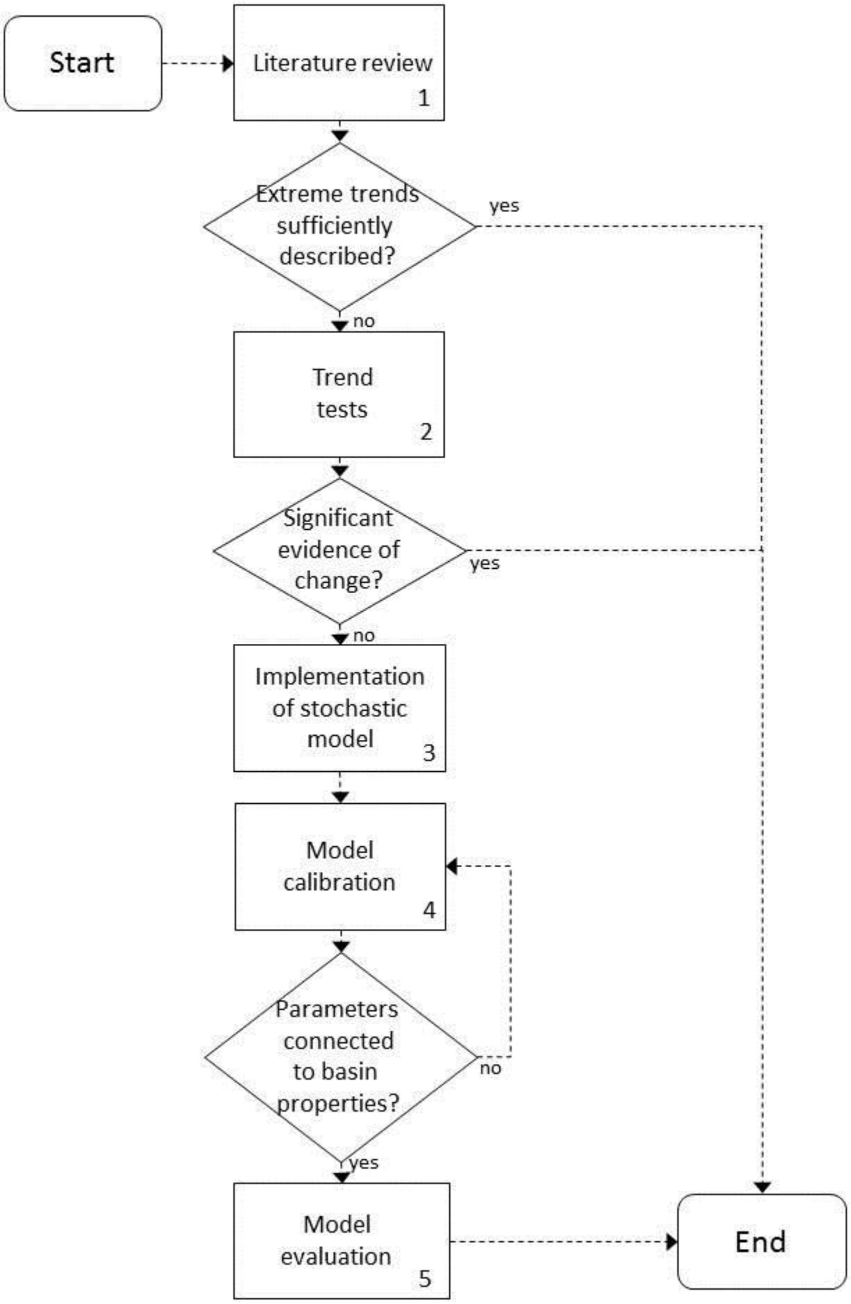
Start
Literature review
1
Extreme trends sufficiently described?
yes
no
Trend tests
2
Significant evidence of change?
yes
no
Implementation of stochastic model
3
Model calibration
4
Parameters connected to basin properties?
no
yes
Model evaluation
5
End

Figure 1. Schematic outline of the proposed work

SECTION 2

CONCEPTUAL DOMAIN

2.1 Fundamentals concepts

This chapter describes the main concepts used in this work and follows the descriptions in Koutsoyiannis (2008), Coles (2001), Gardiner (2004) and Sveshnikov (1966).

2.1.1 Probability

Although many people see probability as a mere branch of mathematics, which provides tools for data analysis, it is actually a more general concept that helps describe and shape a different view of the world, especially in the study of complex systems. In the course of history, scientific views were predominantly deterministic, which left no space for doubts and a law was generally seen as almost absolutely true. The notions of errors or uncertainty in scientific works were hardly considered. Through time, the concept of indeterminism was created and grew more widely accepted, a concept that allowed the existence of distinct outcomes for a problem, given the same initial conditions, which were more or less probable to occur. Although it is nowadays accepted to include the concept of uncertainty and probability, the nature of those concepts in the response of complex systems is still discussed (Koutsoyiannis, 2008; Maldonado, 2009).

Deterministic solutions are valid and good tools for mathematical problems on a microscopic scale, where it is likely to only observe a few objects that need to be described or modeled. However, for problems on a macroscopic scale, it is not so easy to describe them with a deterministic model anymore, because there are many different objects that might not all behave in one given way.

Hydrological processes are complex systems and therefore have to be modeled on a macroscopic level. Describing each object present in a hydrological system would not be possible due to many reasons, such as for example operational limitations or the fact that it is not necessary to describe every aspect of the system in detail (Koutsoyiannis, 2008).

2.1.2 The Axiomatization of Kolmogorov

Many scientific works have described probability and all of its theory. This research is based on the axiomatization presented in 1931 by the soviet mathematician and hydrologist Andrey Nikolaevich Kolmogorov. It is based on three fundamental concepts and three main axioms, which can be described as principles of a theory that are not derived or deducted in the same system.

The base of the axiomatization is the probability space, which is made up of the three main concepts:

a) The *sample space* Ω is a non-empty set that includes the known outcomes ω.
b) A *Sigma-Algebra* (or σ-algebra) Σ, which is a set of all possible subsets of Ω, called events and described as *E*. Based on ordinary set theory, Ω itself and the empty set *Ø* are both subsets contained in Σ, additionally to the other subsets as are the complements and all possible unions of subsets.
c) The *probability function P* assigns each member of Σ a number between 0 and 1, which is equal to the probability of occurrence of the event.

Additionally, the three main axioms describe the properties of *P*:

i. Every event *E* has a probability $P(E) \geq 0$.
ii. The probability of Ω, $P(\Omega) = 1$.
iii. For any incompatible events *A* and *B* ($AB = \emptyset$), $P(A + B) = P(A) + P(B)$

A fourth axiom describes the continuity at zero of decreasing sequences of events and follows from the first three axioms if Σ is finite (Koutsoyiannis, 2008).

2.1.3 Random variables

Random variables are one example of a simple realization of the probability space described by Kolmogorov and can be seen as a function $f(\omega)$ that assigns a number to each possible outcome $\omega \epsilon \Omega$. Following the representation in Koutsoyiannis (2008), random variables will be presented as an underlined lowercase letter, $\underline{x}(\omega)$. It is important to have in mind that a random variable describes the outcome of an experiment, such as for example the average temperature measured in January at a given climate station, which is not a single value, but a function that represents the values of all possible outcomes the experiment can take. These values, the realizations of random variables, are henceforth denoted as non-underlined lower-case letters, which are equal to the letter that denotes the random variable.

Random variables can be fully described by their probability distribution. The distribution function of a variable $\underline{x}$ is defined as

$$F(x) = P(\underline{x} \leq x), \tag{1}$$

which can be described as the probability of non-exceedance of the random variable $\underline{x}$ taking a value *x*. Therefore, *F(x)* is a non-decreasing function, which is also referred to as cumulative distribution function. Its counterpart, *1 – F(x)*, which is used in many hydrological applications, is the function describing the probability of exceedance of *x*, hence is a non-increasing function.

$$1 - F(x) = P(\underline{x} > x) \tag{2}$$

The derivative of the distribution function, *f(x)*, the probability density function (or PDF), describes the concentration of exceedance probability of the random variable $\underline{x}$ in a given interval *dx*. It can be related to probability distribution function *F(x)* as

$$f(x) = \frac{dF(x)}{dx} \tag{3}$$

From this follows that the integral under the complete function accounts to a probability of 100%, therefore

$$\int_{-\infty}^{\infty} f(x)dx = 1 \tag{4}$$

When comparing two or more random variables, the concepts of joint and conditional probability are of fundamental importance.

Joint probability describes the probability that an event ω occurs in both random variables. It can be expressed as in Gardiner (2004)

$$P(\underline{a} \cap \underline{b}) = P\{(\omega \epsilon \underline{a})\ and\ (\omega \epsilon \underline{b})\} \tag{5}$$

This concept is especially important when more than one time is considered, for example when the different random variables represent the same measurement at two different times. Joint probability density functions are *n*-dimensional, depending on the number *n* of random variables considered, which can be easily illustrated if two variables are considered, but becomes more complex for more variables.

Conditional probability is the probability of an event occurring in one random variable given that another event has occurred in another variable. Kolmogorov defines the conditional probability of the event *A* within the sigma-algebra and a probability of the event *B* with $P(B) > 0$, then the conditional probability of A is the quotient of the joint probability and the probability of B, or

$$P(\underline{a}|\underline{b}) = \frac{P(\underline{a} \cap \underline{b})}{P(\underline{b})} \tag{6}$$

Therefore, the probability distribution of one random variable given the distribution of another can be described by an equation that represents the conservation law of a probability density current.

2.1.4 Random processes

Taking into account that random variables represent the possible realizations of a statistical event and their probability of occurrence, random processes can be seen as a set of more than one random variable. A random process is a system $\underline{X}(t)$, in which a time-dependent random variable $\underline{x}(t)$ exists, where the values of $\underline{x}(t)$ are measured at different times t_1, t_2, …, t_n and a set of joint probability densities is given that describes the system completely (Gardiner, 2004). In other words, it is a function, whose values for each time step *t* is a random variable, and which therefore indexes a set of random variables in time. The time argument *t* can assume any value in a given interval. It has to be mentioned that the argument *t* does not automatically represent time, but in the majority of applications, as well as in the present study, it does (Sveshnikov, 1966). Random processes will be denoted as underlined, upper-case letters.

To construct a random process from observed data, the method described by Kolmogorov and presented in Sveshnikov (1966) was applied: Each random process consists of a number of realizations, which are the results of the measurements of a variable of independent experiments. These independent experiments can be the measure of a daily mean at a hydrological station measuring discharge values for each day during a year. This experiment can be repeated for *n* years to obtain a number of *n* realizations. Each realization can be drawn as a curve, connecting all the measured values. Superimposing these curves shows the bundle that reflects the ensemble of observed realizations of the random process.

Since each of the realizations is repeated for the same period of time – in the above mentioned example each day of a complete year – it is possible to consider all values measured on the same day of the year in all of the independent experiments to form a random variable taking the value of the random process at the instant of time *t*. This random variable can be completely described by its probability density function. The random variables of a random process do not necessarily have to be independent from each other, and were not in any occasion in this present study.

For reasons of a better visualization, the above mentioned example is reduced to each realization of the process containing only 12 values, representing the monthly mean values of discharge values. Figure 2 shows the bundle of realizations and the probability density functions of the 12 random variables.

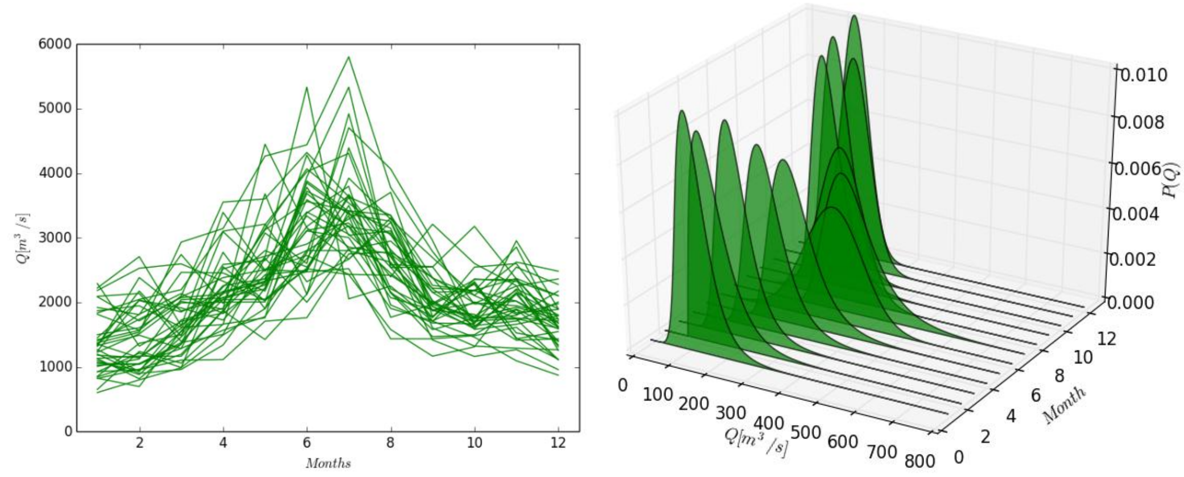

Figure 2. Bundle of realizations and probability density functions of a random process: monthly mean discharges, Pte. Balseadero station, Colombia

As mentioned before, to define completely a random process, the joint probability density function of the probability densities of all instances of time in the given time interval, $t_1, t_2, \dots, t_n$, is sufficient, taking the form of

$$f(x_1, t_1; x_2, t_2; \dots; x_n, t_n) \tag{7}$$

The random processes analyzed in the present work will principally contain 12 time values, representing a monthly statistic of a time series, such as a mean, maximum or minimum value. This way, the process's evolution in time is described by discrete time steps Δt representing the duration of one month. The random process will therefore take the form of

$$\underline{X} = \{\underline{x}_1, \underline{x}_2, \dots, \underline{x}_{12}\} \tag{8}$$

In some cases, each realization will consist of 365 daily values, where the process then takes the form of

$$\underline{X} = \{\underline{x}_1, \underline{x}_2, \dots, \underline{x}_{365}\} \tag{9}$$

A description for the date of February 29th was not considered, because in the majority of the cases, the number of realizations of these values was too little to obtain an amount of values that was statistically sufficiently significant for rational probabilistic analysis.

2.2 Extreme Events

Extreme value analysis is one of the fields in statistics that has gained importance in the past 60 years in different fields of study, including hydrology. Its main objective is to describe the stochastic behavior of a process at unusually large or small levels (Coles, 2001), in particular the estimation of the probability of occurrence of these events.

As described in Coles (2001), in many cases, in which extreme value analysis is applied, existing data does not prove sufficient to describe the statistical behavior of the process to be analyzed with certainty. Therefore, also statistical indices, such as a value that describes an extreme event cannot be obtained exactly. Some methods exist that allow the estimation of mentioned indices assuming the number of data values to approach infinity.

In many cases of statistical analysis, an extreme event is defined as an event within a statistically valid dataset that is rarely found and that lies below or above a defined threshold calculated from said dataset. However, in the sense of the stochastic approach presented in this work, it was not considered feasible to follow the same method. For this purpose, a threshold would be needed for each valid set of values, which is represented by the random variable. The threshold would have been calculated as a probability of exceedance and biased the data to an extent that was not considered to be reasonable.

Hence, in this study, extreme events were exclusively defined as the maximum and minimum values of each month on record obtained from the daily observations in the hydrological time series, which is why it was important to count on daily data.

In the case of monthly maxima, a random process was defined, in which 12 random variables were contained, each representing the monthly maximum values of the respecting month for each of the realizations of the process.

$$\underline{X}(t) = \{\underline{x}_1, \underline{x}_2, \dots, \underline{x}_{12}\} \tag{10}$$

with

$$\underline{x}_i = \{max(x'_{1,i}), max(x'_{2,i}), \dots, max(x'_{n,i})\} \tag{11}$$

where *x'* denotes the subset of daily values of the month *i* and *n* the total number of observed years.

The same procedure was applied for minimum values.

Two random variables containing the annual mean, maximum and minimum value, respectively, were created to represent the data on the annual level and only used in trend analysis described in Section 3 of this work.

2.3 Statistical Methods

2.3.1 PDF Fitting and Kolmogorov goodness of fit test

After proving its randomness, theoretical probability density functions were fitted to the empirical distribution of the random variables analyzed in this work to evaluate the best fit in each occasion. For these tests, the empirical distribution function was built following the Kritskiy – Menkel equation (Moreno, 2011)

$$F(Q_i) = \frac{i}{n+1} \tag{12}$$

where *i* is the position of plotting data from highest to lowest value and n is the number of available data in each dataset. To this empirical distribution, 12 different theoretical probability density functions were fitted, which were all included in the *Scipy Stats* Package described in section 2.4.1.

The following 12 distributions were used:

- Normal
- Lognormal
- Gamma
- Loggamma
- Gumbel with positive skew
- Gumbel with negative skew
- Weibull Min
- Weibull Max
- Powerlaw
- Pareto
- Exponential
- Logistic

The best fit was determined using the Kolmogorov goodness-of-fit test (Moreno, 2011), which is a non-parametric test to compare the equality of the empirical distribution F*(x) with the theoretical probability distribution F(x). The statistic λ is determined by the maximum difference between the theoretical and the empirical function:

$$\lambda = (max|F^*(x) - F(x)|)\sqrt{n} \tag{13}$$

where *n* represents the number of elements contained by the empirical distribution. In the next step, the critical value λ_q of the Kolmogorov distribution is determined. If $\lambda < \lambda_q$, the null hypothesis is accepted and the theoretical and empirical distribution are considered to match. The distribution of the Kolmogorov criteria can be approximated as follows (Moreno, 2011):

$$\lambda_q = 1 - \sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2\lambda^2} \tag{14}$$

The theoretical distribution resulting in the least mean squared error between empirical and theoretical distributions was chosen out of all those that passed the Kolmogorov goodness-of-fit test.

2.3.2 Mann-Kendall trend test

The Mann-Kendall trend test, also referred to as Kendall τ is a non-parametric test that determines the significance of a trend using consecutive pairs of data values in the time series to compare for a positive or negative difference, which does not take into account the magnitude of this difference. This test is resistant to outliers, can be applied to samples with small sizes and is well-suited for variables that are not necessarily normally distributed (Helsel and Hirsch, 2002; Kunkel et al., 2010; Morin, 2011).

The median of the slopes of all consecutive data pair values *P* is represented by the statistic

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sign\left(P_j - P_i\right) \tag{15}$$

where *n* is the sample size. For a sample size of *n > 8*, *S* has an approximate normal distribution (Morin, 2011), with mean zero and a variance that depends on the sample size and the number of ties *q*,

$$Var(S) = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_{i=1}^{n} q_i i(i-1)(2i+5)\right] \tag{16}$$

The final test statistic *z*,

$$z = \frac{S - sign(S)}{\sqrt{Var(S)}} \tag{19}$$

determines if the null hypothesis of no existing trend in the observed data is rejected if it is larger than the critical value calculated from the normal distribution with a given probability of exceedance.

2.3.3 Thomas Algorithm

The Thomas Algorithm is a numerical method used in linear algebra, which permits solving tridiagonal systems of equations (Weickert et al., 1998) and was in this applied to the implicit solution of the Fokker-Planck-Kolmogorov equation algorithm, as described in Section 4.2.1. The tridiagonal equation system has the structure of

$$a_i x_{i-1} + b_i x_i + c_i x_{i+1} = d_i \tag{20}$$

In this case, *i* might represent an index in the observation interval of a modeled variable, or time. The whole system can be rewritten in matrix form, where the coefficients *a*, *b* and *c* form a three-diagonal matrix that is multiplied by the vector *x*, resulting in the vector *d*.

The Thomas algorithm consists of two modification steps, one in forward and another in backward direction. In the first step, the coefficients are modified recursively in the forward direction, changing the values of *c* and *d*, where the *** marks the modified coefficient (Hoffman and Frankel, 2001).

$$c_i^* = \begin{cases} \dfrac{c_1}{b_1} & if \quad i = 1 \\ \dfrac{c_i}{b_i - c_{i-1}^* a_i} & if \quad i > 1 \end{cases} \tag{21}$$

$$d_i^* = \begin{cases} \dfrac{d_1}{b_1} & if \quad i = 1 \\ \dfrac{d_i - d_{i-1}^* a_i}{b_i - c_{i-1}^* a_i} & if \quad i > 1 \end{cases} \tag{22}$$

The second modification step assigns values to the variable x, which is modified starting from the last value in the vector and advancing in backward direction.

$$x_n = d_n^* \tag{23}$$

$$x_i = d_i^* - c_i^* x_{i-1} \quad from\, i = n-1, \dots , 1 \tag{24}$$

where *n* represents the number of values contained in the vector *x*. The resulting vector *x* is the solution of the Thomas algorithm.

2.3.4 Multiple Regression Analysis

In this study, multiple linear regression analysis was used to estimate the degree to which some variables influence others. Therefore, correlation coefficients were calculated for each independent variable in order to describe the dependent variable as a linear combination that represents a type of weighted sum, using the coefficients as weight. This takes the form of

$$y = a + \ b_1 x_1 + b_2 x_2 + \ldots + b_n x_n \tag{25}$$

In equation 25, *y* is the dependent variable, x_i are the dependent variables, b_i the assigned weights and *a* is an additive constant or intercept. With only one independent variable, this can be seen as a point cloud in 2-dimensional space, to which a line is fit, but with each additional variable, the space increases by one dimension.

Regression analysis was solved using the Ordinary Least Squares method, which is used to estimate the regression coefficients by minimizing the sum of the squared vertical distances between the observed data and the predicted one by the linear approximation. The exact methodology is described in Feldman and Valdez-Flores (2009). For each multiple regression, a regression value is output to describe the overall coherency between the variables.

For each of the coefficients, it has to be determined if it is statistically valid and may therefore be used to serve as an indicator of the dependent variable. Therefore, the variances of the errors are calculated for each coefficient, which allows for the construction of Student t distributed random variables that may be used for hypothesis testing and building confidence intervals (Feldman and Valdez-Flores, 2009). Therefore, for each independent variable, the confidence interval of 95% was regarded to accept the result of the regression.

To establish the correlation between the dependent and the independent variables, only the statistically valid ones were identified with a correlation analysis of all variables in a first run. The analysis was repeated in a second run with only those independent variables that were statistically valid.

2.3.5 Kullback-Leibler divergence criteria

The Kullback-Leibler divergence (Kullback and Leibler, 1951) is a measure to describe the difference of two probability distributions. In other words, it can be described as the measure of information that is lost when one of the distributions is used to model the other. Relying on Shannon's theory of information, this difference is the relative entropy of one probability distribution with respect to the other, which is expressed in bits.

In most applications, the observed probability distribution is named *p*, and the modeled one *q*. The divergence *d* between the two distributions is then calculated as

$$d = \sum_i p_i log_2 \left(\frac{p_i}{q_i}\right) \tag{26}$$

2.4 Computational Tools

Computation was necessary in many different areas of this research. This did not only include the analysis, but also the acquisition of data and the visualization and presentation of the results.

2.4.1 Python programming language

The principal computational tool used in this research was the Python programming language. All information stated in the following paragraphs was obtained from the project pages of the respective modules, as well as their documentation.

Python is a high-level programming language ("Python.org," 2014), which focuses on easy code readability and implementation, and allows creating short and clear programs. Therefore, Python is used mainly for scripting purposes and in the scientific environment. Python is distributed under the Python Software Foundation License, which is comparable to the GNU General Public License used for free software distribution and available for cross-platform use. Python has a large standard library, but especially for scientific purposes, there is a wide range of additional packages. By June 2014, more than 44.000 packages were available at the official repository, the Python Package Index ("PyPI - the Python Package Index," 2014). For easier software installation, a number of different distribution collections are available, which include the standard library and selected packages. For this investigation, the Anaconda Python distribution was used, which is compiled by Continuum Analytics and focuses especially on scientific computing and large-scale data processing. The principal components of Python used in this research are described below and can be referred to in later sections that describe the functionality of some of the scripts.

The Python standard library provides the commonly used functions of operation in Python. From this library, principally basic mathematical operations, date and time modules, Input/Output (I/O) modules to read and write different file types and the list data type were used. This data type is a collection of values of any data type, which do not have to be equal within the same list. Lists were the most basic data type used in the analyses and principally applied to save IDs or station names, as well as the creation and stepwise extension of time series. The main I/O libraries used were for reading or writing text and comma-separated value (CSV) files, browsing and downloading files on FTP servers and reading ZIP files. As of June 2014, the most recent Python version was 3.4, in this research however, version 2.7 was used.

NumPy and *SciPy* ("NumPy — Numpy.org," 2014, "SciPy.org," 2014) are the two most widely used packages for scientific computing with Python, where *NumPy* is the more fundamental and *SciPy* the extensive collection of mathematical and statistical tools. The most frequently used component of *NumPy* in this research is the array data format, which allows the processing of matrices and provides functions associated with their use. Furthermore, a wide range of mathematical and statistical function, especially those that permitted the presence of Null values in a matrix was used in all of the analyses. *SciPy* complements the range of functions not included in *NumPy* and provides the essential tool for probabilistic analysis, the *stats* module. This module includes over 80 probability distributions, which all offer efficient methods for the identification of parameters with the maximum-likelihood method ("Statistics (scipy.stats) — SciPy v0.14.0 Reference Guide," 2014). Furthermore, *SciPy* offers a big variety of mathematical optimization functions.

Pandas ("pandas: Python Data Analysis Library," 2014) is a library that provides additional data structures and data analysis tools. Its main component used in this work is the DataFrame structure, which can be described as an indexed array object. It indexes its rows and columns, which permits easy appending of new columns with the same row indices as an existing DataFrame and therefore an automated reorganization of the data. This was especially useful for the creation of data files of multiple time series, where the date was used as the row index and each time series was saved as a column of the DataFrame.

Matplotlib ("matplotlib: python plotting," 2014) is a Python plotting library, which is specialized for scientific graphics. *Matplotlib* allows the creation of charts of all types both in 2 and 3 dimensions. All results of the analyses of this study were plotted with this module.

Python script files were created containing all the data analysis and data operation tools that were used in this research, which were not already included in one of the Python packages. These script files are used as various modules that contain approximately 50 functions and were imported as external libraries in all of the other scripts used for data analysis. A list of all functions created for this study and included in the modules is provided in Annex B.

2.4.2 ArcSWAT

SWAT stands short for *Soil and Water Assessment Tool* and is a software tool developed and distributed by United States Department of Agriculture: Agricultural Research Service (USDA-ARS) and Texas A&M University system. The tool is developed to execute quantitative and qualitative analysis of environmental impacts on small watersheds and river basins. SWAT has been used in a large number of scientific works and other projects around the globe ("SWAT | Soil and Water Assessment Tool," 2012).

ArcSWAT is an extension for the ESRI ArcGIS software to implement SWAT's functionality in custom GIS software and was operated in this work with ArcGIS version 9.3. For this work, ArcSWAT was used to delimit the watersheds for the test basins according to the obtained elevation information described in section 2.5.2.

2.5 Data collection

Apart from creating the software tools for hydrological analysis, a large part of the preliminary work in this study was dedicated to the retrieval of high-quality data. Both hydrometeorological, as well as geospatial data were assembled from a number of different data sources on the internet. One of the most important factor of this work was to use exclusively data that was offered free of charge, whenever this was possible.

In a first step, a collection was created that contained all the institutions that provided data. For this reason, an extensive search was conducted, which lead to a rough overview of the availability of data. In the next step, data samples were downloaded and Python scripts generated to automatically read the data in the formats, in which the information is saved and made available. Finally, completeness and quality checks were performed and the information was saved in a unified format, which facilitated further work with the information.

In the following sections, the data that was used in the investigation will be described. However, it has to be stated that this is merely a small part of the information that was actually downloaded and analyzed. A significant part of the data was not fit for further use, due to a multitude of reasons, including high percentages of missing values, short periods of observation or access restrictions that would have resulted in extensive manual preparation and therefore would have caused delays in the process.

For the sake of better readability of the following sections, all consulted hydrometeorological and geographical data sources can be found in Annex A and will not be quoted in the text.

2.5.1 Hydrometeorological data

Hydrometeorological data had to meet two criteria to be used in this research. First, due to the types of analysis, which were foreseen to be executed, only data on a daily level were downloaded. Monthly and hourly data were not obtained from data sources, although in most analysis, monthly and annual datasets were derived from the daily data in later processing steps. Second, the series had to provide at least 80% of completeness to guarantee the valid base for the analyses to be performed.

Hydrometeorological data is collected by a lot of organizations worldwide, many of which provide it free of charge in databases that can be accessed via their web appearance. The hydrometeorological variables of interest for this investigation were discharge, precipitation and temperature, which are the most common information collected and therefore accessible from most of the institutions that provide data.

In the majority of the cases, databases are provided on a national or worldwide basis, although data collections from projects in smaller areas are also available. The fact that data collected on a bigger geographical scale implies the existence of quality control was the main reason why databases on a worldwide and national level were primarily accessed. Another reason was the unified data format, which allowed retrieving bigger amounts of information with a single interface used to download and evaluate the data. This way, it was aspired to conform a consistent data set with a comparable level of data quality.

On a worldwide level, two main databases were consulted that provided data for all three variables. Firstly, data provided at the National Climatic Data Center (NCDC) of the National Oceanic and Atmospheric Center (NOAA) in the United States was accessed. The Climate Data Online Search and the Global Historical Climatology Network (GHCN), which were later combined into one central data search, provide historical and real-time data from stations around the world. Especially data from the GHCN was accessed, which provides information for more than 85.000 stations worldwide and more than 50 variables on a daily and monthly resolution, which include precipitation, maximum and minimum temperature. The data can be downloaded as automatically generated text files for each station that all have the same data structure. Access to the database is possible via HTTP as well as FTP, the latter of which enabled an automatic download process with a Python script.

Secondly, the Global Runoff Data Centre (GRDC) is a data collection of over 8.000 discharge data stations from over 150 countries on a monthly and daily level, which is operated by the German Federal Institute of Hydrology. For this database, an automated download is not possible. The data access procedure requires a request by email, in which the specification of the data has to be made. The data files are delivered in return per email, which also contains separate data files for each station with a unified data format.

Most of the discharge data, as well as some precipitation information to complement the GHCN data, was obtained from national data providers. Data from the following institutions was downloaded:

- Argentina: Sub secretary of Hydrological Resources (Subsecretaría de Recursos Hídricos): BDHI database
- Australia: Bureau of Meteorology
- Australia: Government of Queensland
- Austria: Ministry of Life: eHyd database

- Brazil: National Agency for Water (Agência Nacional de Águas, ANA): Hidroweb database
- Canada: Environment Canada (HYDAT database)
- Mexico: National Commission for Water (Comisión Nacional del Agua, CONAGUA)
- South Africa: Department of Water Affairs
- United Kingdom: Centre of Ecology and Hydrology
- United States: United States Geological Survey (USGS): National Water Information System

For Colombia, discharge, precipitation and temperature data was provided by the Institute of Hydrology, Meteorology and Environmental Studies (Instituto de Hidrología, Meteorología y Estudios Ambientales, IDEAM) for research purposes to Efraín Domínguez and could also be used for this study.

The time frame of available data varied between the sources of information, the longest series dated back into the 1700s, which could be found in the GHCN database for a meteorological station in Italy. At the same time, the data sources of US-American institutions usually provide data until almost the current date, where quality control takes place in larger periodic cycles, such as quarterly or yearly. Data from other national organizations are usually updated in yearly intervals, where data can be obtained until the last complete year before the current date or two years back. In all data collections, information of the operation period of already closed stations is also still included in the database.

For most data providers, station lists were available, which provided at least the name or code of the hydrometeorological station, along with its geographic coordinates. Additional information that was provided in the station lists was the period of observation, altitude and operator of the station, as well as the percentage of missing values in some cases. This information was consulted first and used to create a list of stations, which was analyzed in a Geographic Information System and filtered using the given information as decision criteria for choosing a sample of stations whose data was to be downloaded. If no station list was available, it was created from the metadata provided.

During the data retrieval, a variety of challenges surfaced, mainly the fact that each data provider used a different format for saving data, which was the case for the file format, size and the data structure. File formats included text files, many of which used proper file extensions, such as for example *.dly, CSV files, or in the case of the Colombian data a proper file format that needed scripts to prepare. In some cases, it was possible to access data from different stations in one file, but principally the information from each station was provided separately in single files. The biggest challenge, however, was the difference in the file structure. In most cases, the time series were provided as a two-column list including the date (and hour) in one and the data value in the other with a single line for each observation. In some cases, especially for files including multiple stations, an additional column for the station ID was available, as well as for data variables in case of multiple variables or columns indicating data quality. If data was not listed as one observation per line, it was presented as a monthly table, with 31 data values representing the observations of each day of the month in the same line. These tables were sometimes tab-separated, in other cases some separation character was used, and in some cases an additional data quality flag was added for each observation, as it was the case for the GHCN data. Additionally, in some cases, files include lines with metadata at the beginning of the file. This metadata, if present, had to be extracted separately or in most cases was skipped.

Facing all the above mentioned challenges, it was quickly clear that it was necessary to create an interface for each separate data format, which extracted the data from the raw data files into a unified structure that could be used in further analysis. All interfaces were written in Python code and collected in a file called dataops.py, which served as a module for data extraction for each data source. In the cases where it was possible, it also included the automatic data download from a list of given stations. The Python code uses the basic module, as well as *NumPy* and *Scipy* functions and the different I/O interfaces to download, load and save the data files. After reading the data files, the *Pandas* module was the essential tool to unify the data in DataFrames, where the function of data indices served to readily reorganize the data with the row indices representing the observation dates and the column names the codes of the hydrometeorological stations. This way, a table of all stations could be generated, which was saved as a CSV file for each data source separately, due to file size. These CSV files could again be loaded easily with the Pandas module and a desired section of the data extracted for different analysis, both for specific time periods and also to select a subset of stations.

2.5.2 Geospatial data

Geospatial data of different types was obtained or created for this work and apart from serving as a base for decisions it was used to geographically locate the hydrometeorological data used. In the following paragraphs, the types of data used and their sources are briefly described.

Firstly, all hydrometeorological stations were displayed according to the coordinates provided in the station information. From the station lists obtained from the data providers or metadata, point shapefiles were created for further use. Additionally, a shapefile displaying the world country borders was downloaded for free at Geocommons.com.

For the delimitation of watershed boundaries with ArcSWAT, the elevation data provided by the HydroSHEDS project was used. This data is published by the USGS free of charge and is based on high-resolution data from NASA's Shuttle Radar Topography Mission (SRTM). A variety of datasets are provided, from which the digital elevation model (DEM) with a resolution of 3 degree-seconds was chosen for being the one with the highest resolution, which corresponds to approximately 93 meters along the equator.

Geographical data representing the river networks in the study areas was found at the national or regional agencies for hydrological information. Although HydroSHEDS also provides river networks derived from the elevation data, it was considered that the datasets from local institutions are more accurate. For the display of river data for Europe, especially for the Enns catchment in Austria, the river layer from the Ecrins dataset was obtained from the European Environmental Agency (EEA). For the US, the river dataset from NOAA was used. The dataset used for Colombian rivers originated from the IDEAM for the doctoral thesis of Efraín Domínguez and contained the permission to be used for scientific projects.

Other data, such as the USGS's hydrological units dataset, were useful for the study of available stations to determine the study area in the USA. Different spatial datasets were obtained for other countries, which were not used in the analysis for obtaining results presented in this work.

A satellite image for the whole world was used as a Web Map Service from the NASA website.

2.6 Study areas for extreme event analysis

The study of extreme events was conducted in 4 river basins in different parts of the world. The intention was to use basins with comparable sizes in both hemispheres, as well as close to the equator. A total of 5 candidates with satisfactory data availability were selected, out of which 4 were chosen. The main criteria for the choice of the basins were the existence and sufficiently complete data, as well as a correlation between runoff and precipitation series. Therefore, the data had to include at least a completeness of 80% of all time series on a daily level. Also, a data structure of the random processes of monthly maximum and minimum discharges was required to indicate a Markovian process by the analysis of its cross correlation, as is described in section 4.1.5.

The selected basins were the Enns River (Austria), the Upper Magdalena River (Colombia), the Upper Great Miami River (USA) and the Brisbane River (Australia), as displayed in figure 3. The Crocodile River Basin in South Africa was among the candidates, but was not used because of reasons in its data structure. Therefore it will not be described in detail.

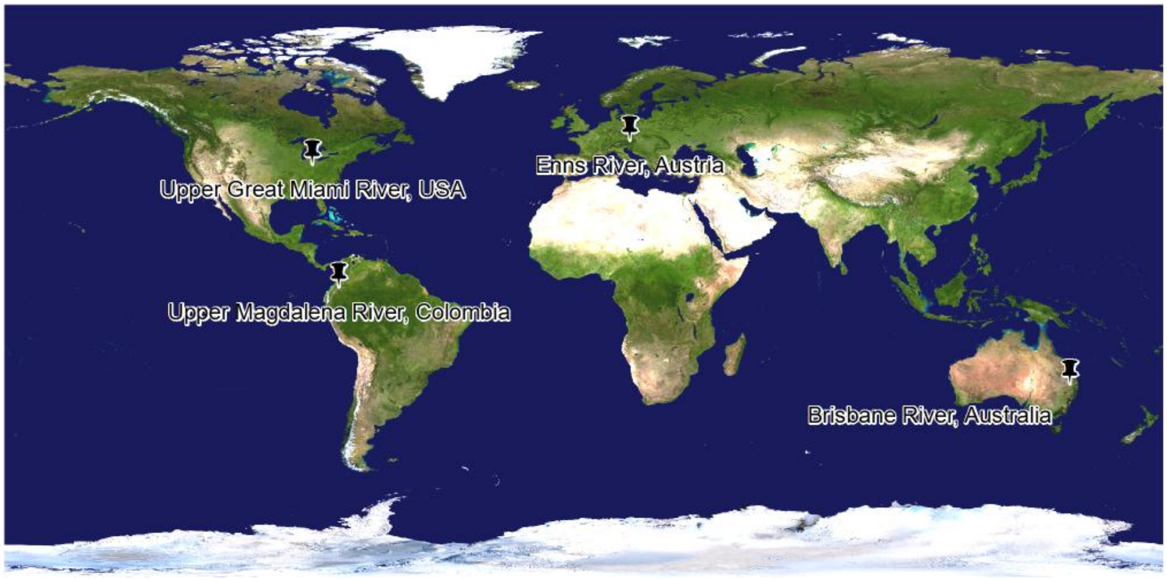
Upper Great Miami River, USA
Enns River, Austria
Upper Magdalena River, Colombia
Brisbane River, Australia

Figure 3. Location of the study areas

For each basin, only the discharge data from the station representing its outlet was used for further analysis in order to evaluate the changes in the complete watershed. Among all precipitation stations in each basin, those with a correlation of over 60% with the discharge data for the same date or one of up to 5 lag days was considered. Precipitation time series were constructed calculating the mean of the observed values of all those stations. Due to the lower availability of freely available temperature stations in the regions of most test basins, the temperature station closest to the basin outlet was used.

The information given in the descriptions of each river basin in the following paragraphs was retrieved from the datasets and their metadata.

2.6.1 Enns River Basin (Austria)

The Enns is a river in central Austria, which has a basin area of approximately 6100 square kilometers. It embarks parts of the northern range of the Alps in its upper reaches and flows into the Danube River from the South. The Enns River itself has a total length of 253 kilometers, and the basin also includes two other major rivers, the Steyr of 68 km and the Salza of 90 km, and the basin includes parts of the 4 federal states of Lower Austria, Upper Austria, Salzburg and Styria. The highest elevation in the basin is approximately 2600 meters above sea level and the mouth of the river at 240 meters, where the river is about 100 meters wide and has an average discharge of over 200m^3/s.

The lowest discharge station is located in the city of Steyr, after the union of the Enns and Steyr rivers and about 30 km from the outfall of the Enns into the Danube at an elevation of 283 meters. Its drainage area is 5915 square kilometers big and was used for this investigation.

Within the basin of the Steyr station, 34 precipitation stations are located, which meet the criteria established for data completeness and were all obtained from the eHyd data portal. 23 of the precipitation stations met the correlation criteria and were used for further analysis. The records length is 40 years from 1971 to 2010. Figure 4 shows the area of the river basin, as well as the location of the hydrometeorological stations. Precipitation stations with a significant correlation with the discharge series are marked separately.

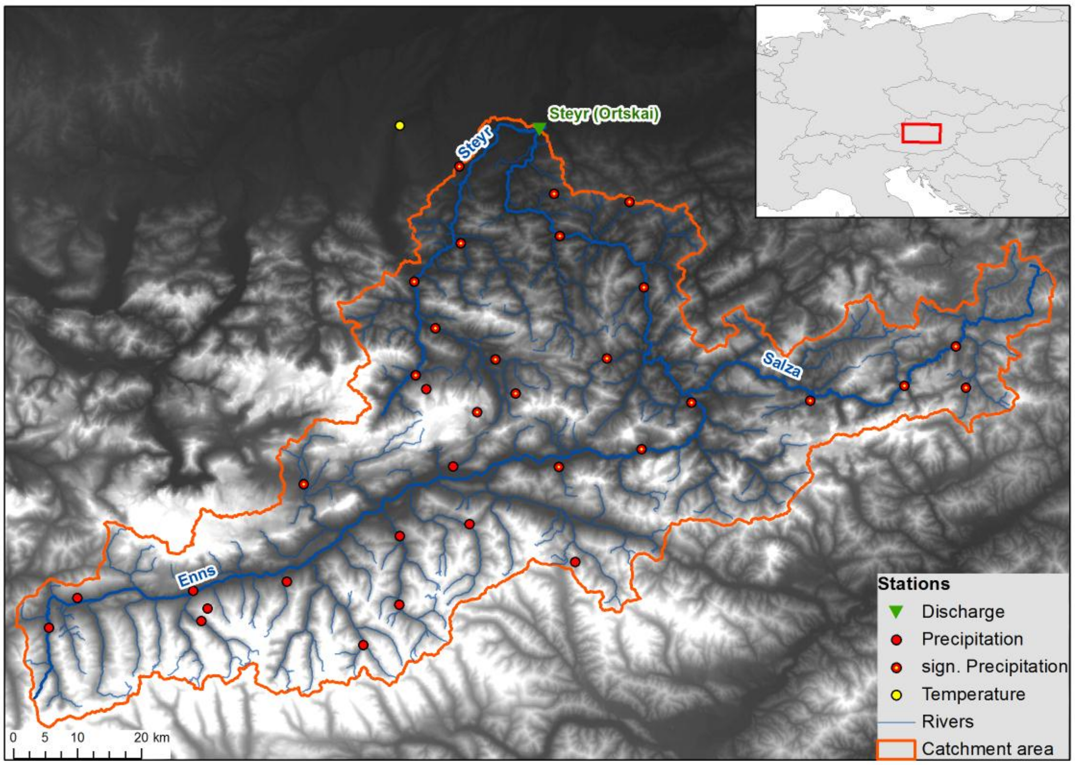

Figure 4. Illustration of the Enns River basin

The following table gives an overview of the characteristics of each station used. Data completeness refers only to the analyzed period from 1971 to 2010 on a daily level.

Variable	Station ID	Data Source	Start Date	End Date	Completeness
Discharge	205922	eHyd	01.01.1951	31.12.2010	100%
Precipitation	105643	eHyd	01.01.1971	31.12.2010	100%
Precipitation	106021	eHyd	01.01.1971	31.12.2010	100%
Precipitation	106153	eHyd	01.01.1971	31.12.2010	100%

Variable	Station ID	Data Source	Start Date	End Date	Completeness
Precipitation	106161	eHyd	01.01.1971	31.12.2010	100%
Precipitation	106203	eHyd	01.01.1971	31.12.2010	100%
Precipitation	106229	eHyd	01.01.1971	31.12.2010	100%
Precipitation	106237	eHyd	01.01.1971	31.12.2010	100%
Precipitation	106245	eHyd	01.01.1971	31.12.2010	100%
Precipitation	106252	eHyd	01.01.1971	31.12.2010	100%
Precipitation	106278	eHyd	01.01.1971	31.12.2010	100%
Precipitation	106286	eHyd	01.01.1971	31.12.2010	100%
Precipitation	106310	eHyd	01.01.1971	31.12.2010	100%
Precipitation	106328	eHyd	01.01.1971	31.12.2010	100%
Precipitation	106336	eHyd	01.01.1971	31.12.2010	100%
Precipitation	106351	eHyd	01.01.1971	31.12.2010	100%
Precipitation	106377	eHyd	01.01.1971	31.12.2010	100%
Precipitation	106401	eHyd	01.01.1971	31.12.2010	100%
Precipitation	106419	eHyd	01.01.1971	31.12.2010	100%
Precipitation	106427	eHyd	01.01.1971	31.12.2010	100%
Precipitation	106435	eHyd	01.01.1971	31.12.2010	100%
Precipitation	106443	eHyd	01.01.1971	31.12.2010	100%
Precipitation	106450	eHyd	01.01.1971	31.12.2010	100%
Precipitation	106484	eHyd	01.01.1971	31.12.2010	100%
Temperature	AU000005010	GHCN-D	01.01.1876	31.12.2013	100%

Table 1. Hydrometeorological stations used in the Enns River basin

2.6.2 Upper Magdalena River Basin (Colombia)

The Magdalena River is the biggest river in Colombia with a length of 1530 km, which flows into the Caribbean Sea at Barranquilla. Its source is located in the Andean mountains at an elevation of almost 3700 m and has a drainage area of 257.500 square kilometers, including also the sub basin of the Cauca River.

The Betania reservoir is located about 200 km from the source in the Upper Magdalena Basin, and the last discharge station before it is the Puente Balseadero station, located approximately 150 km from the source at 688 m above sea level and has a drainage basin of 5.850 square kilometers. The two other major rivers in the basin are the Suaza River with a length of 136 km and the Guarapas River. Seven IDEAM precipitation stations are within the basin area, which range over 39 years from 1972 to 2010, and out of which 5 were in accordance with the completeness and correlation criteria and therefore used.

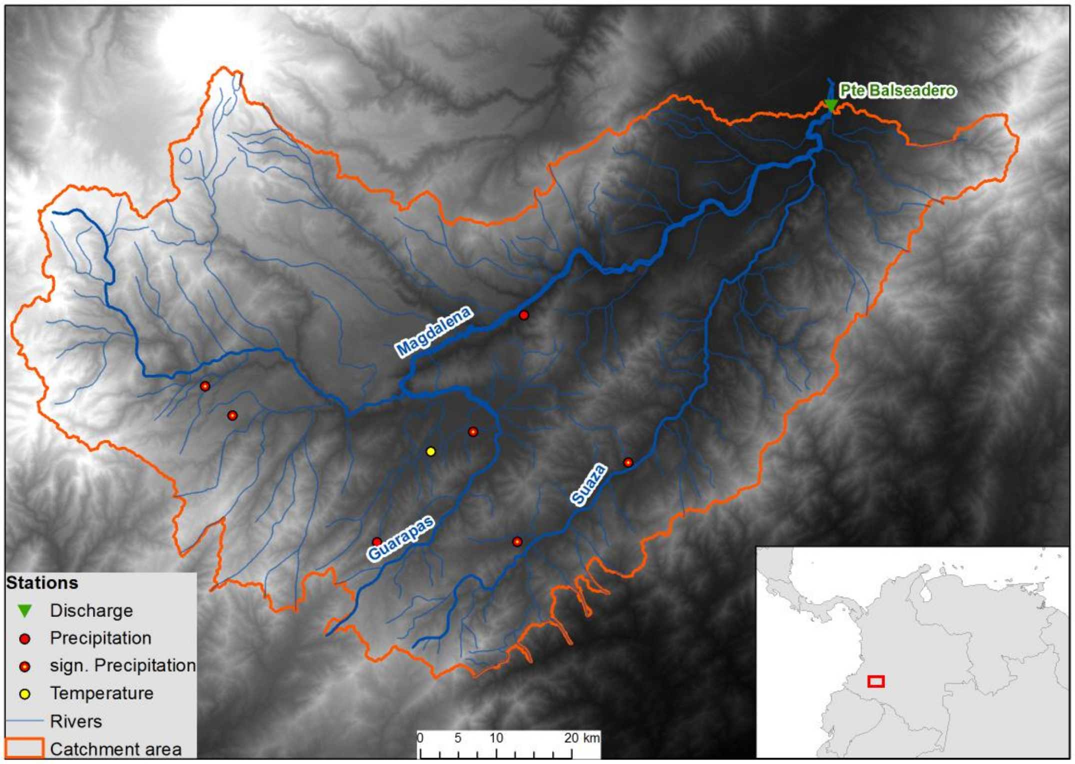

Figure 5. Illustration of the Upper Magdalena River basin

The following table gives an overview of the characteristics of each station used. Data completeness refers only to the analyzed period from 1972 to 2010 on a daily level.

Variable	Station ID	Data Source	Start Date	End Date	Completeness
Discharge	21047010	IDEAM	01.01.1972	31.12.2010	99.76%
Precipitation	21010110	IDEAM	01.01.1972	31.12.2010	99.39%
Precipitation	21010140	IDEAM	05.11.1975	31.12.2010	89.69%
Precipitation	21010160	IDEAM	01.12.1975	31.12.2010	87.46%
Precipitation	21030060	IDEAM	01.01.1972	31.12.2010	98.98%
Precipitation	21030080	IDEAM	01.01.1972	31.12.2010	98.94%
Temperature	21015020	IDEAM	01.01.1976	31.12.2010	84.11%

Table 2. Hydrometeorological stations used in the Upper Magdalena River basin

2.6.3 Upper Great Miami River Basin (USA)

The Great Miami River is a northern tributary to the Ohio River in the United States of America with a total basin area of almost 14.000 square kilometers. Its basin is located in the states of Ohio and Indiana.

For this study, only the basin of the Upper Great Miami River was used, an area that was also established as a cataloging unit of the Hydrologic Units by the USGS. It is delimited by the discharge station located at Dayton, OH, which is exactly midway between the source and the mouth of the river, approximately 128 km from both. The station is located about 1 km downstream of the union with Mad River, one of the two other major rivers in the basin with a length of 106 km. The other major river is the Stillwater with a length of 111 km. The total basin of the Upper Great Miami River is 6.500 square km big and includes 14 GHCN precipitation stations, out of which 8 fulfilled the correlation criteria. The data chosen are the 65 years from 1948 to 2012.

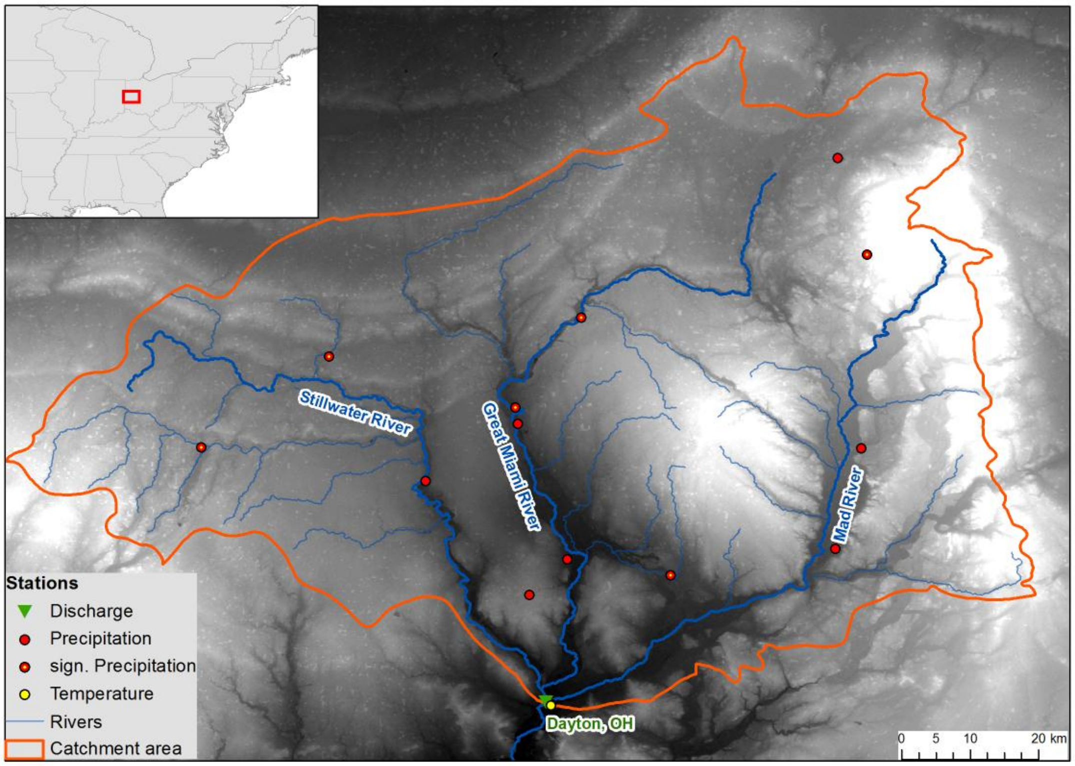

Figure 6. Illustration of the Upper Great Miami River basin

The following table gives an overview of the characteristics of each station used. Data completeness refers only to the analyzed period from 1948 to 2012 on a daily level.

Variable	Station ID	Data Source	Start Date	End Date	Completeness
Discharge	3270500	USGS	01.04.1913	31.12.2013	99.97%
Precipitation	USC00330563	GHCN-D	02.04.1894	31.12.2013	96.55%
Precipitation	USC00332067	GHCN-D	01.06.1893	31.12.2013	98.95%
Precipitation	USC00333375	GHCN-D	01.06.1893	31.12.2013	99.07%

Variable	Station ID	Data Source	Start Date	End Date	Completeness
Precipitation	USC00335786	GHCN-D	01.01.1914	31.12.2013	97.76%
Precipitation	USC00336645	GHCN-D	01.07.1893	31.12.2013	89.44%
Precipitation	USC00337693	GHCN-D	01.05.1948	31.12.2013	98.67%
Precipitation	USC00338642	GHCN-D	01.01.1914	31.12.2013	98.36%
Precipitation	USW00093815	GHCN-D	01.01.1948	31.12.2013	99.99%
Temperature	USC00332067	GHCN-D	01.06.1893	31.12.2013	95.15%

Table 3. Hydrometeorological stations used in the Upper Great Miami River basin

2.6.4 Brisbane River Basin (Australia)

The Brisbane River is located in eastern Australia in the territory of Queensland, flowing into the Pacific Ocean near the city of Brisbane. The whole basin has an area of 13.541 square kilometers and its altitude ranges from 2320 meters to sea level. The river has a total length of 345 kilometers.

For this study, the discharge station located at Savages Crossing is used, which is located approximately 130 kilometers from the mouth of the river and at an altitude of 42 meters. The basin area draining this station is 10.000 square kilometers big with the main sub basins being those of Lockyer Creek and Stanley River. It is located 18 kilometers downstream of Lake Wivenhoe and the Wivenhoe dam, just below which the Lockyer Creek flows into the Brisbane River. In the basin, 38 precipitation stations are located, out of which 13 fulfilled the completeness and correlation criteria. The data could be used during the period of the 52 years from 1961 to 2012, which also includes the data from the big flood in the Brisbane region at the beginning of 2011 mentioned before. In the map displayed in figure 7 on the next page, the rivers resulting from the watershed delineation in ArcSWAT are displayed, due to the lack of a freely available river dataset for the region.

Despite its location below a dam, no significant changes in discharge regimes could be found analyzing the time before and after its construction in the early 1980s. Therefore the station was considered suitable for further analysis.

Table 4 gives an overview of the characteristics of each station used. Data completeness refers only to the analyzed period from 1961 to 2012 on a daily level.

Variable	Station ID	Data Source	Start Date	End Date	Completeness
Discharge	143001C	Queensland Government	28.11.1958	31.12.2013	97.10%
Precipitation	ASN00040020	GHCN-D	01.01.1900	31.12.2013	98.15%
Precipitation	ASN00040056	GHCN-D	01.01.1916	31.12.2013	97.54%
Precipitation	ASN00040075	GHCN-D	01.01.1887	31.12.2013	95.50%

Variable	Station ID	Data Source	Start Date	End Date	Completeness
Precipitation	ASN00040079	GHCN-D	01.01.1894	31.12.2013	92.95%
Precipitation	ASN00040082	GHCN-D	01.01.1897	31.12.2013	99.55%
Precipitation	ASN00040083	GHCN-D	01.01.1894	31.12.2013	94.80%
Precipitation	ASN00040145	GHCN-D	01.01.1909	31.12.2013	96.28%
Precipitation	ASN00040169	GHCN-D	01.01.1915	31.12.2013	91.44%
Precipitation	ASN00040188	GHCN-D	01.01.1937	31.12.2013	89.85%
Precipitation	ASN00040189	GHCN-D	01.01.1936	31.12.2013	98.10%
Precipitation	ASN00040205	GHCN-D	01.01.1909	31.12.2013	92.67%
Precipitation	ASN00040247	GHCN-D	01.01.1928	31.12.2013	99.49%
Precipitation	ASN00040289	GHCN-D	01.01.1946	31.12.2013	90.85%
Temperature	ASN00040004	GHCN-D	01.01.1941	31.12.2013	100.00%

Table 4. Hydrometeorological stations used in the Brisbane River basin

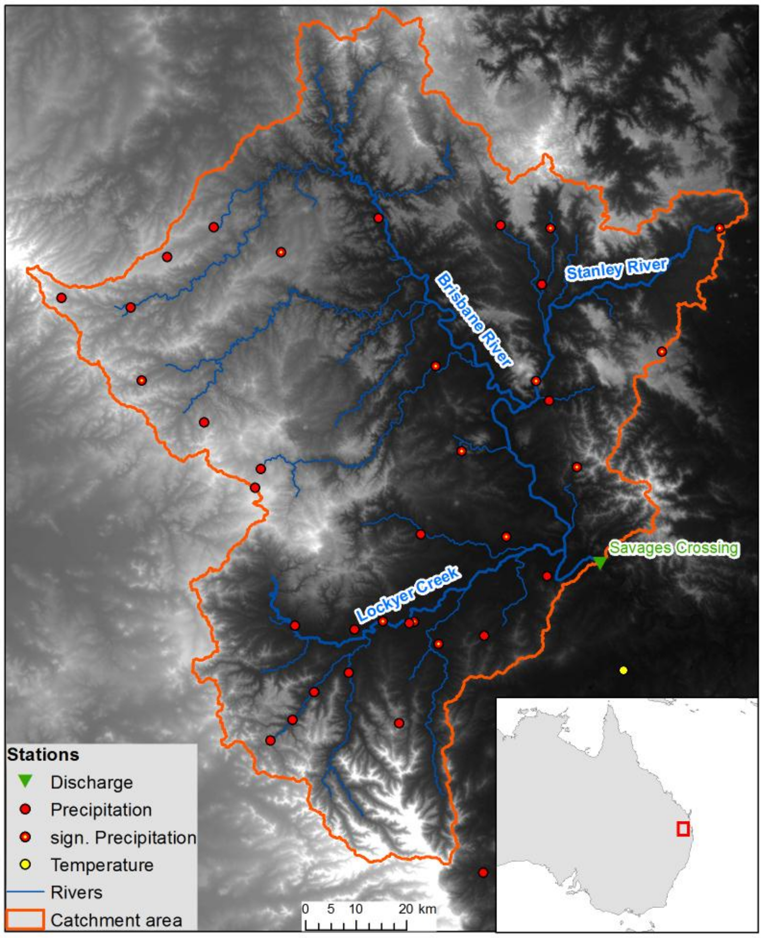

Figure 7. Illustration of the Brisbane River basin

SECTION 3

WORLDWIDE TREND ANALYSIS

In order to statistically analyze the change in the patterns and behavior of extreme events, it is essential to determine if patterns of change can be found in the global climate. Only then is the execution of such an analysis justified and reasonable. It was important to know if the changes in extreme events can sufficiently be described by trends in hydrometeorological time series or if an additional, more in-depth analysis is necessary for this purpose. For this reason, a worldwide trend analysis of time series for different hydrometeorological variables was conducted. Both trends for mean time series and for extreme value series were calculated. This analysis was conducted with the data previously gathered and organized, as described in Figure 8.

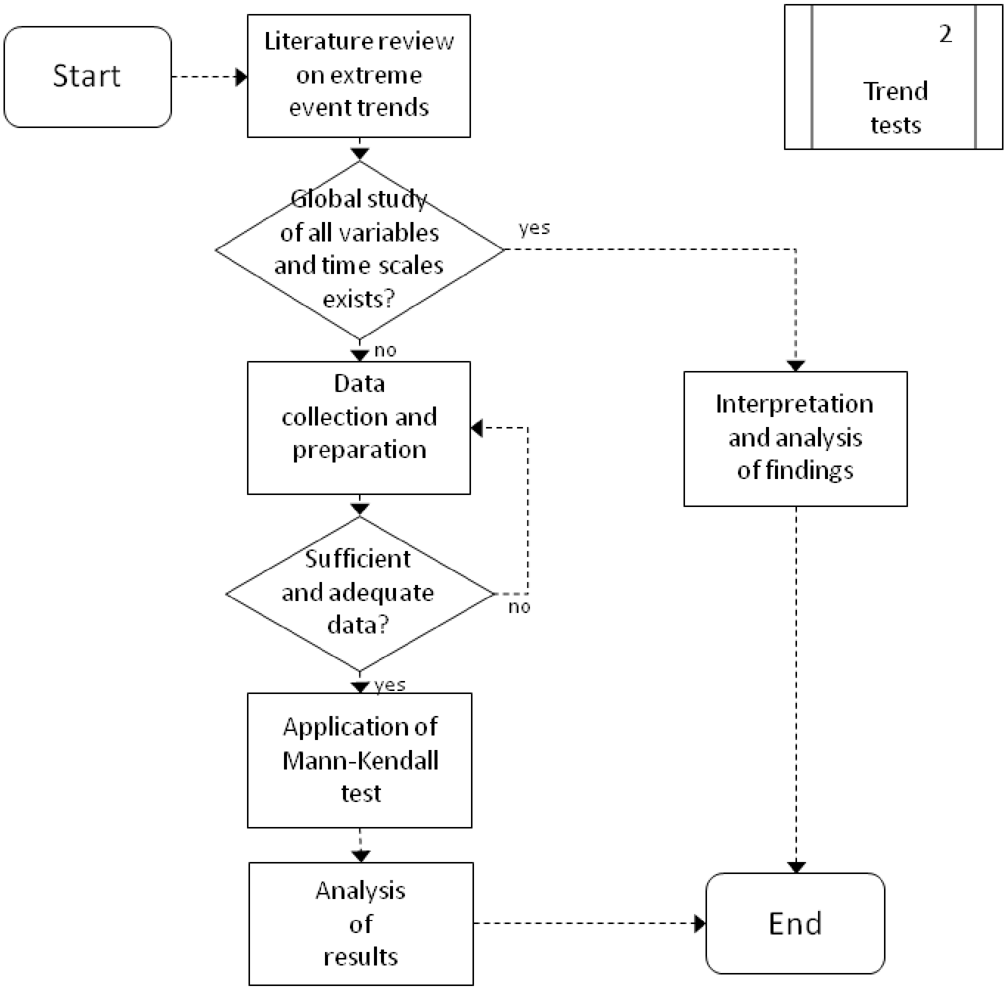
Start
Literature review on extreme event trends
2
Trend tests
Global study of all variables and time scales exists?
yes
no
Data collection and preparation
Interpretation and analysis of findings
Sufficient and adequate data?
no
yes
Application of Mann-Kendall test
Analysis of results
End

Figure 8. Schematic outline of worldwide trend analysis

3.1 Background and methodology

A large number of authors provide accounts of the existence of statistically significant trends in hydrometeorological time series, which might cause a change in the hydrometeorological regime of the area. However, the majority of these studies conducted the research in a small study area and usually did not study all relevant variables in the present work. The goal of this trend analysis was to extend the study area to the whole globe where possible and to calculate the trends for three variables on different time resolutions.

Taking into account the previous studies, it can generally be said that statistically significant trends could be found for all of the studied variables all around the world, both for the time series of the variables and their extreme value series. For mean values, trends in temperature were found to be positive in all latitudes (Aguilar et al., 2005; Del Río et al., 2011; Falvey and Garreaud, 2009; Nicholson et al., 2013), where minimum temperatures have been found to increase more frequently than maximum ones (Hu et al., 2012; Sonali and Nagesh Kumar, 2012; Xu et al., 2010). Precipitation trends were observed fewer and were usually positive (Barros et al., 2000; Vargas et al., 2002; Xu et al., 2010), in some cases no significant trends were found at all (Abghari et al., 2012; Mass et al., 2011). For discharge series, trends depended heavily on the studied area, but Dai et al. (2009) showed that over 30% of the major rivers worldwide show statistically significant trends. These trends could also be related to human activities in some studies (Wang et al., 2009; Woo et al., 2008). In a study for discharge trends in Sweden, a 5% increase was found over the 20th century, which was not statistically significant (Lindström and Bergström, 2004). In a study for Colombia (Moreno, 2011), it was found that more than 70% of temperature stations indicate a significant positive trend with no negative trends, whereas only close to a fourth of all precipitation stations show significant precipitation trends, which locally vary between positive and negative ones. These trends in all cases are consistent with discharge trends in the same region.

The majority of studies on trends in extreme events are concentrated on precipitation events, where generally an increase of heavy rainfall events could be observed. For example, Min et al. (2011) found that for two thirds of all precipitation stations in the northern hemisphere, extreme events intensified in the study period. Extreme events in discharge and temperature were not studied as intensively but usually also show an increase or intensification (Delgado et al., 2010; Nyeko-Ogiramoi et al., 2013).

In this trend study, analysis was conducted for the four hydrometeorological variables discharge, precipitation, maximum and minimum temperature on an annual, monthly and daily level. Therefore, time series were prepared as random processes of monthly and daily means as described in Section 2.1.4, in order to study the trends in each of the random variables, as well as a random variable of annual means. For precipitation series, monthly and annual sums were calculated instead of means.

Trends in extreme value time series were identified using random processes representing the monthly and annual minimum and maximum values for each variable. Precipitation minima were not considered. For this reason, only the change in the magnitude of the extreme events were analyzed and not the trends in their number of occurrence.

The time period analyzed were the 41 years from 1970 to 2010, which resulted in each of the random processes consisting of 41 realizations. The goal of extending the geographical coverage of stations as far as possible was achieved well for precipitation and temperature series, which originated from the data of the global GHCN database. However, for discharge data, due to the lack of available data in many areas of the world, especially Asia and Africa, it was far more difficult to accomplish, which resulted in a reduced study area focusing principally on the Americas and Australia.

For the selection of stations to be tested for trends, a procedure was created that chose stations following different criteria and assuring the geographically uniform distribution of the stations. For this purpose, stations were selected randomly among all available stations for the same variable. The station was used if it met the criteria of providing at least 80% of data in the required time period, and passed a test of homogeneity in time, avoiding time series with heavy changes caused by human activity, which was especially necessary for discharge stations. For this same variable, however, it was difficult to ensure a uniform distribution of stations in space due to the different station density for each data provider. Therefore it was tried to assure uniformity at least among the station collection of each of the providers.

For temperature and precipitation, principally data from the GHCN database was used, as well as 19 precipitation stations for Brazil, 7 from Argentina and 3 from Colombia, totaling 471 precipitation stations, 462 for maximum temperature and 444 for minimum temperature. Discharge data was selected from the information obtained from the national agencies mentioned in section 2.5.1. In total, 421 stations were used. The location of the stations can be seen in figures 9 to 11.

Each daily, monthly and annual random variable was tested for trends, resulting in 1 test per hydrometeorological station on the annual level, 12 on the monthly and 365 on the daily level. For the tests, the Mann-Kendall trend test was used at a 95% confidence interval and the percentage of statistically significant positive and negative trends compared to the total number of tests was evaluated. The magnitude of each trend was not calculated, since the main goal of the study was to show the percentage of stations, for which the Mann-Kendall test indicates a statistically significant trend in the observed time period.

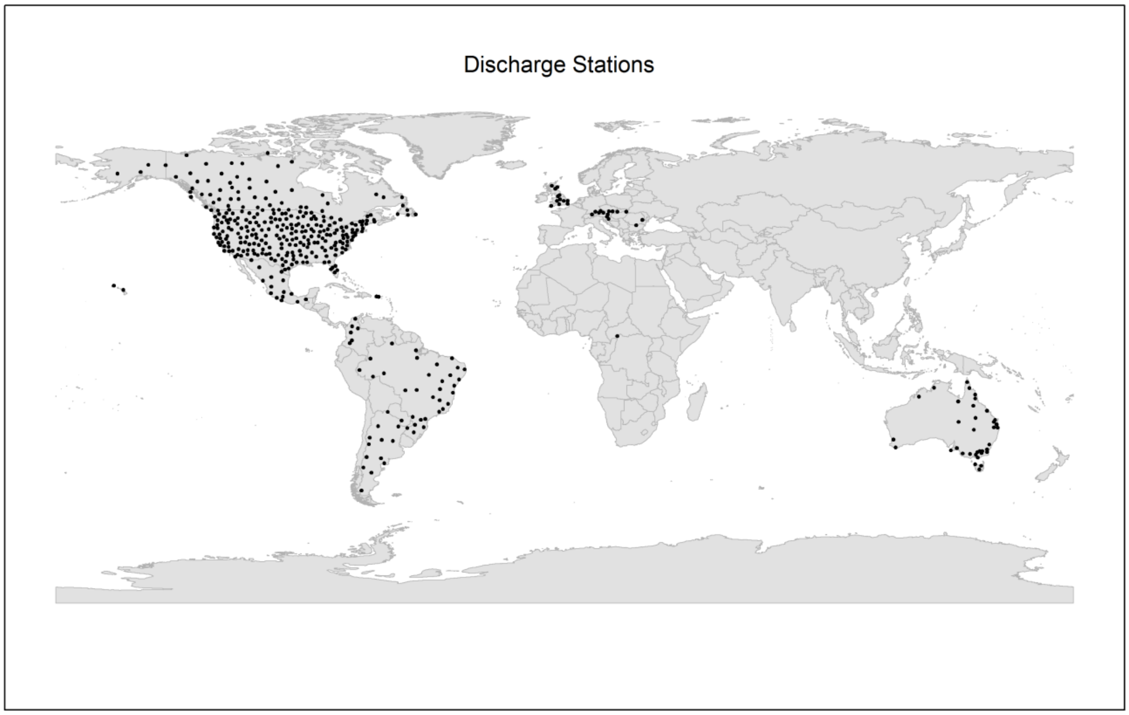

Figure 9. Location of the discharge stations used in trend analysis

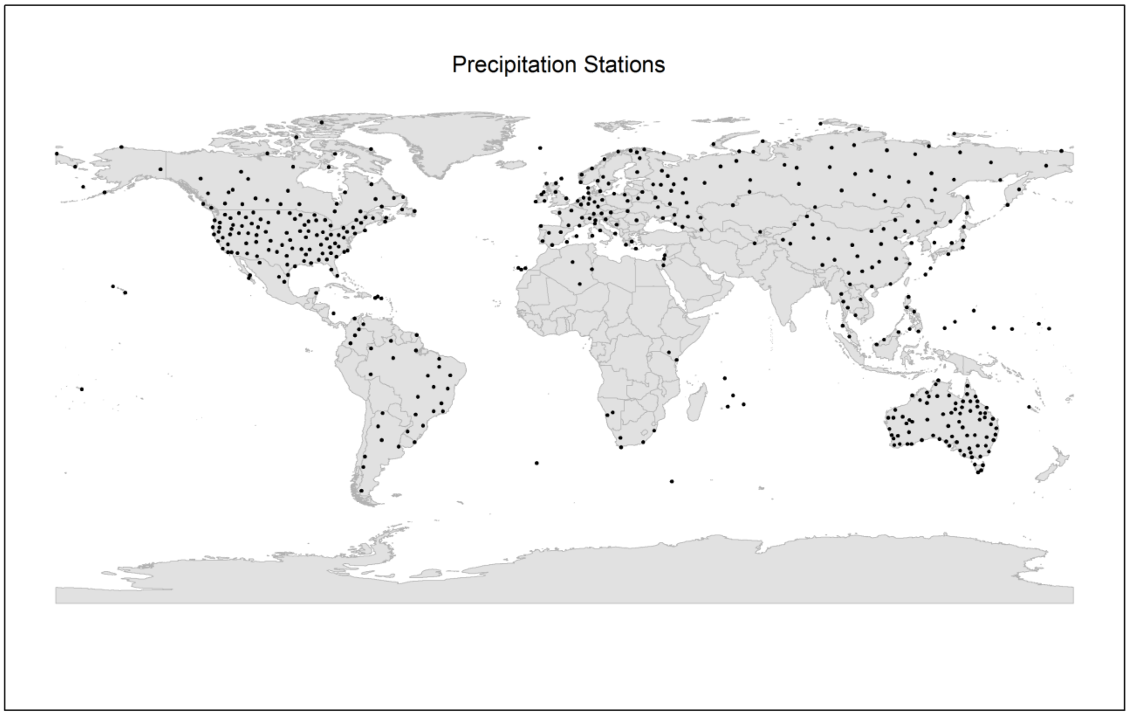

Figure 10. Location of the precipitation stations used in trend analysis

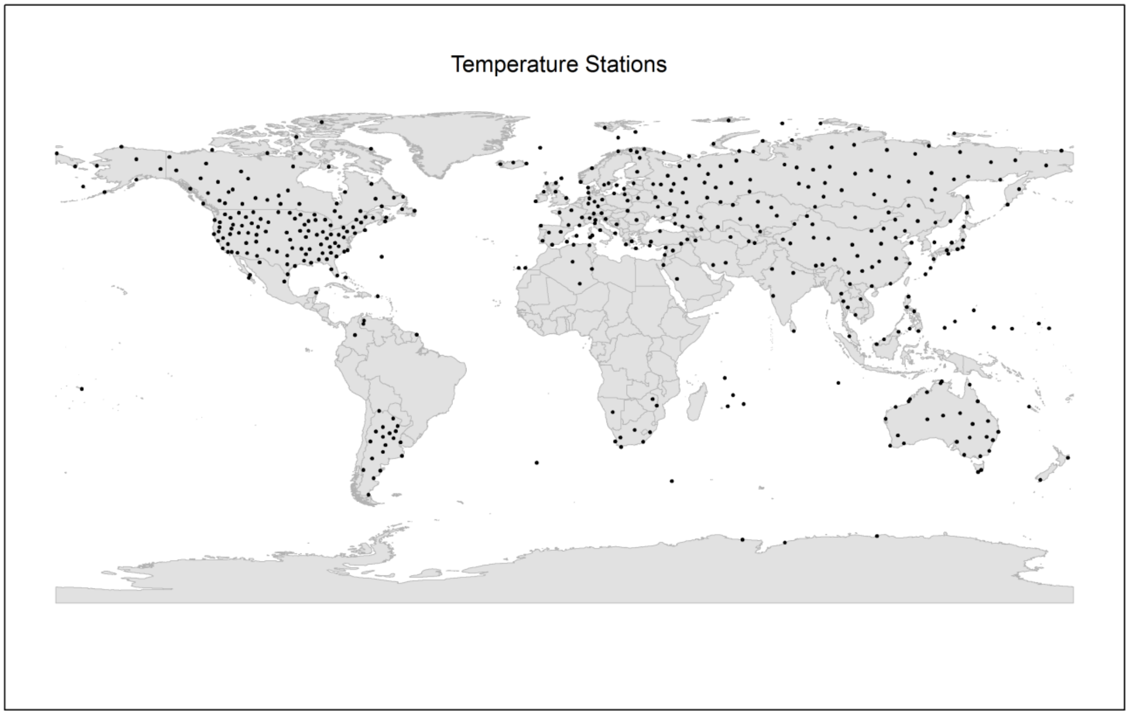

Figure 11. Location of the temperature stations used in trend analysis

3.2 Results of the global trend analysis

In accordance with the previous studies, trend analysis results indicate that for all the variables, statistically significant alterations can be found in a considerable number. Comparing the number of trends between different time resolutions, it can be stated that most of them were observed on the annual level, with decreasing numbers until the daily level. This phenomenon can be clearly seen for temperature series, and also for precipitation, although much fewer trends were observed. The differences between time resolutions for discharge series are hardly visible. In general, a lot of trends were found for discharge and temperature series and very few for precipitation series. Figure 12 on the next page shows the results of trend detection for each of the mean time series, divided into results for the northern and southern hemisphere. This figure shows that an opposing pattern can be found for precipitation between the northern hemisphere, where mainly positive trends are found, and the southern one, where the majority is negative. This can also be seen in Figure 16 later on.

As stated in previous works, it could also be observed that minimum temperature shows most positive trends of all variables, which can also be seen in Figure 13. All of the trends shown in this graphic are statistically significant according to the Mann-Kendall trend test.

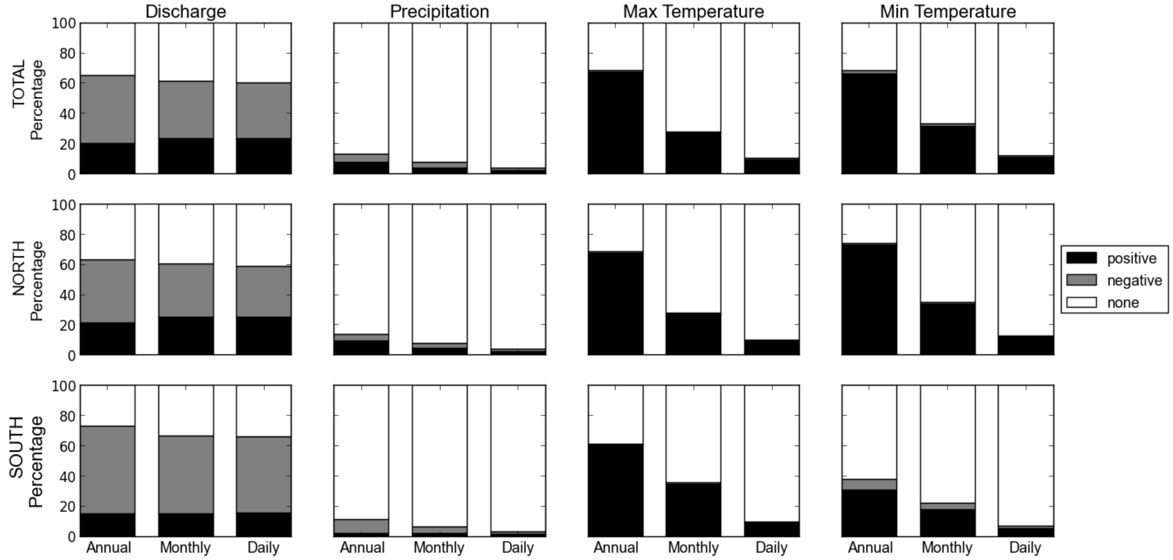

Figure 12. Results showing statistically significant trends for mean values

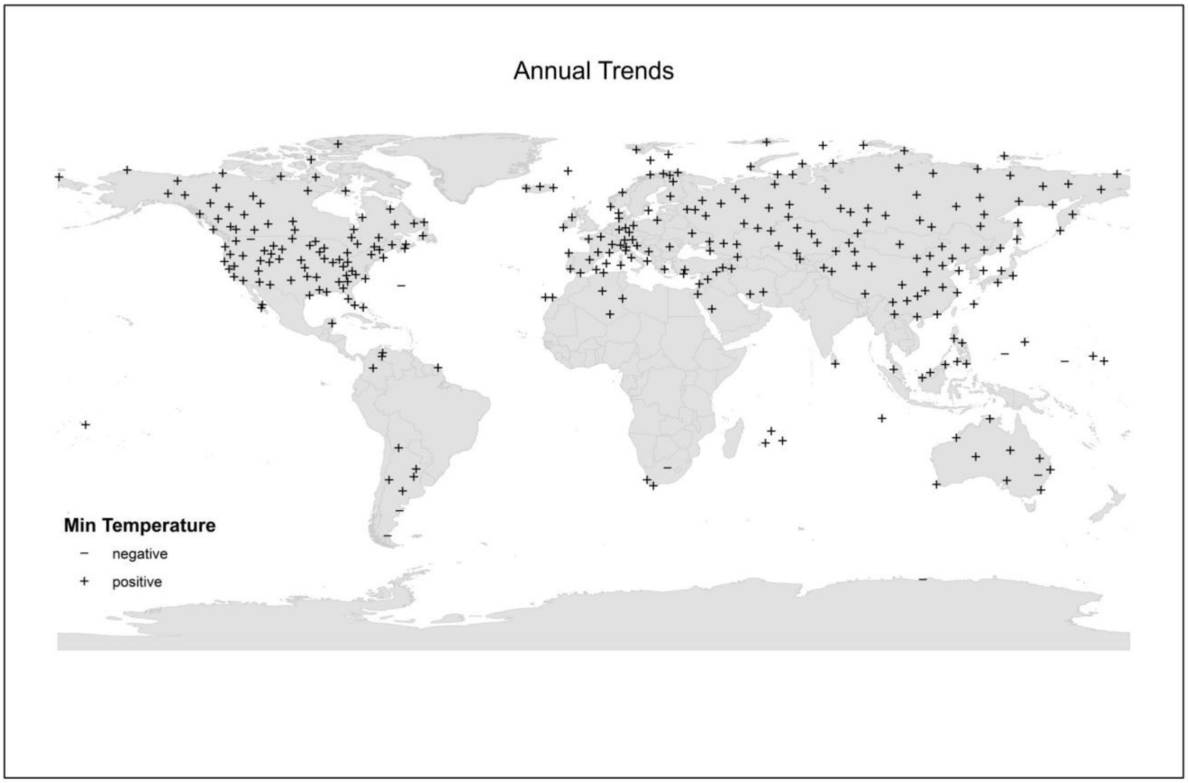

Figure 13. Annual statistically significant trends in minimum temperature

For extreme value series, fewer trends were found although generally the same tendencies as in the trends in mean time series can be observed. The results reflect indicate that while discharge minima tend to get more intense, maxima are becoming less intense. The opposite is found for temperature extremes where especially minima tend to become less intense. Very few trends were found in precipitation maxima, which principally include more evidence of intensifying events in the northern and more weakening ones in the southern hemisphere.

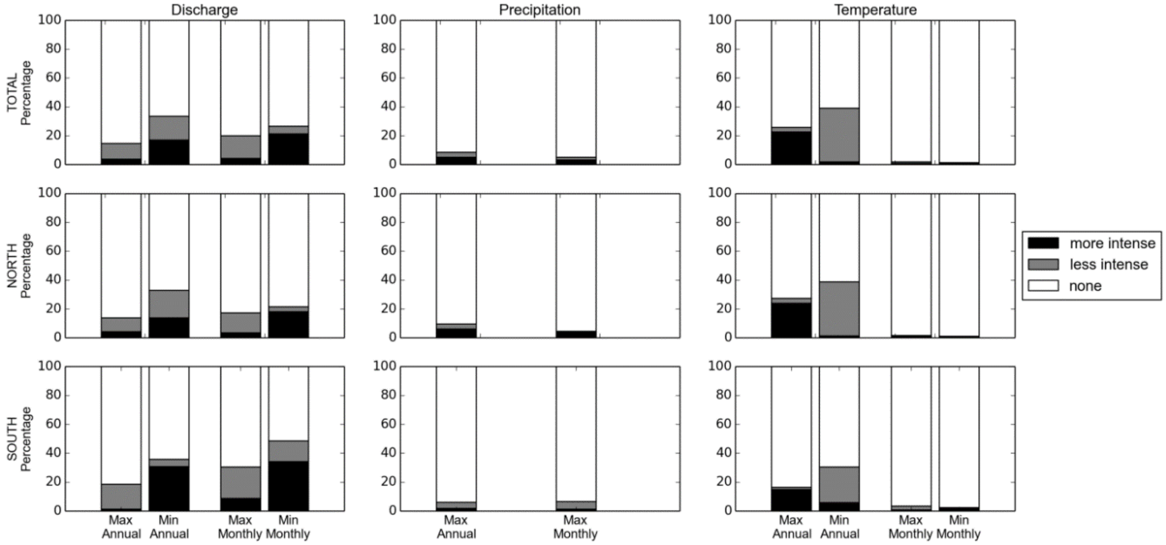

Figure 14. Results showing statistically significant trends for extreme values

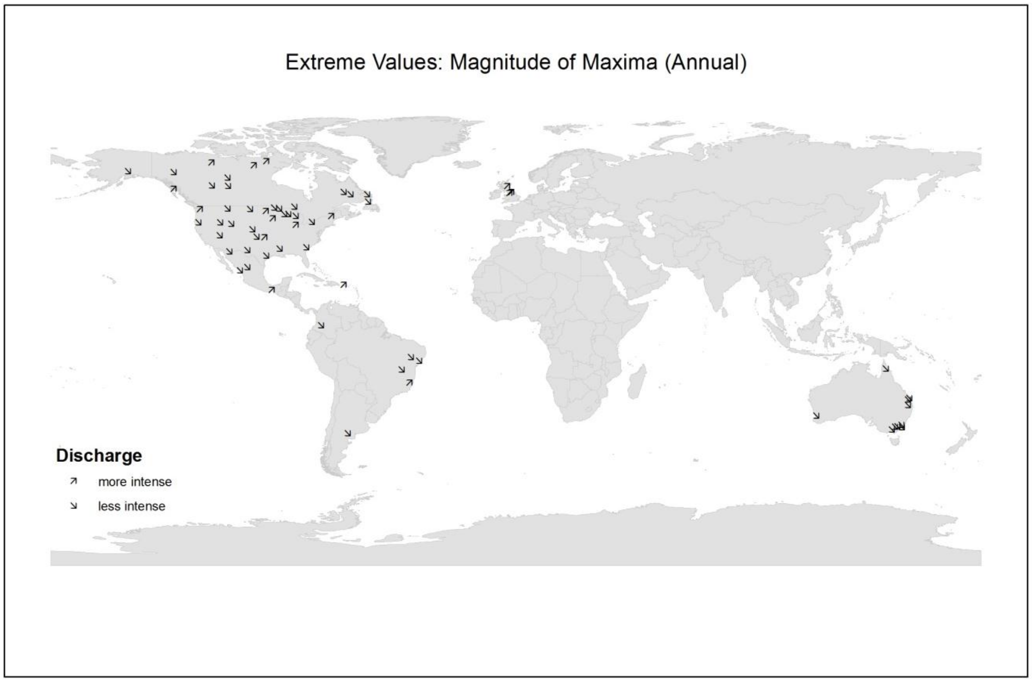

Figure 15. Trends in discharge maxima

While the majority of the findings in trend analysis is in accordance with previous findings, the additional fact not mentioned in other studies is the existence of opposing patterns in the precipitation trends in northern and southern hemisphere, since few previous studies conducted a global sample of precipitation stations. Mainly negative trends, as well as more intensifying minima and weakening maxima in discharge series and at the same time very few trends in precipitation could lead to the conclusion that human activities influence on the flow regimes of most of the rivers worldwide, although this was not specifically proven by the study.

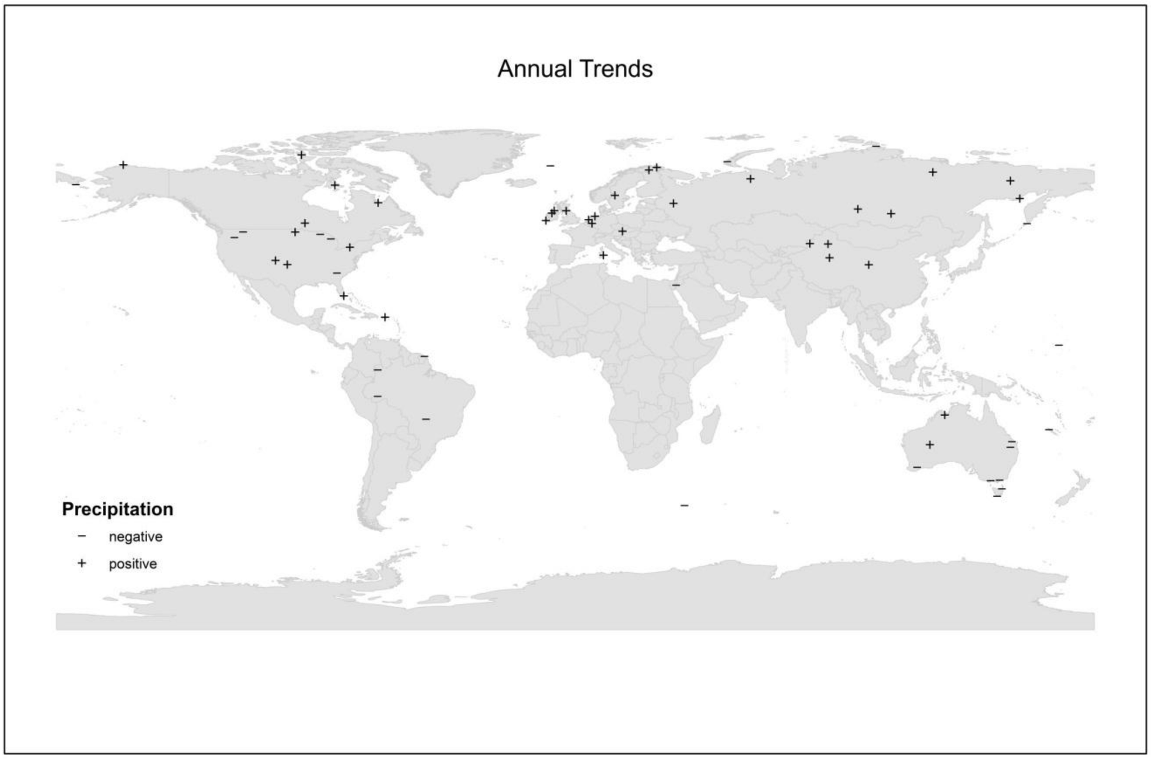

Figure 16. Trends in annual precipitation

Another finding that has not been reported before is the difference in the number of trends detected on different time resolutions. This can be noted especially for temperature series, where a strong decrease in the number of observed trends from annual to daily resolution can be seen. While a similar pattern is present for precipitation series, the percentage of trends for discharge series is constant over all time resolutions. As explained before by various authors (Morin, 2011; Yue et al., 2002), the variability of the data can influence on the ability of the Mann-Kendall test to produce Type I and Type II errors. The power of the test to correctly detect a trend is described to be a decreasing function of the coefficient of variation of the time series. For a sample size of approximately 40 years, the power of the test rapidly decreases at a coefficient of variation of approximately 0.2 (Yue et al., 2002).

In an additional test, it was found that temperature series show a coefficient of variation of below that limit on the annual level and above or close to it for the other resolutions. This fact could be one of the possible explanations of the decrease in the number of trends detected. The coefficient of variation of discharge data is always far above the 0.2 level and therefore could explain the constant number of trends detected. The histograms for the number of trends in percentage of the total number of tests are shown in Figure 17.

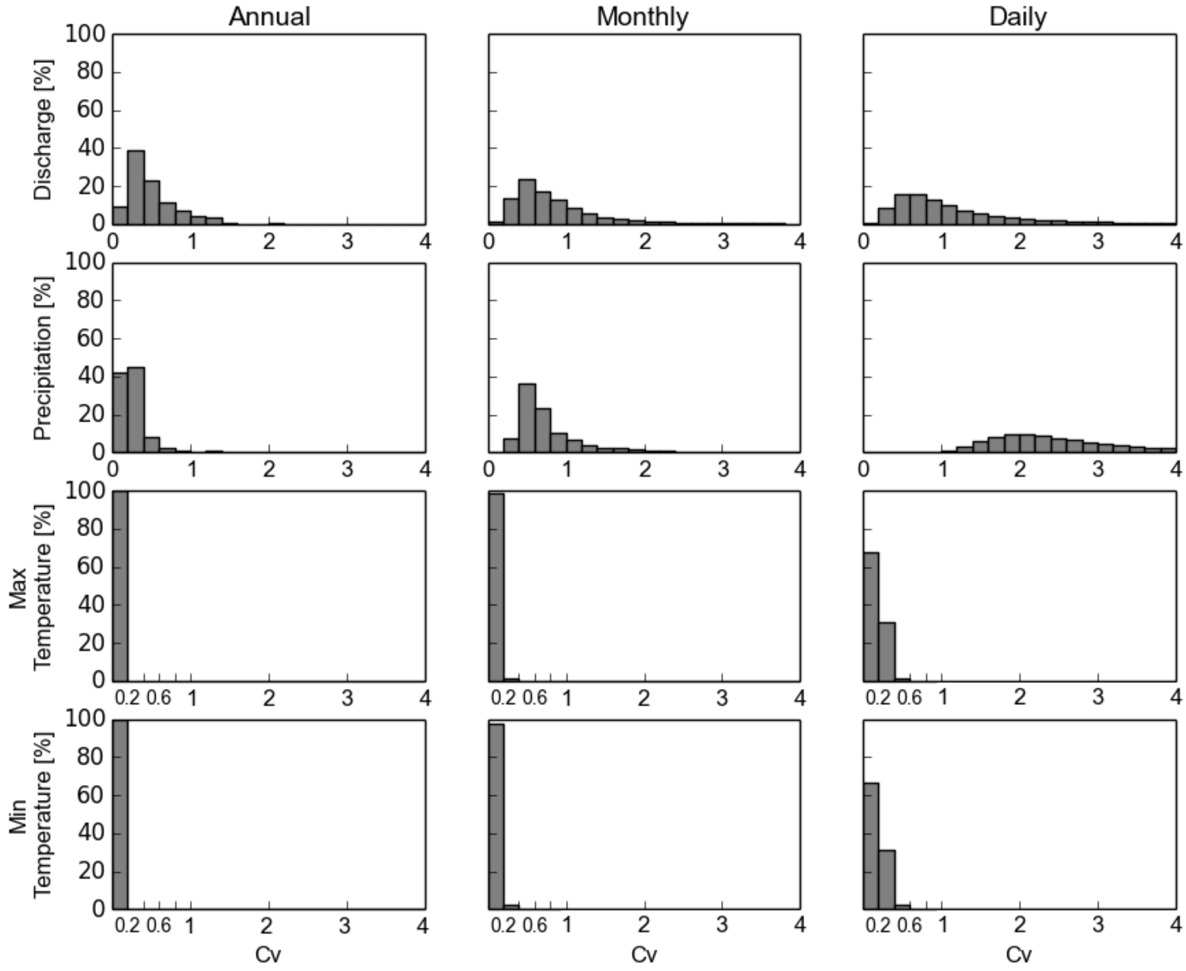

Figure 17. Histograms of coefficients of variation for different variables

From the study and previous findings it can be concluded that statistically significant trends can be found for all variables on all time scales during the 41 years taken into account for this analysis. As stated in Bordi et al. (2009), "due to the shortness of the time records […], it is difficult to objectively estimate trends and their statistical significance, as well as to discern between linear trend and long-term periodicity". The 41 years of this investigation were too short to make general conclusions about change in global climate, but the results most definitely give an indication about how the climate may continue behaving in the next 10 to 15 years, especially the overwhelming amount of stations with rising temperature and the predominantly negative trends in discharge and its extremes, while precipitation trends hardly exist.

However, because of the above stated problems and also the fact that a difference in numbers of trends indicate a different statistical behavior on different time resolutions, the results of this analysis led to the conclusion that it could not be considered to sufficiently describe the change in the regimes of extreme events only with linear trends. For this reason, a more profound statistical analysis had to be conducted that uses stochastic methods to analyze the existing mechanisms that influence this behavior and might cause possible alterations.

SECTION 4

STOCHASTIC MODEL APPROACH

As shown in the results of trend analysis, it is unfeasible to conclude about the changes in extreme event regimes from trends in short hydrometeorological time series. Therefore, in this section a model is proposed to describe the ensembles that represent the extreme events of discharge time series in depth with a solid statistical base, which was later applied to the test basins presented in Section 2.6. This model is based on a stochastic differential equation, in order to represent the evolution of changes of statistical properties in time and encompass also the principles of uncertainty in the form of fluctuations around a deterministic kernel.

Since a lot of data is available from the observations of hydrometeorological variables, but the physical parameters of such systems are largely unknown, an inverse modeling approach is proposed in this work. With the measurements of discharge data and those variables that influence the river basin, precipitation and temperature, a set of parameters is developed that is afterwards linked to the physical parameters of the basin. The outline of works realized related to the development of the model are shown in Figure 18.

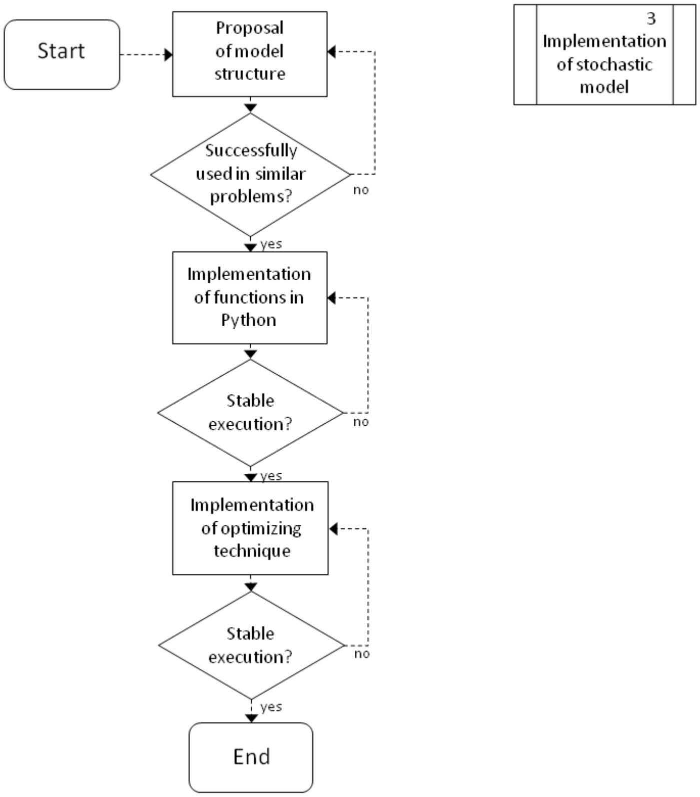
Start
Proposal of model structure
Successfully used in similar problems?
no
yes
Implementation of functions in Python
Stable execution?
no
yes
Implementation of optimizing technique
Stable execution?
no
yes
End
3
Implementation of stochastic model

Figure 18. Schematic outline of the development of the stochastic model

4.1 Stochastic Hydrological Modeling

4.1.1 Complex systems

When modeling natural systems, it is important to bear in mind the nature of these systems. Chaitin (1969) proposed, based on previous works by Kolmogorov (1965) and Solomonoff (1964), the concept of algorithmic complexity, which is also known as Kolmogorov-Chaitin complexity.

Algorithmic complexity describes the measure of how much computation resources are necessary to describe a system. A system that describes the swing of a pendulum can be easily described because of repeated patterns and therefore shows minimum algorithmic complexity. On the other side, the result of the throw of a die is completely random and therefore presents maximum complexity. The throw of the die can certainly be modeled, but only with a more complex algorithm and requires the full transmission of the system states in order to describe the evolution of the system making it impossible to describe the system at any level of compression (Gell-Mann, 1995).

Hydrological systems show neither maximum nor minimum algorithmic complexity, but could be described as of significant algorithmic complexity. In the majority of cases, these systems can be described by a deterministic model, but this kind of model does not represent the system in a truly ideal way. In a hydrological system, as in almost all other natural systems as well, fluctuations are present that are caused by the chaotic nature of the system, and therefore do not allow for purely deterministic descriptions. Therefore, it is impossible to model or predict a certain outcome with complete certainty and it is necessary to provide a probability associated to all possible outcomes as a measure of accuracy of the model.

4.1.2 The Langevin Equation

Paul Langevin originally developed the equation eventually named after him to describe the motion *v* of a Brownian particle in a liquid,

$$\dot{v} = -\gamma v + L(t), \tag{27}$$

including the notions of a constant term γ damping the particle and an added irregular and unpredictable motion *L(t)*, which has simple averaged properties (Denisov et al., 2009).

In many previous scientific works, the Langevin equation has been applied to describe the stochastic dynamics of systems with a fluctuating environment, which describes very well the above mentioned complex systems. The Langevin equation is a random differential equation that describes the stochastic evolution of the observed variable in time.

Because of the mentioned reasons that it combines a deterministic kernel with a modulated noise part, the Langevin equation was chosen to represent the hydrological model used in this study. Applying the equation to hydrological time series, the equation can be rewritten as follows, which is also called the overdamped Langevin equation (Denisov et al., 2009).

$$\frac{d\underline{q}(t)}{dt} = \psi\left[\underline{q}(t), t\right] + \zeta\left[\underline{q}(t), t\right]\xi(t) \tag{28}$$

The constant damping term is replaced by $\psi\left[\underline{q}(t), t\right]$, a deterministic function that describes the kernel of the hydrological system and $\zeta\left[\underline{q}(t), t\right]$ is the deterministic function that modulates the noise $\xi(t)$, which was chosen to be white noise, and represents the fluctuating force of the system. In the equations, discharge is not, as commonly used, denoted as a capital Q in order not to confuse the notions of random variables and random processes.

Stochastic differential equations such as the Langevin equation can be solved using different techniques, for example the Ito calculus, the Stratonovich integral, or the Fokker-Planck-Kolmogorov (FPK) equation (Gardiner, 2004). Since the first two methods are analytical solutions, it is only possible to solve problems of a limited difficulty with them. However, as shown in Dominguez and Rivera (2010), a numerical solution of the Fokker-Planck-Kolmogorov equation is possible that also permits to tackle more difficult problems.

While the Langevin equation is a stochastic differential equation that describes the evolution of an observable in time, the FPK equation is a deterministic equation that describes the evolution of its probability density function in time. The probability density of the solution is one of the most important statistical characteristics of random differential equations. If the noise term of the Langevin function is produced by a noise-generating function, then the probability density is seen as a closed equation that can be represented by the FPK equation (Denisov et al., 2009). Therefore, each Langevin equation that uses a certain type of noise function has its corresponding FPK equation.

The complete derivation of the FPK equation that corresponds to the overdamped Langevin equation is described in detail by Denisov et al. (2009).

4.1.3 The Fokker-Planck-Kolmogorov (FPK) Equation

Also called Forward Kolmogorov Equation, the Fokker-Planck-Kolmogorov equation is based on the works developed by Adriaan Fokker and Max Planck in the early 20th century and its mathematical base was described analytically in depth later by Kolmogorov (Kolmogorov, 1931). This deterministic, partial differential equation is the conservation law used to describe the time evolution of the probability density function of a random process in one or multiple dimensions, containing a drift vector and a diffusion matrix.

In its base, the FPK equation is a diffusion process representing the conditional probability density of two random variables measured at different times, taking the form (Gardiner, 2004)

$$\frac{\partial p\left(\underline{x},t\middle|\underline{y},t'\right)}{\partial t}=-\sum_{i}\frac{\partial}{\partial x_i}\left[A_i(\underline{x},t)p\left(\underline{x},t\middle|\underline{y},t'\right)\right]+\frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_i\partial x_j}\left[B_{ij}(\underline{x},t)p\left(\underline{x},t\middle|\underline{y},t'\right)\right] \qquad (29)$$

where $A_i(\underline{x},t)$ is the drift vector, $B_{ij}(\underline{x},t)$ the diffusion matrix and t' represents a previous time step. This is the multidimensional form of the equation.

In its one-dimensional form, the Fokker-Planck-Kolmogorov equation is used to describe the evolution of probability density during the translation from one time step to the next and takes the form (Dominguez and Rivera, 2010)

$$\frac{\partial p(\underline{q},t)}{\partial t}=-\frac{\partial}{\partial x}\left[A\left(\underline{q},t\right)p\left(\underline{q},t\right)\right]+\frac{1}{2}\frac{\partial^2}{\partial x^2}\left[B\left(\underline{q},t\right)p\left(\underline{q},t\right)\right] \qquad (30)$$

where the diffusion matrix reduces to the diffusion vector $B\left(\underline{q},t\right)$ and $p\left(\underline{q},t\right)$ represents the probabilistic density of the hydrological variable at time t.

The choice of using only the one-dimensional FPK equation and not the multidimensional one was taken after an analysis of the processes' correlation moment. In order not to interrupt the flow of reading, the results of this analysis will be presented after the complete description of the methodology in section 4.1.5.

The drift and diffusion vectors $A\left(\underline{q},t\right)$ and $B\left(\underline{q},t\right)$ are defined by deterministic functions that control the movements of the PDF, where the drift term alone principally describes the sideward movement of the curve and the diffusion term its flattening and sharpening. The simplest form of the influence of these vectors on the probability density curves is displayed in Figure 19. In general terms, the drift term describes the overall comportment of the process that can be used as an approximation of the behavior observed in nature. The diffusion term describes the fluctuation around this approximation in an attempt to more precisely represent the naturally present deviations in the physical process.

The parameters of the functions $A\left(\underline{q},t\right)$ and $B\left(\underline{q},t\right)$ are linked to the parameters of the system that is described and in this case is a river basin (Kovalenko et al., 1993). These parameters can be internal factors, such as the morphometry or the land cover of the basin or external factors, such as precipitation or temperature.

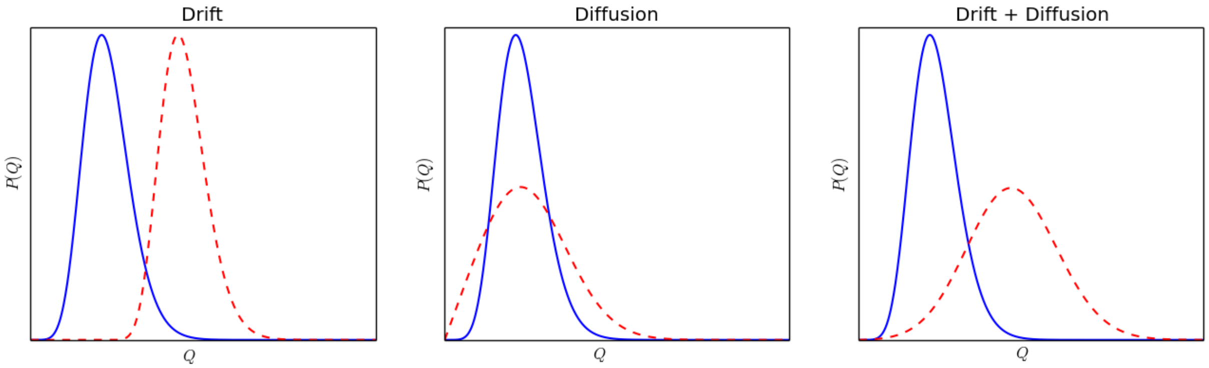

Figure 19. Drift and diffusion of probability density functions

$A\left(\underline{q},t\right)$ and $B\left(\underline{q},t\right)$ were chosen to be high order polynomial functions of the form of

$$A\left(\underline{q},t\right) = k_1\underline{q}^{\alpha_1} + k_2\underline{q}^{\alpha_2} + k_3\underline{q}^{\alpha_3}, \tag{31}$$

$$B\left(\underline{q},t\right) = g_1\underline{q}^{\beta_1} + g_2\underline{q}^{\beta_2} + g_3\underline{q}^{\beta_3} \tag{32}$$

In this study, the main goal was to understand the dynamics of the process, so to obtain an equation that was sufficiently easy to implement while taking into account the linear nature of the drift term and the quadratic one of the diffusion term, the parameters were set to

$$\alpha_1 = \beta_1 = 1;\ \alpha_2 = \beta_2 = 2;\ \alpha_3 = \beta_3 = 0 \tag{33}$$

This resulted in the simplified form of

$$A\left(\underline{q},t\right) = k_1\underline{q} + k_2\underline{q}^2 + k_3, \tag{34}$$

$$B\left(\underline{q},t\right) = g_1\underline{q} + g_2\underline{q}^2 + g_3 \tag{35}$$

4.1.4 *Relationship between Langevin and FPK equation*

The relationship between the parameters of the Langevin and FPK equations can be expressed as (Sveshnikov, 1966)

$$A\left(\underline{q},t\right) = \psi\left[\underline{q}(t),t\right] + \frac{1}{2}\zeta\left[\underline{q}(t),t\right]\frac{\partial\zeta\left(\underline{q},t\right)}{\partial\underline{q}} \tag{36}$$

$$B\left(\underline{q},t\right) = \zeta^2\left[\underline{q}(t),t\right] \tag{37}$$

Since the equation defining the diffusion term was chosen to be $B\left(\underline{q},t\right) = g_1\underline{q} + g_2\underline{q}^2 + g_3$, equation 37 can be rewritten as follows:

$$g_1\underline{q} + g_2\underline{q}^2 + g_3 = \zeta^2\left[\underline{q}(t),t\right] \tag{38}$$

$$\zeta\left[\underline{q}(t),t\right] = \sqrt{g_1\underline{q} + g_2\underline{q}^2 + g_3} \tag{39}$$

The same way, equation 36 becomes

$$k_1\underline{q} + k_2\underline{q}^2 + k_3 = \psi\left[\underline{q}(t),t\right] + \frac{1}{2}\zeta\left[\underline{q}(t),t\right]\frac{\partial\zeta\left[\underline{q}(t),t\right]}{\partial q} \tag{40}$$

Equation 39 can be employed into equation 40 to obtain

$$\begin{aligned} k_1\underline{q} + k_2\underline{q}^2 + k_3 = \psi\left[\underline{q}(t),t\right] \\ + \frac{1}{4}\left(g_1\underline{q} + g_2\underline{q}^2 + g_3\right)^{\frac{1}{2}}\left(g_1\underline{q} + g_2\underline{q}^2 + g_3\right)^{-\frac{1}{2}}\left(g_1 + 2g_2\underline{q}\right) \end{aligned} \tag{41}$$

which can be simplified as

$$\psi\left[\underline{q}(t),t\right] = k_1\underline{q} + k_2\underline{q}^2 + k_3 - \frac{1}{4}\left(g_1 + 2g_2\underline{q}\right) \tag{42}$$

Therefore, the Langevin equation in terms of the parameters of the drift and diffusion equations takes the following form

$$\frac{d\underline{q}(t)}{dt} = k_1\underline{q} + k_2\underline{q}^2 + k_3 - \frac{1}{4}\left(g_1 + 2g_2\underline{q}\right) + \sqrt{g_1\underline{q} + g_2\underline{q}^2 + g_3\xi(t)} \tag{43}$$

4.1.5 Determination of Markov process structure

In order to take a decision whether it is sufficient to use the one-dimensional form of the Fokker-Planck-Kolmogorov equation or if it is necessary to use its higher-dimensional form, it had to be determined if the processes could be assumed to have a Markovian structure (Kovalenko et al., 1993). Therefore, the cross correlation function of the processes representing the discharge data for monthly maxima and minima was used to evaluate if a lag-one cross correlation could be proven. For this purpose, the cross correlation of the random processes was determined and compared with the standard error of the autocorrelation coefficient, as proposed in Druzhinin and Sikan (2001). The critical autocorrelation radius of the process is the x-value of the cross correlation function, at which it intersects the standard error function of the autocorrelation coefficient. If this lag value lies between 1 and 2, the structure of the process can be assumed to be a lag-one correlation and therefore it is valid to use the one-dimensional FPK equation, since no other random variable of a higher lag has a statistically significant influence.

The standard error function was constructed as described in Druzhinin and Sikan (2001),

$$e = \frac{1 - r^2}{\sqrt{n-1}} t_{1-\alpha/_2} \tag{44}$$

where n is the total number of observations in each random variable and t is the critical value of the Students t distribution with significance level α, which was chosen to be 5%.

Although the results of the analysis indicate that the processes of minima show a higher critical autocorrelation radius than those of maxima, both processes could still be assumed to have a Markovian structure.

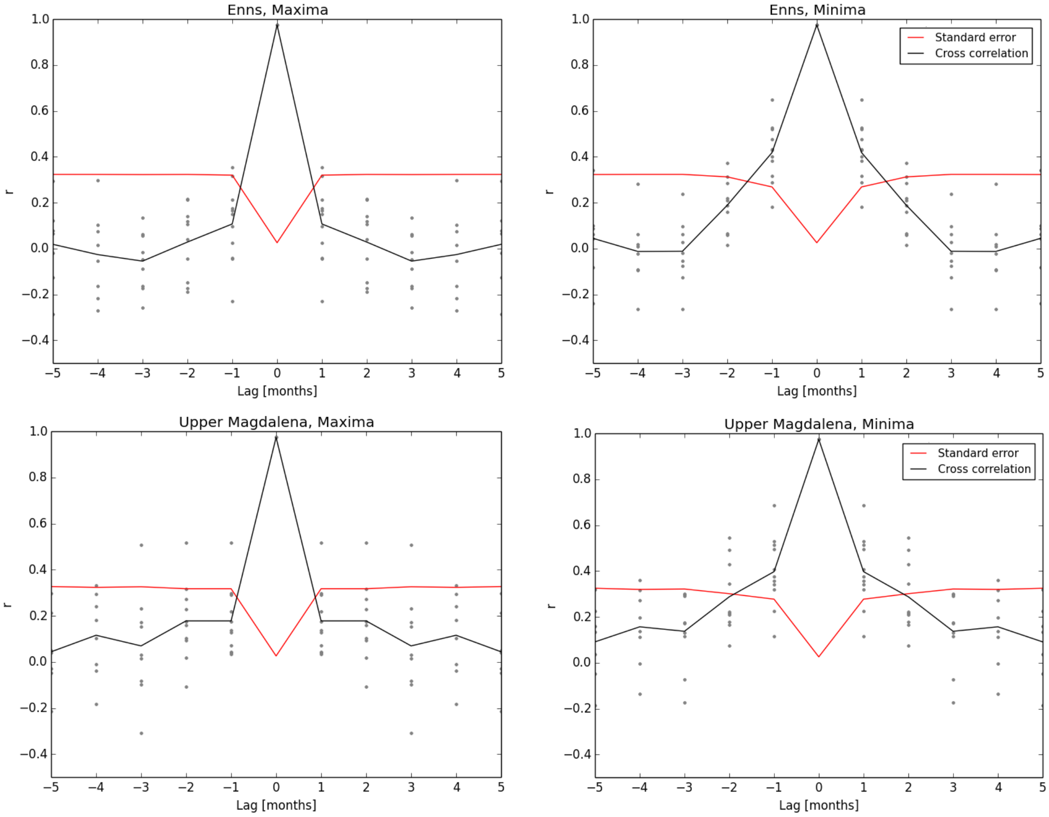

Figure 20. Determination of the processes' cross correlation in the Enns and Upper Magdalena basins

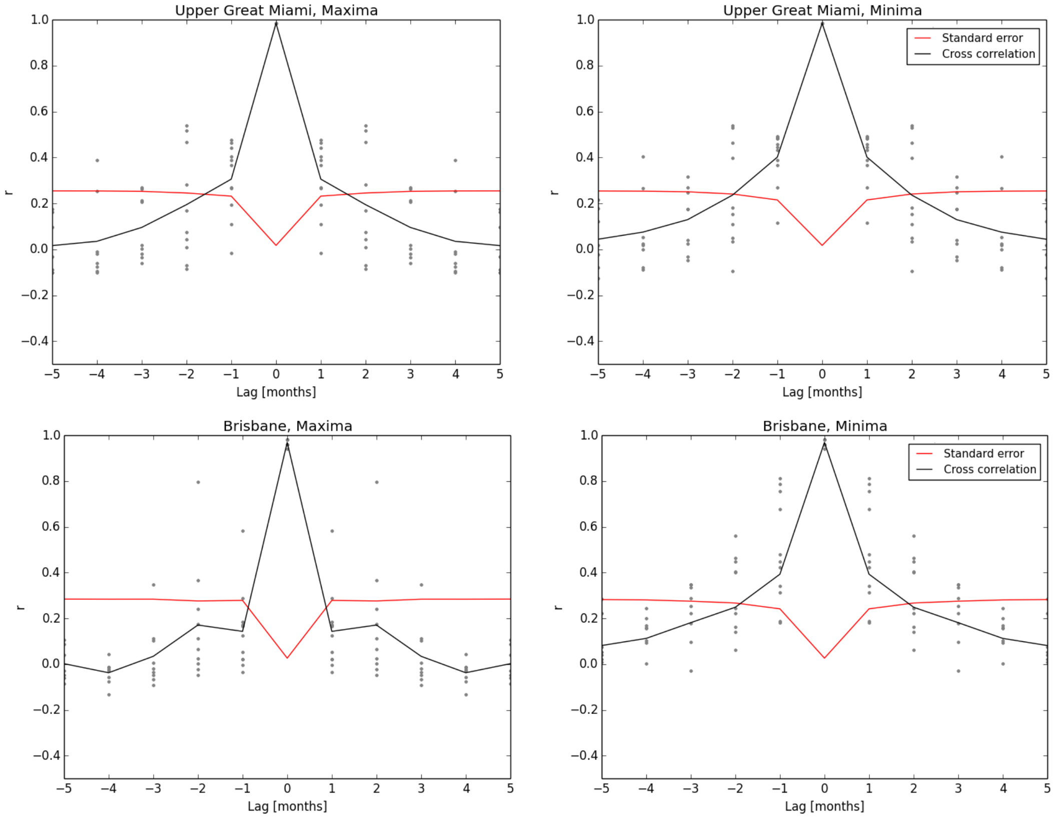

Figure 21. Determination of the processes' cross correlation in the Upper Great Miami and Brisbane basins

The choice between the two candidate basins in the southern hemisphere, the Brisbane and Crocodile River basins, was made due to its process structure. Since data from the Crocodile River did not indicate a Markovian structure, this basin was not included and the Brisbane basin was preferred, despite its location downstream of a dam.

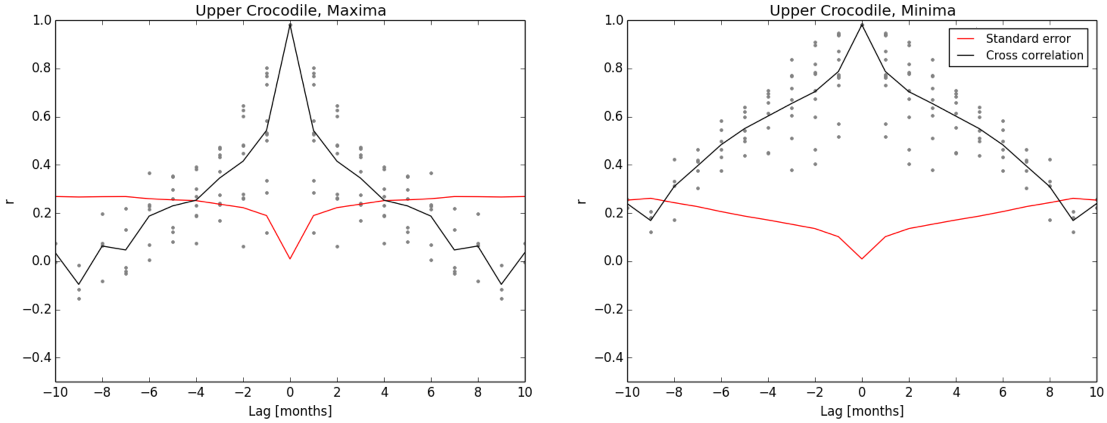

Figure 22. Determination of the processes' cross correlation for the Crocodile River basin

4.2 Implementation of the Fokker-Planck-Kolmogorov Equation

4.2.1 Finite-difference system

For the implementation of the one-dimensional Fokker-Planck-Kolmogorov (FPK) equation, two different numerical models were established. An explicit and an implicit scheme were developed and implemented in close accordance with the method proposed by Dominguez and Rivera (2010).

As stated in Dominguez and Rivera (2010), an analytical solution to the non-stationary FPK equation needs strong restrictions on the types of drift and diffusion coefficients, which makes it not as convenient as the numerical solution. Therefore, a numerical solution to the equation is proposed that includes a bidirectional approach for the drift term, which enables a drift in both directions. Equation 45 shows the finite difference approximation that includes directional weights that permit both the backward and forward solutions and a time layer weight that controls if the equation is solved explicitly or implicitly.

$$
\begin{aligned}
\frac{p_j^{i+1}-p_j^i}{\Delta t} = -\Bigg\{ & \sigma\left[\frac{\varphi_L\left(A_{j+1}^{i+1}p_{j+1}^{i+1} - A_j^{i+1}p_j^{i+1}\right)}{\Delta q} + \frac{\varphi_R\left(A_j^{i+1}p_j^{i+1} - A_{j-1}^{i+1}p_{j-1}^{i+1}\right)}{\Delta q}\right] \\
& + (1-\sigma)\left[\frac{\varphi_L\left(A_{j+1}^{i}p_{j+1}^{i} - A_j^{i}p_j^{i}\right)}{\Delta q} + \frac{\varphi_R\left(A_j^{i}p_j^{i} - A_{j-1}^{i}p_{j-1}^{i}\right)}{\Delta q}\right] \\
& + \frac{\sigma}{2}\left[\frac{\left(B_{j+1}^{i+1}p_{j+1}^{i+1} - 2B_j^{i+1}p_j^{i+1} + B_{j-1}^{i+1}p_{j-1}^{i+1}\right)}{\Delta q^2}\right] \\
& + \frac{(1-\sigma)}{2}\left[\frac{\left(B_{j+1}^{i}p_{j+1}^{i} - 2B_j^{i}p_j^{i} + B_{j-1}^{i}p_{j-1}^{i}\right)}{\Delta q^2}\right]\Bigg\}
\end{aligned}
\qquad (45)
$$

where

j interval descriptor for discharge

i interval descriptor for time step

φ_L, φ_R directional weights for the bidirectional drift, if $A_j^i < 0, \varphi_R = 0, \varphi_L = 1$ and otherwise $\varphi_R = 0, \varphi_L = 1$

σ time layer weight implementing the numerical scheme totally implicitly when equal to 1 and totally explicitly when equal to 0

In order to execute the model stably, a stability condition is required. The dimensionless Peclet number gives the ratio between the drift and diffusion component along a characteristic length, which is in this case the step size of *q*, *Δ*q. Since the drift and diffusion are defined by vectors and not numbers, the respective maximum value of the vector was used in the calculation of the Peclet number.

$$Pe = \frac{max\left[\left\|A\left(\underline{q},t\right)\right\|\right]\Delta q}{max\left[\left\|B\left(\underline{q},t\right)\right\|\right]} \tag{46}$$

In those cases, in which $Pe \leq 3$, a Courant-Friedrichs-Lewy stability condition was defined (Dominguez and Rivera, 2010)

$$max\left[\left\|B\left(\underline{q},t\right)\right\|\right]\frac{\Delta t}{\Delta q^2} < \frac{1}{2} \tag{47}$$

In the rest of the cases, the complete diffusion term was neglected, because in this case the diffusion would not be noticeable anymore. The step size for time was defined by

$$\frac{\Delta q}{\Delta t} > max\left[\left\|A\left(\underline{q},t\right)\right\|\right] \tag{48}$$

The condition established in equation 48 was merely used formally. Since the diffusion component produces much higher values for the resulting vector B than the drift vector A, it was not applied in the models, but served for the complete implementation of the model. The difference in magnitude of the values can be shown with some random example values obtained by the implementation of the model in two of the test basins in Table 5.

Comparison of values of Drift and Diffusion components					
Upper Great Miami Basin			Upper Magdalena Basin		
Q	A	B	Q	A	B
77.59	8.12	1075.62	1284.39	-324.79	-68199.12
77.93	8.00	1085.23	1286.64	-321.82	-68620.99
78.26	7.87	1094.87	1288.89	-318.84	-69043.47
78.60	7.74	1104.56	1291.15	-315.84	-69466.55
78.93	7.62	1114.29	1293.40	-312.83	-69890.25
79.26	7.49	1124.07	1295.65	-309.80	-70314.54
79.60	7.36	1133.88	1297.90	-306.76	-70739.45
79.93	7.23	1143.74	1300.15	-303.71	-71164.96
80.27	7.09	1153.64	1302.40	-300.63	-71591.07
80.60	6.96	1163.59	1304.66	-297.55	-72017.79
80.94	6.82	1173.58	1306.91	-294.44	-72445.12
81.27	6.69	1183.61	1309.16	-291.33	-72873.06
81.61	6.55	1193.68	1311.41	-288.19	-73301.60
81.94	6.41	1203.80	1313.66	-285.05	-73730.75
82.27	6.27	1213.95	1315.91	-281.88	-74160.50

Table 5. Comparison of the magnitudes of values resulting from drift (A) and diffusion (B) components

For the implementation of the explicit method, equation 45 was used with $\sigma = 0$, which cancels out its first and third line. The number of time steps Δt is defined by the above mentioned stability conditions, and therefore the number of iterations repeated in every calculation is *1 / Δt*.

For a large number of cases, this version proved to be a sufficient solution, but in other cases, especially those presenting negative diffusion terms, it was necessary to fully solve the equation, providing both implicit and explicit components. Therefore, as proposed in Dominguez and Rivera (2010) equation 45 is rewritten as

$$\xi_j^{i+1} p_{j-1}^{i+1} + \psi_j^{i+1} p_j^{i+1} + \gamma_j^{i+1} p_{j+1}^{i+1} = R_j^i \tag{49}$$

with the four variables defined as

$$\xi_j^{i+1} = -\omega_2 A_{j-1}^{i+1} - \omega_5 B_{j-1}^{i+1} \tag{50a}$$

$$\psi_j^{i+1} = \omega_2 A_j^{i+1} - \omega_1 A_j^{i+1} + \omega_5 B_j^{i+1} + 1 \tag{50b}$$

$$\gamma_j^{i+1} = \omega_1 A_{j+1}^{i+1} - \omega_5 B_{j+1}^{i+1} \tag{50c}$$

$$\begin{aligned} R_j^i = {} & -p_j^i + \omega_3 A_{j+1}^i p_{j+1}^i - \omega_3 A_j^i p_j^i + \omega_4 A_j^i p_j^i - \omega_4 A_{j-1}^i p_{j-1}^i \\ & - \omega_6 B_{j+1}^i p_{j+1}^i + 2\omega_6 B_j^i p_j^i - \omega_6 B_{j-1}^i p_{j-1}^i \end{aligned} \tag{50d}$$

where

$$\omega_1 = \frac{\sigma \varphi_L \Delta t}{\Delta q} \tag{51a}$$

$$\omega_2 = \frac{\sigma \varphi_R \Delta t}{\Delta q} \tag{51b}$$

$$\omega_3 = \frac{(1-\sigma) \varphi_L \Delta t}{\Delta q} \tag{51c}$$

$$\omega_4 = \frac{(1-\sigma) \varphi_R \Delta t}{\Delta q} \tag{51d}$$

$$\omega_5 = \frac{\sigma \Delta t}{2\Delta q^2} \tag{51e}$$

$$\omega_6 = \frac{(1-\sigma) \Delta t}{2\Delta q^2} \tag{51f}$$

The tridiagonal matrix that results from the algebraic system in equation 49 was solved implementing the Thomas algorithm as described in 2.3.3.

Each model run represented a transition from one month to the next, where the initial condition was the probability density function of the parting month, from which the translation started to predict the probability of the following month.

The boundary conditions of the model were chosen to be of the type of absorbing barrier. For this kind of condition, it is assumed that any value that is at or outside the boundary values of the PDF has a probability of 0 (Gardiner, 2004).

$$p\left(\underline{q}, t\right) = 0, \qquad if\ Q \epsilon \{\alpha, \beta\} \tag{52}$$

where α and β are the boundary values.

With this kind of boundary condition it is easy that some of the probability contained by each PDF is lost at the boundaries during a translation from one time step to the next as shown in Figure 23. In order to conserve this probability, the lower and upper boundary values have to be chosen far enough apart to include the whole area contained under the curve. For each of the random processes, the boundary values were determined according to the extension of its PDF.

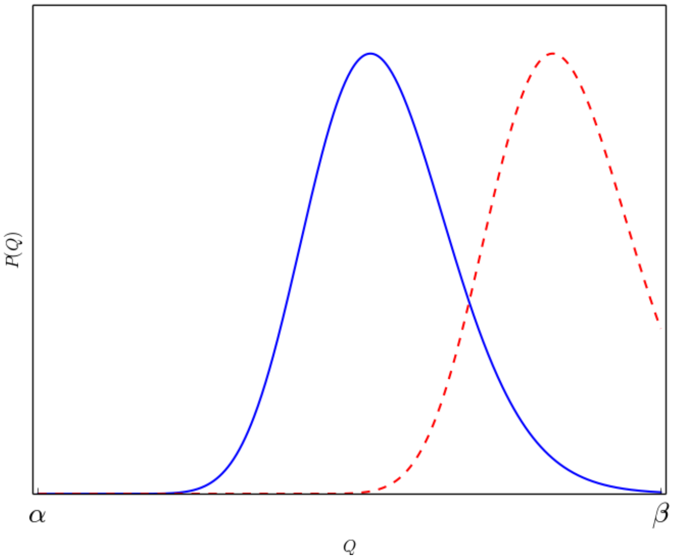

Figure 23. Scheme of an absorbing boundary

As mentioned before, the number of iterations needed to complete the calculation of one translation of probability distribution from one time step to the next using the explicit scheme depends completely on the size of *Δt*. Therefore, also the execution time of the algorithm depends on this variable and can take more or less time. On the contrary, the implicit method always consists of the execution of the Thomas algorithm function to solve the linear system and therefore the only difference in computation time is the size of the boundary interval of the initial condition. In almost all of the translations, the implicit scheme was therefore faster than the explicit one.

The algorithm was implemented in Python in a way that it can easily be used for different applications, but using the same core functionality of the Fokker-Planck-Kolmogorov equation. It consists of various different functions that control the use of the explicit and implicit methods and the calculation of the drift and diffusion vectors. This way, each module can be changed or extended easily, or additional functions can be implemented to complement the existing functionality if this is required. This way, the method of defining the diffusion and drift vectors of the FPK equation using the noise intensities of the basin's parameters proposed by Kovalenko (1993) and applied by Dominguez and Rivera (2010), which was not used in this study, can be implemented by a simple additional function and a change of method to calculate the values of these vectors.

The original code of the explicit method was created looping through all the nodes of the finite differences system with for-loops to calculate each value. This approach, however, was rather time-consuming, especially for translations that required a smaller value of Δt to execute stably. The computing time could be significantly improved by the use of Numpy slices ("Performance Python: Solving The 2D Diffusion Equation With numpy | t-square," 2012), which are parts of matrices temporarily saved in memory for rapid access, and afterwards applying vector multiplications. This way, the execution time of the functions could be improved by approximately 100 times.

4.2.2 Optimization of model parameters

An inverse modeling approach was used to implement the FPK equation model where each of its parameters was optimized. For this purpose, a function was needed to find the optimal model parameters that define the equation for any given transformation of the probability density of one time step to that of the next one. It is important to highlight, that the optimization process did not optimize the fit of probability density functions to the data, but the parameters of the deterministic drift and diffusion equations that describe the transformation of probability density.

For this task, *SciPy*'s optimize package was consulted, which offers a large number of different optimizing functions. A function previously used successfully for optimization purposes is the *curve_fit* function, which had proved to work very well for a variety of different optimization problems. Additionally, error-minimizing algorithms were analyzed, as well as the simulated annealing technique ("scipy.optimize.anneal — SciPy v0.14.0 Reference Guide," 2014).

The *minimize* function minimizes a given scalar function and finds the optimal parameters using a variety of different solving algorithms, out of which the Nelder-Mead, the Powell, Broyden, Fletcher, Goldfarb, and Shanno (BFGS), and the Newton Conjugate Gradient methods were tested ("scipy.optimize.minimize — SciPy v0.14.0 Reference Guide," 2014).

The optimization was tested on 2 different computers with the goal of evaluating the efficiency of the algorithms and the quality of the results. One was a laptop with a Pentium Dual Core CPU with 2.1GHz and 4GB RAM, and the other one desktop computer with an Intel Core i5 CPU with 3.1GHz and 8GB RAM. Both computers executed the optimization on a Windows 7 operating system.

The only function that was able to find parameters that provided a result with less than 50% of mean average error was the *curve_fit* function. All others failed to reach this criterion or took more than one hour to execute for the explicit method. Due to the above mentioned dependency on the model parameters to define the stability criteria in the explicit method, it was necessary to limit the computation time, because otherwise a calibration of the model would have been too time consuming. Since the other tested functions did not provide satisfying results and were therefore not used in the investigation, they are not described in detail.

The curve fitting algorithm ("scipy.optimize.curve_fit — SciPy v0.13.0 Reference Guide," 2014) was chosen to optimize the parameters used to define functions of $A\left(\underline{q},t\right)$ and $B\left(\underline{q},t\right)$. This function uses a nonlinear least squares technique to fit a function to the passed data. To solve this technique, a Levenberg-Marquardt algorithm is implemented in the *curve_fit* function, which is an algorithm commonly used for solving this kind of problem. The *curve_fit* function uses as input parameters the function to be optimized, which in this case is the created Fokker-Planck-Kolmogorov algorithm, the initial data, which is the probability density function of the initial month, and the expected result of the optimization, which is the PDF of the final month. Furthermore, an initial guess of the model parameters to be optimized is passed to the function. Therefore the optimization process was run 12 times for each test area, once for each transition from one month to the next.

The *curve_fit* algorithm includes an exit condition that terminates the procedure after a number of iterations passes without finding the optimal solution. This number is determined by the number of parameters n_p to be optimized and calculated as

$$200 * \left(n_p + 1\right) \qquad (53)$$

SECTION 5

DATA PREPARATION AND RESULTS

In this section, the proposed model is applied to the data of the 4 presented test basins. In a first step, the underlying data structure is presented. Afterwards, the results of the optimization of model parameters is presented, where both an unsupervised and a supervised optimization strategy will be used. After a relation between the model parameters and the physical parameters of the basin are established as described in Section 5.3, the model will be applied to estimate the change in the regimes of hydrometeorological extreme events in future scenarios.

5.1 Initial conditions

In a first step, three monthly random processes were constructed for each test basin's outlet discharge stations from the daily data obtained:

- Monthly means: The mean of all existing values of each month were averaged if a month contained more than 70% of data, otherwise the value for the monthly mean was considered to be missing to ensure a representative mean value.
- Monthly maxima: The month's maximum day, in other words the highest among all observed daily mean values in each month was taken as the monthly maximum.
- Monthly minima: Likewise, the lowest value among all daily means was used as the monthly minimum.

All three time series were rearranged in the form of a stochastic process and indexed in time from January to December. After revising the randomness of each monthly random variable of the processes, the best fit theoretical probability density function was determined among the 12 functions described in 2.3.1. In order to provide the best representation of probability for each monthly random variable, the best fit was used as initial condition for the model, also if different functions resulted for different months of the same random process. It was considered more important to use the best fit than to use the same distribution for all 12 random variables and therefore use a more inaccurate function for some months. For some months, the distribution that resulted as best fit most often for the random process, did not pass the Kolmogorov goodness of fit test, which was another reason why a general best fit for each random process was not found. However, if this is not the case, it might be an interesting topic for future investigations to evaluate the differences that result from using a general best fit for all the random variables of the same process.

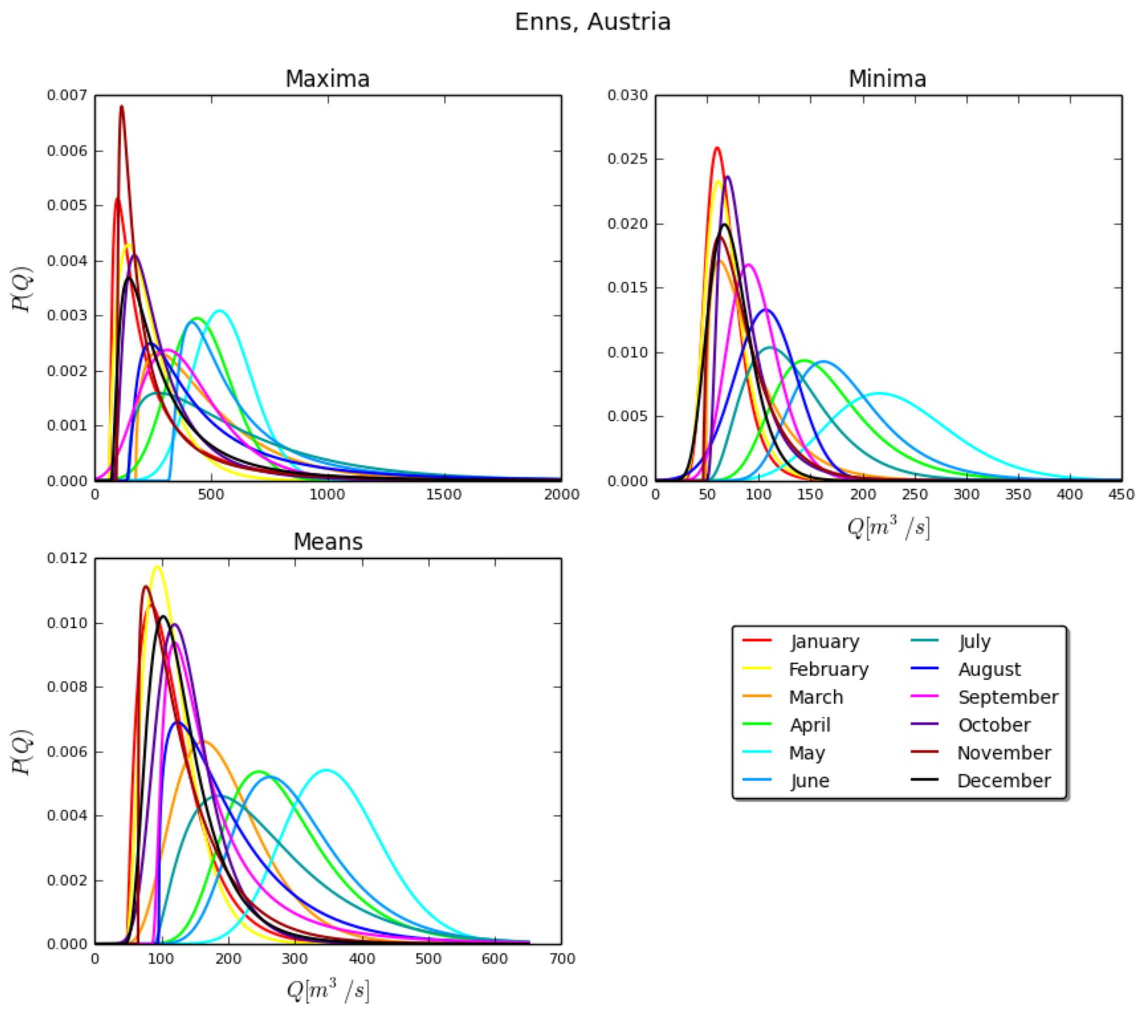

Figure 24. Probability density functions of the random variables in the Enns River basin

Month	Means	Maxima	Minima
January	Gamma	Lognormal	Gumbel pos. skew
February	Gamma	Gamma	Gumbel pos. skew
March	Gamma	Gamma	Gamma
April	Gumbel pos. skew	Gamma	Gumbel pos. skew
May	Gamma	Gamma	Gamma
June	Gumbel pos. skew	Lognormal	Gumbel pos. skew
July	Gamma	Gamma	Gamma
August	Gamma	Lognormal	Normal
September	Lognormal	Gumbel pos. skew	Lognormal
October	Gumbel pos. skew	Lognormal	Lognormal
November	Gamma	Lognormal	Gamma
December	Lognormal	Lognormal	Lognormal

Table 6. Best probability density function fit for each random variable in the Enns River basin

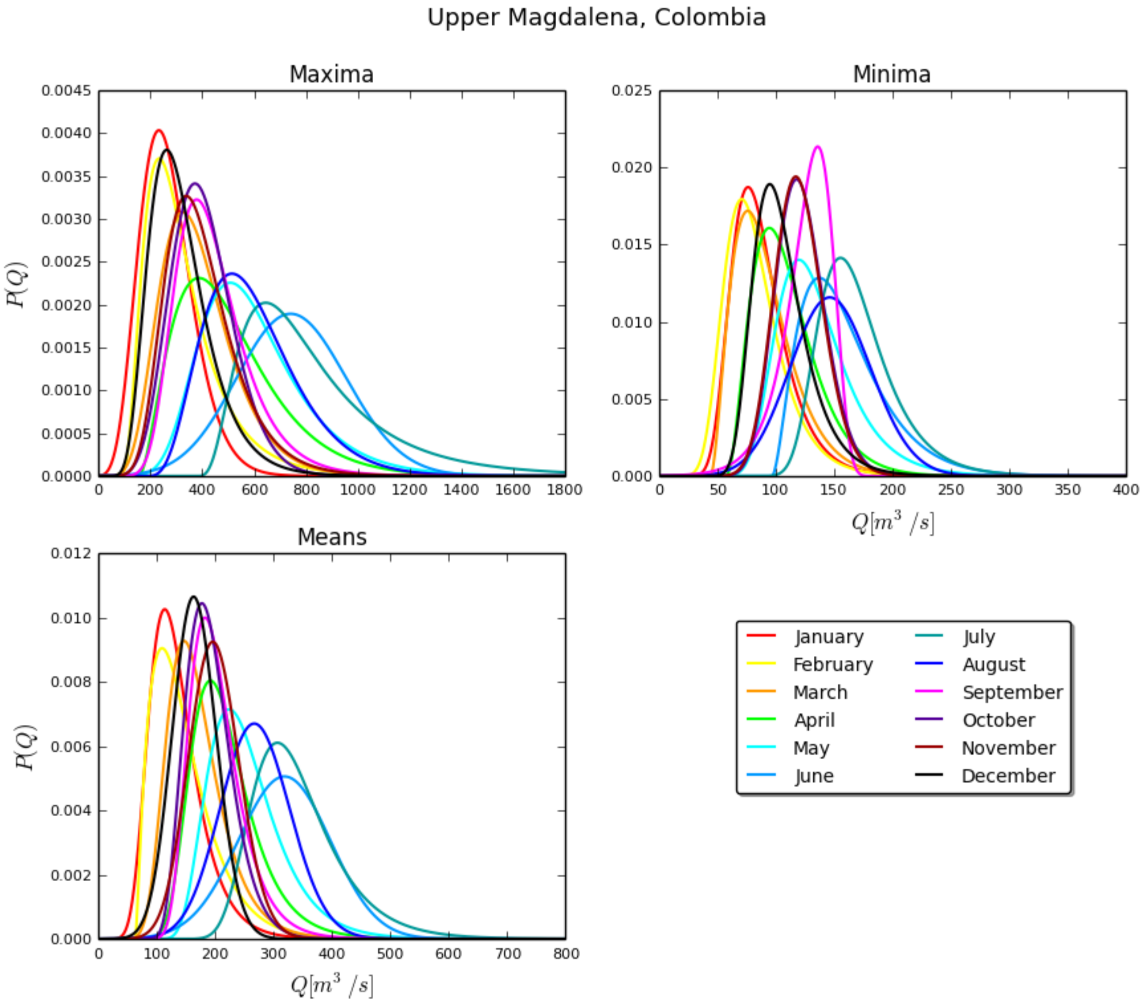

Figure 25. Probability density functions of the random variables in the Upper Magdalena River basin

Month	Means	Maxima	Minima
January	Gumbel pos. skew	Gamma	Gumbel pos. skew
February	Gamma	Lognormal	Gumbel pos. skew
March	Gumbel pos. skew	Lognormal	Gamma
April	Gumbel pos. skew	Gamma	Gamma
May	Lognormal	Gumbel pos. skew	Gumbel pos. skew
June	Loggamma	Gamma	Gamma
July	Gumbel pos. skew	Lognormal	Gumbel pos. skew
August	Lognormal	Gamma	Gamma
September	Gumbel pos. skew	Gumbel pos. skew	Gumbel neg. skew
October	Gamma	Gamma	Gamma
November	Lognormal	Gumbel pos. skew	Normal
December	Lognormal	Lognormal	Gumbel pos. skew

Table 7. Best probability density function fit for each random variable in the Upper Magdalena River basin

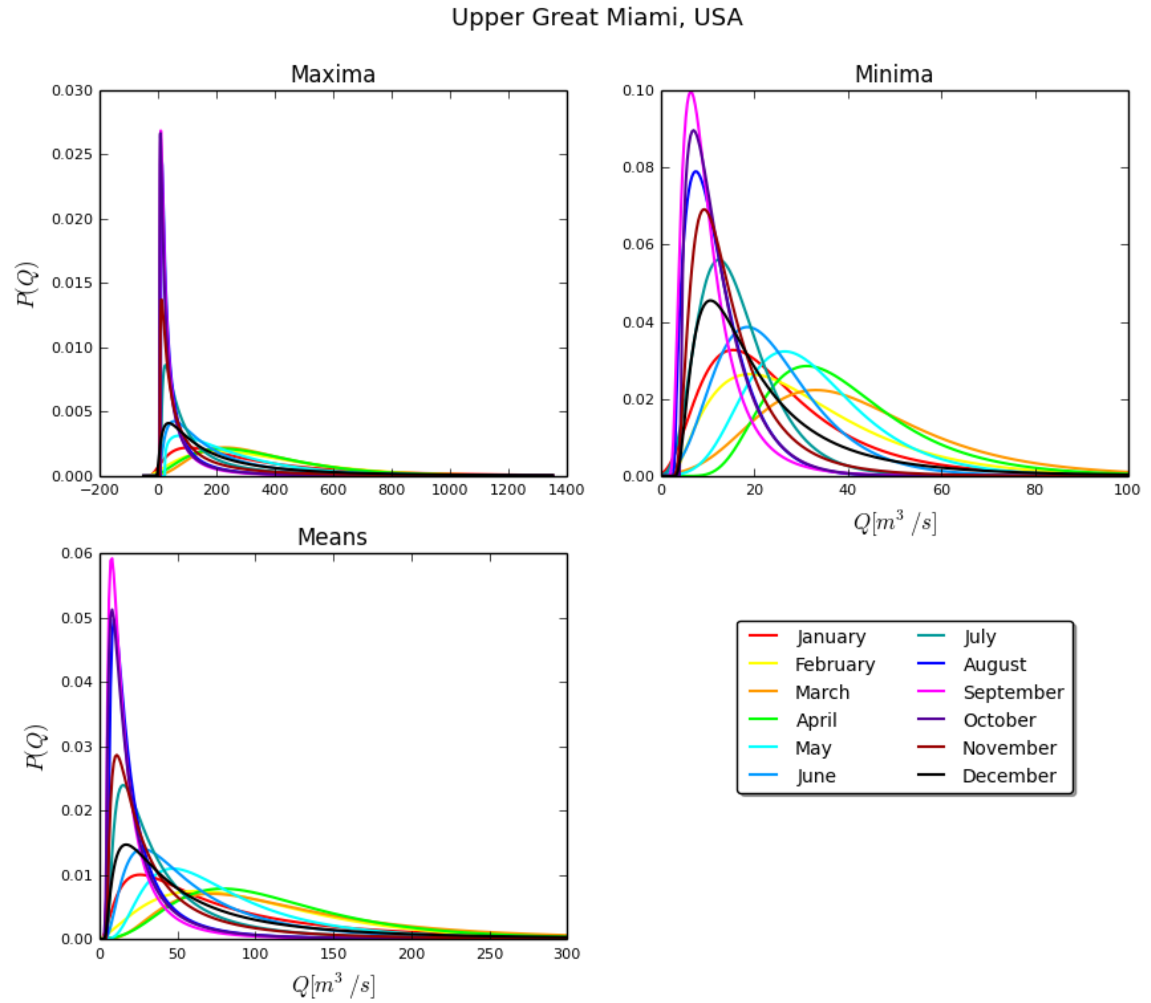

Figure 26. Probability density functions of the random variables in the Upper Great Miami River basin

Month	Means	Maxima	Minima
January	Lognormal	Lognormal	Lognormal
February	Gamma	Gamma	Gamma
March	Lognormal	Lognormal	Gumbel pos. skew
April	Lognormal	Gamma	Lognormal
May	Lognormal	Gamma	Gumbel pos. skew
June	Lognormal	Lognormal	Lognormal
July	Lognormal	Lognormal	Lognormal
August	Lognormal	Lognormal	Gamma
September	Lognormal	Lognormal	Lognormal
October	Lognormal	Lognormal	Gamma
November	Lognormal	Lognormal	Lognormal
December	Lognormal	Lognormal	Lognormal

Table 8. Best probability density function fit for each random variable in the Upper Great Miami River basin

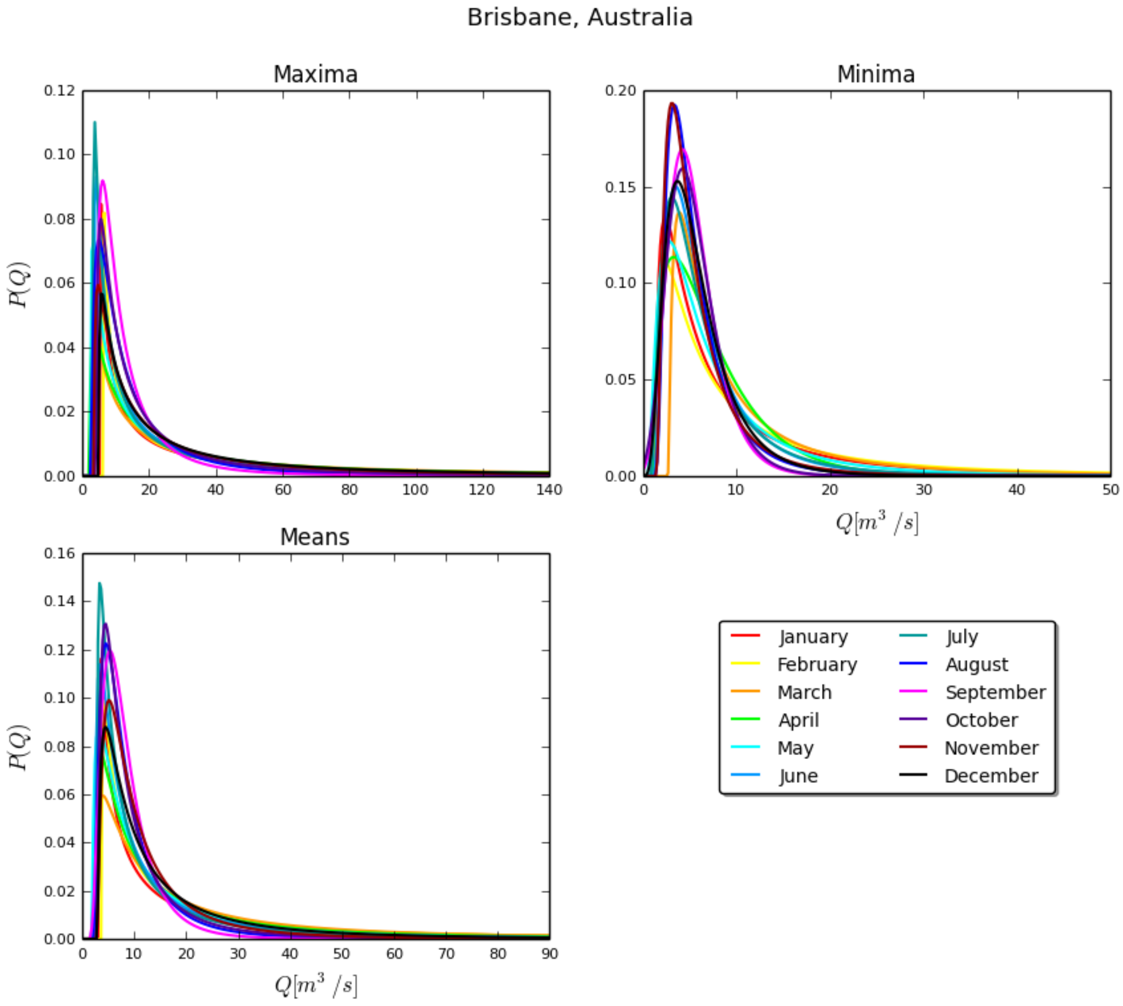

Figure 27. Probability density functions of the random variables in the Brisbane River basin

Month	Means	Maxima	Minima
January	Lognormal	Lognormal	Lognormal
February	Lognormal	Lognormal	Lognormal
March	Lognormal	Lognormal	Lognormal
April	Lognormal	Lognormal	Gamma
May	Lognormal	Lognormal	Lognormal
June	Lognormal	Lognormal	Lognormal
July	Lognormal	Lognormal	Lognormal
August	Lognormal	Lognormal	Lognormal
September	Lognormal	Lognormal	Gumbel pos. skew
October	Lognormal	Lognormal	Gumbel pos. skew
November	Lognormal	Lognormal	Lognormal
December	Lognormal	Lognormal	Gamma

Table 9. Best probability density function fit for each random variable in the Brisbane River basin

The upper and lower boundary limits were all set according to the values contained in the observed data, in order not to cut off parts of the curve and lose part of the probabilities contained under the curves this way. Consequently, for the maxima in the Great Miami basin, the lower border of the boundary had to be extended into the range of negative numbers.

5.2 Optimization of model parameters

The calibration of the model consisted of the optimization of the parameters defining the Fokker-Planck-Kolmogorov equation and was conducted with respect to the physical interpretation of the outcome as shown in Figure 28.

For each random process, the Fokker-Planck-Kolmogorov model was applied to describe the translations from each monthly probability density function to the one of the next month. For this purpose, the inverse problem approach was used, optimizing each of the model parameters to obtain the best possible simulation. The optimization was executed for both the explicit and the implicit method proposing a supervised and an unsupervised approach.

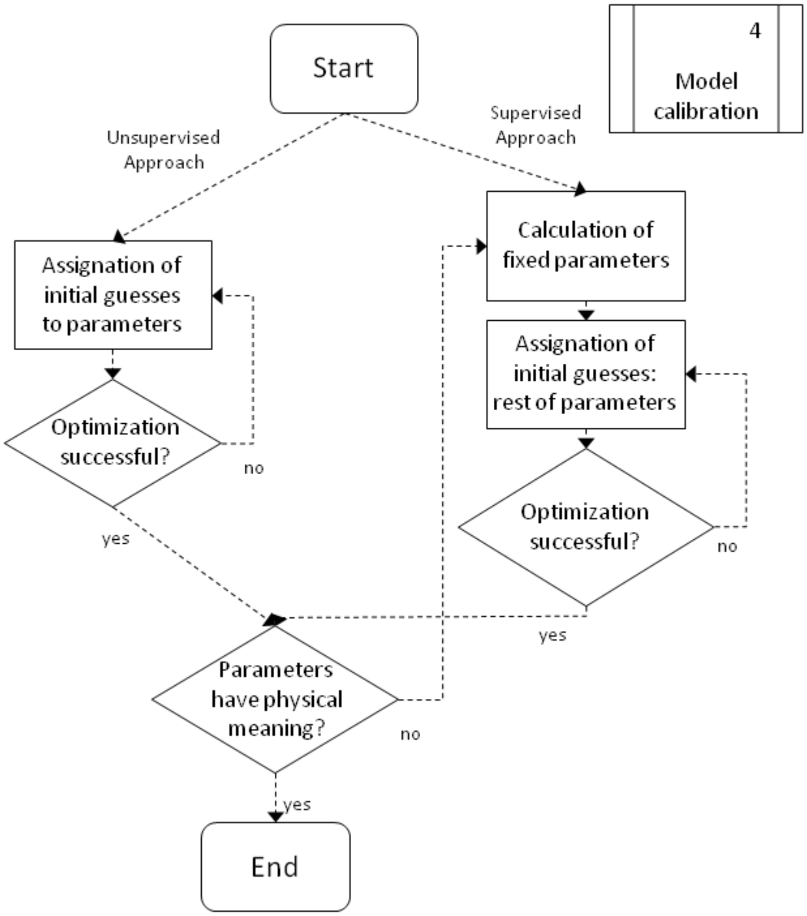
Start
4
Model calibration
Unsupervised Approach
Supervised Approach
Assignation of initial guesses to parameters
Calculation of fixed parameters
Assignation of initial guesses: rest of parameters
Optimization successful?
no
yes
Optimization successful?
no
yes
Parameters have physical meaning?
no
yes
End

Figure 28. Schematic outline of model calibration

The six parameters defining the drift and diffusion vectors of the FPK equation, k_1, k_2, k_3, g_1, g_2 and g_3 were optimized. The process of optimization was automated with a Python script that loaded the data of the 12 monthly probability density curves, assigned an initial guess to each of the six parameters and applied the optimization procedure for both the implicit and then the explicit scheme for each PDF translation, and at the end saved a plot of the result of the simulated against the observed data for visual inspection and a list of the optimized parameters with their respective mean average error. The exit condition used by the *curve_fit* algorithm caused the optimization to terminate automatically after 1400 iterations and a set of empty parameters was returned.

The whole optimization procedure was computationally intensive and, depending on the parameters used, took between a few minutes and 4 hours to complete for the 12 translations of one random process on the laptop with a Pentium Dual Core CPU with 2.1GHz and 4GB RAM, on which the majority of the calculations were run.

The first results indicated that a multitude of optimized solutions was possible and that the procedure was very much dependent on the initial guesses used for the model parameters. Changing them in some cases resulted in a completely different set of optimal parameters obtained and therefore indicated that a large number of local minima existed for each translation, but that it was very difficult to find the global minimum of the optimization function that provided the optimal model parameters. However, it also showed that it was not necessary to find these optimal model parameters to achieve a perfect fit of the simulated data to the observed one. The following table shows the optimized parameters with different initial guesses of minima in the Great Miami River basin from November to December with the implicit method. All sets of optimized parameters result in an almost perfect fit with a mean average error of fewer than 3%. It can be seen, however, that some of the initial guesses produce similar results, but others are very distinct, especially for the parameters related to diffusion.

Initial Guess						***Optimized Parameters***					
k1	***k2***	***k3***	***g1***	***g2***	***g3***	***k1***	***k2***	***k3***	***g1***	***g2***	***g3***
0	0	1	0	0	1	0.484	-0.0023	-1.278	0.185	0.033	-2.213
0	0	0	0	0	0	0.556	-0.0047	-1.202	-0.468	0.122	-0.576
0.1	0.01	1	0.1	0.01	1	0.578	-0.0055	-1.060	-0.685	0.159	-0.061
0.1	0.01	0	0.1	0.01	0	0.603	-0.0063	-1.037	-0.944	0.191	0.708
0.01	0	-1	0.01	0.001	1	0.482	-0.0019	-1.589	0.361	-0.004	-2.715
0.01	0	1	0.01	0.001	1	0.584	-0.0059	-0.769	-0.873	0.196	0.586
0.01	0.001	1	0.01	0.001	1	0.578	-0.0054	-1.059	-0.685	0.159	-0.061
1	0.1	10	1	0.1	10	0.505	-0.0031	-1.165	-0.016	0.067	-1.783

Table 10. Example of optimized parameters using different initial guesses

To evaluate the result, it was established that a simulation that counted with a mean average error of 10% or less was considered to be an acceptable result.

Another result that appeared in some tests was the fact that parameters could be optimized more easily when there were fewer nodes to be calculated, or in other words, when the limits of the boundary condition were closer together. The idea of trimming the boundary condition to cut off long tails was considered, but discarded in order to ensure the same conditions for each translation of the same random process. In any case, this topic could be studied in more detail in future investigations. Due to these insights, a stepwise procedure was implemented in order to automatically achieve a higher rate of successful optimizations. Therefore, in the first step, the rapid optimization of the implicit scheme was conducted using a predefined initial guess, followed by the explicit method with the same initial guess. If one of the optimizations failed to produce a result, it was rerun using the result of the other method as the initial guess, if this was available. This way the percentage of successful optimizations could be raised from about 40% initially to over 80%. For the remaining translations, the initial guesses were changed manually to obtain a satisfactory result for each translation with at least one of the methods. With the exception of some translations in the Enns basin, this could be achieved and produced a simulation for the random processes representing the monthly means, maxima and minima in each basin. The bad fit observed in the Enns River data in the transition from September to October (Figure 29) is an example of the fact that the model generally had difficulty fitting the result in translations where the probability density experiences a significant sharpening of the curve.

Below, the results of the optimization are shown. For the sake of brevity, only the result of one of the variables is shown for each basin. All results were achieved using the optimization of the implicit scheme. Perfect fits were also obtained with the explicit scheme for over 60% of all translations, but due to the longer computation time, not all of the missing ones were optimized manually.

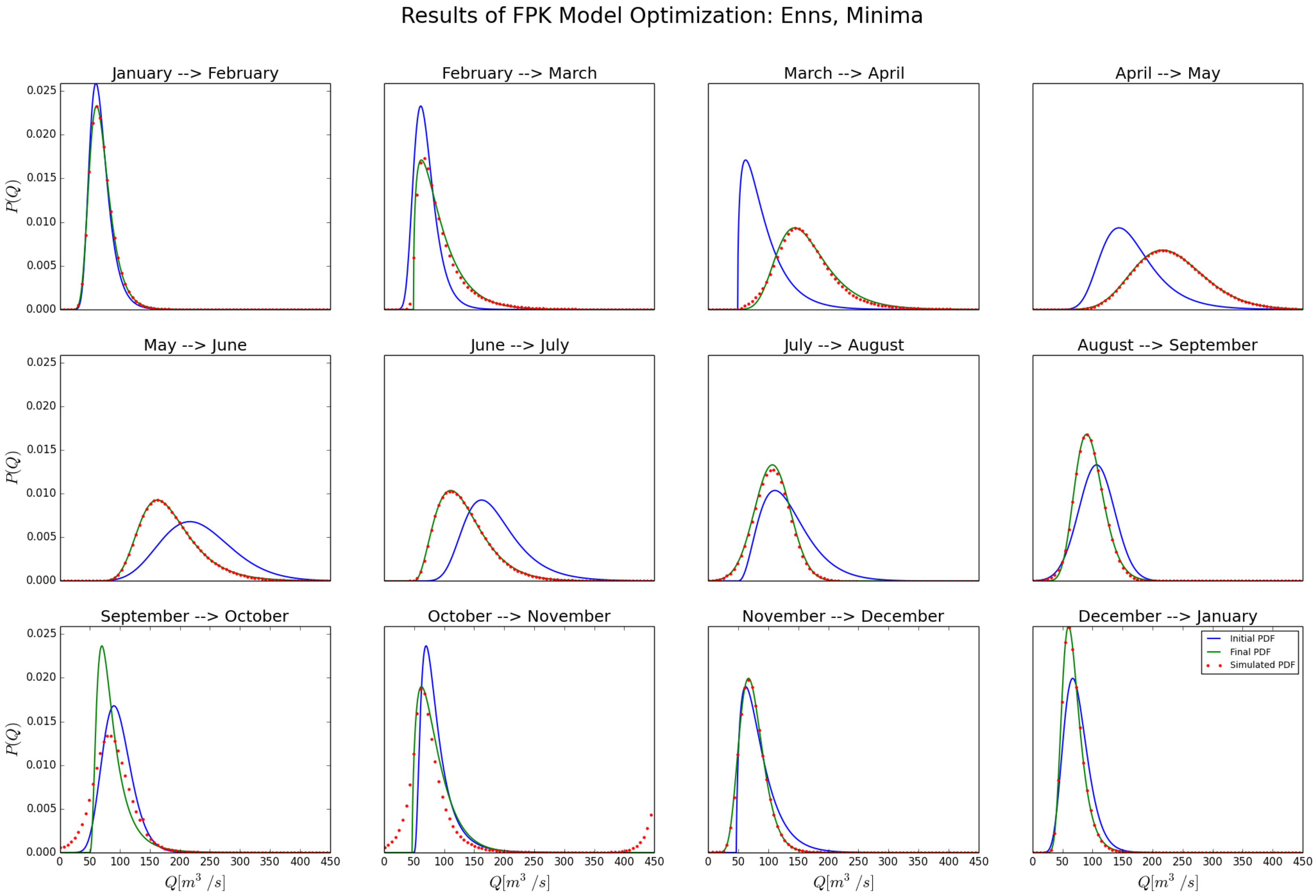

Figure 29. Results of optimization in the Enns River basin

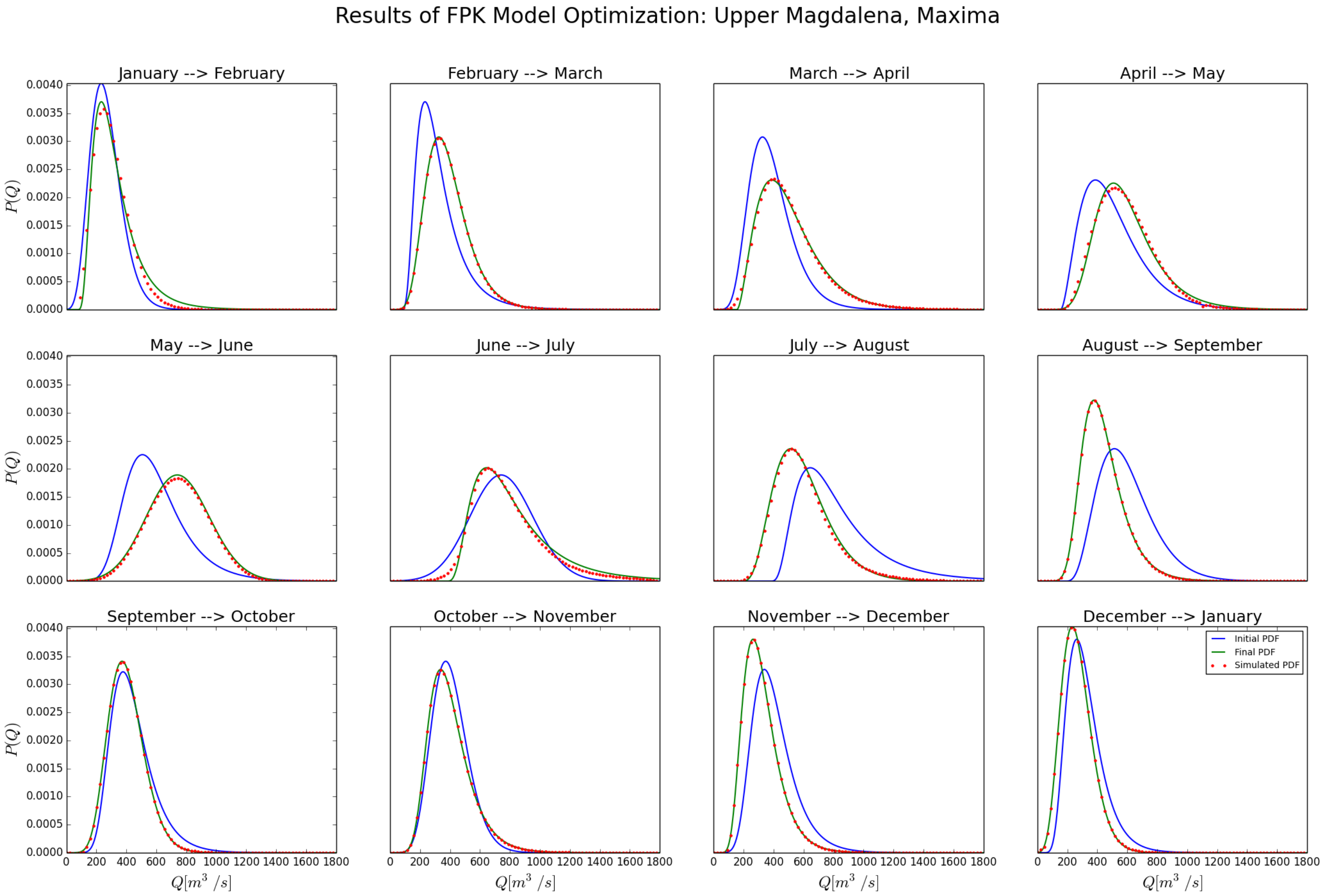

Figure 30. Results of optimization in the Upper Magdalena River basin

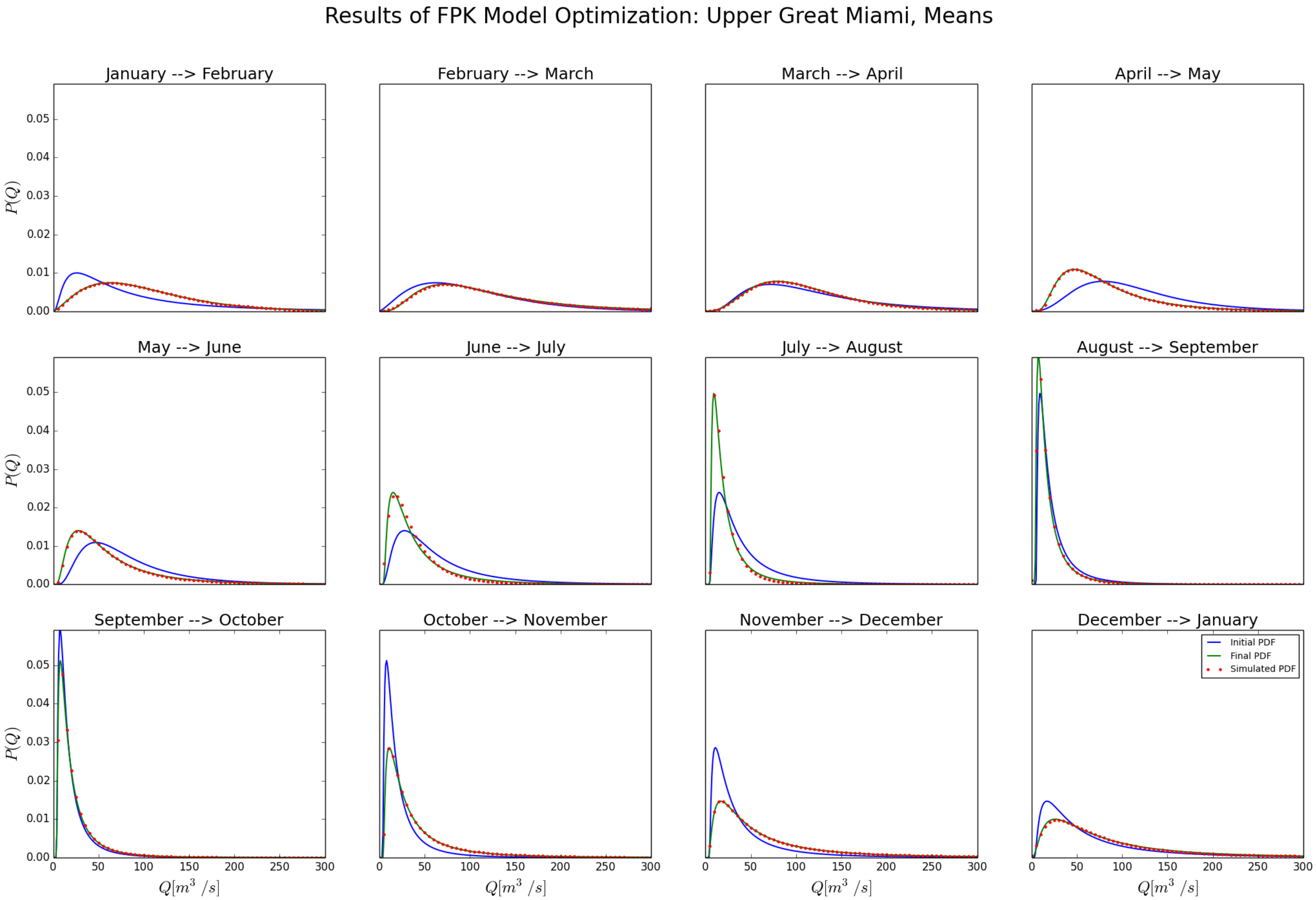

Figure 31. Results of optimization in the Upper Great Miami River basin

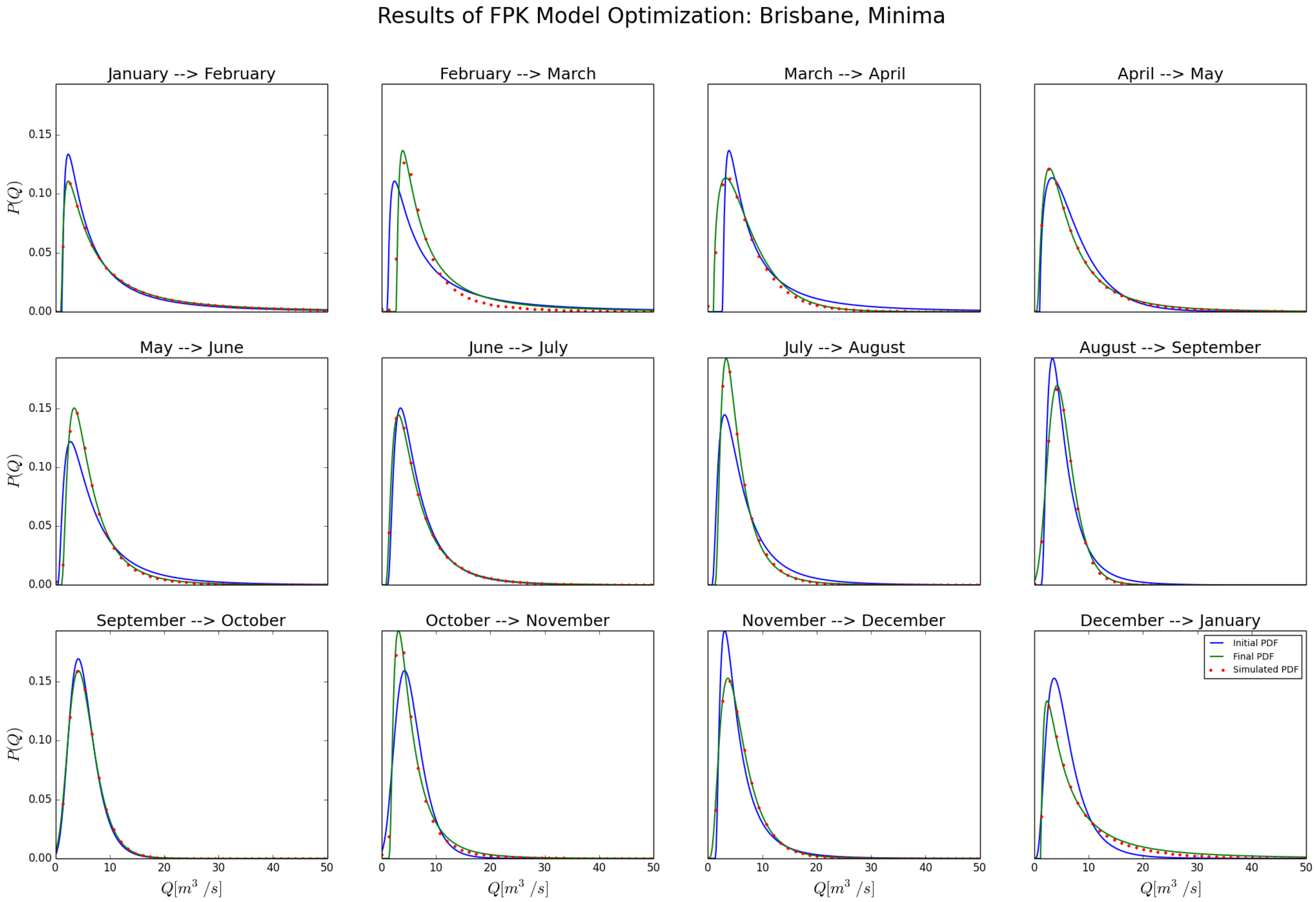

Figure 32. Results of optimization in the Brisbane River basin

The results showed that the optimized parameters using the explicit scheme never matched those using the implicit scheme. Explanations for this phenomenon are either that two different local minima were found by the optimizing functions, or that there was numerical diffusion present in the implicit scheme changed the parameters. It was found that the implicit method added numerical diffusion to the translation, although it was not determined to which degree. Figure 33 shows the application of the optimized parameters of the explicit scheme both to the explicit and the implicit function. The green PDF curve of observed data in the final month is displayed as a reference.

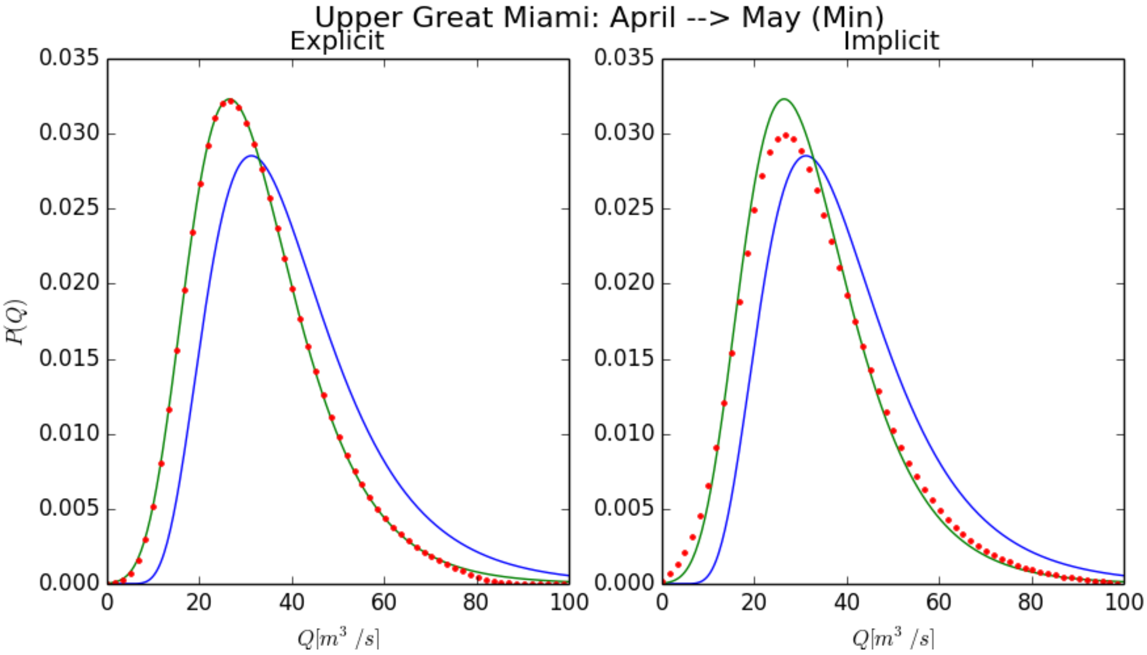

Figure 33. Numerical diffusion using the explicit scheme

5.3 Relation of model parameters to physical properties

To apply the model to be able to make simulations for the changes in probability, it was necessary to relate the optimized parameters of the drift and diffusion equations of the FPK equation to the parameters of the watershed.

Without knowing the detailed characteristics of the river basin, it is difficult to determine the internal parameters and therefore link them to the parameters of the Langevin equation. With the external parameters, it is easier to determine, also to establish a relationship between these and the discharge values. Two external parameters, for which data was available, were considered to influence on the discharge characteristics, first precipitation, and second temperature, which again drives evaporation in the basin.

In order to get an idea of the influence precipitation has on the extreme discharge series, the correlation between the two series was calculated. Therefore, the values of monthly discharge maxima and minima were correlated with the precipitation values of the same day and the averaged precipitation amount of up to 20 days prior. This analysis was conducted in three different ways: first, the discharge value was correlated with only the precipitation value of lag 0 to 20; second, the mean of precipitation values of the same day as the discharge maximum and up to 20 days before was correlated and third, the same procedure was repeated without taking into account the same day, averaging only precipitation values preceding the date of the discharge extreme value measured.

For monthly maxima, a high correlation could be established in all of the 4 basins, with correlation values of 0.6 and above. In most of the cases, the averaged value over the lag time including the current date shows the highest correlation with the discharge data.

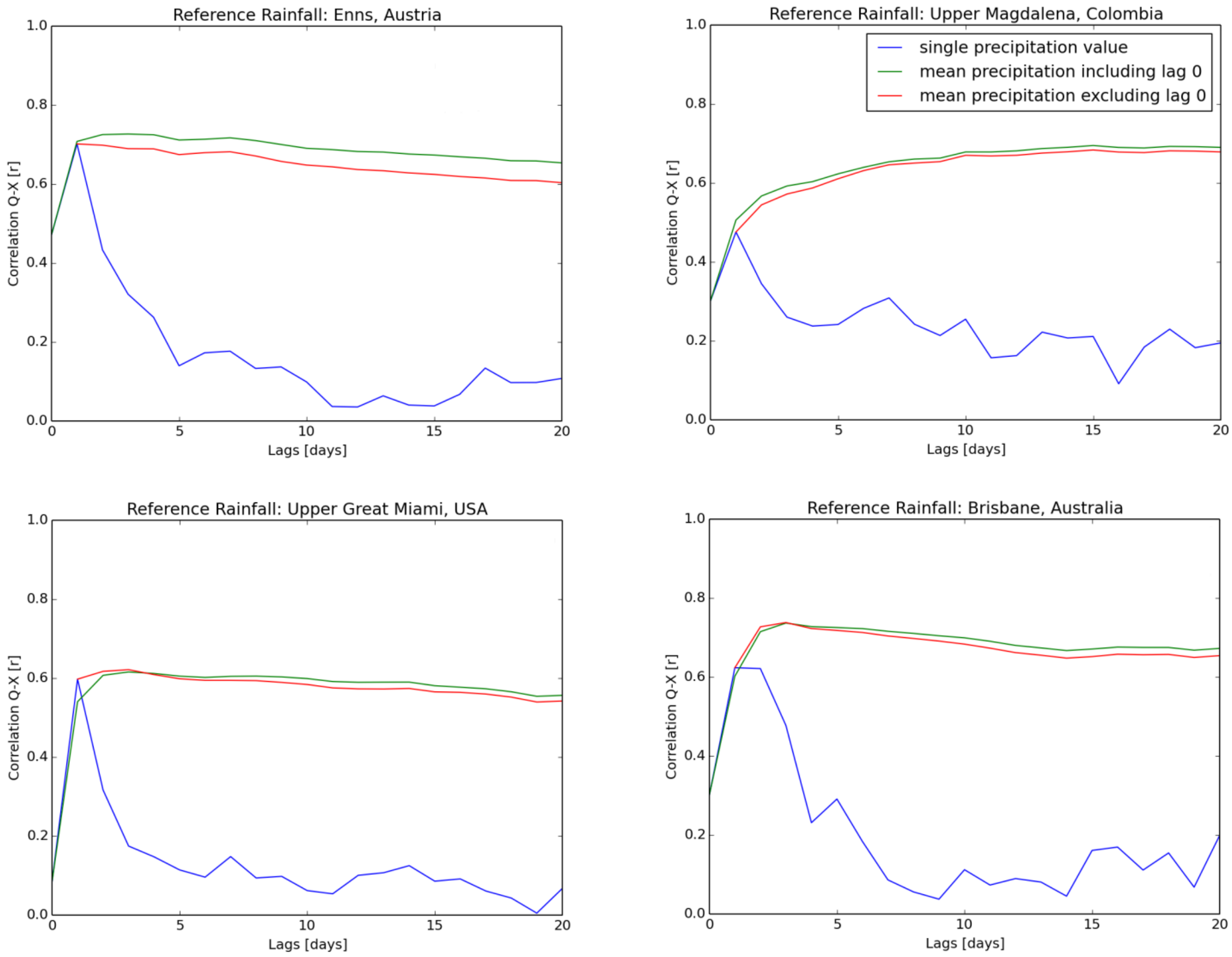

Figure 34. Correlation between discharge maxima and precipitation

Figure 34 shows that the best correlation between the values of monthly discharge maxima can be achieved with the average of the daily precipitation values of a varying lag time. The correlations with the methods including or excluding the precipitation value of the same day as the discharge extreme are usually very similar, with the exception of the Enns basin. A correlation between only the single precipitation value and the discharge maximum applying a lag time was highest for a lag of one day, but not significant any more with a lag time of more than 3 days. However, in all basins, the best correlation with a single precipitation value was still exceeded by that using the average of an interval of previous days.

Using the averaged precipitation values during the lag time with the highest correlation in each basin, monthly reference rainfall value series were created for the series of monthly maxima. The specifications with the highest correlation are given in table 11. In all cases the averaged value over the interval of days within the lag time period gave the best result. For the Great Miami and Brisbane basins, this was achieved excluding the precipitation value measured the same day as the discharge maximum, in the other two it was included.

Basin	Lag time [days]	Method	Correlation [r]
Enns	3	mean including same date	0.727
Upper Magdalena	15	mean including same date	0.695
Upper Great Miami	3	mean excluding same date	0.622
Brisbane	3	mean excluding same date	0.734

Table 11. Lag times and method with the highest correlation between discharge maxima and precipitation values for each basin

Due to the fact that only precipitation stations with a correlation of more than 60% compared with discharge series were used, monthly precipitation averages could be used as a referencefor monthly mean discharge series. The reason why the commonly applied precipitation sums were not used was the difference in reference time, since the lag times were shorter than the duration of a month and therefore comparability between the results could not be given.

Discharge minima are not correlated as highly with the precipitation series. Although in some cases correlation results as high as 60%, it was not considered to be a substantial enough relation with the discharge minima, especially because of the low values observed in the Enns and Great Miami basins. Again, it could be seen that correlations with the mean of the interval of previous days was generally higher than that of only a single lagged precipitation value.

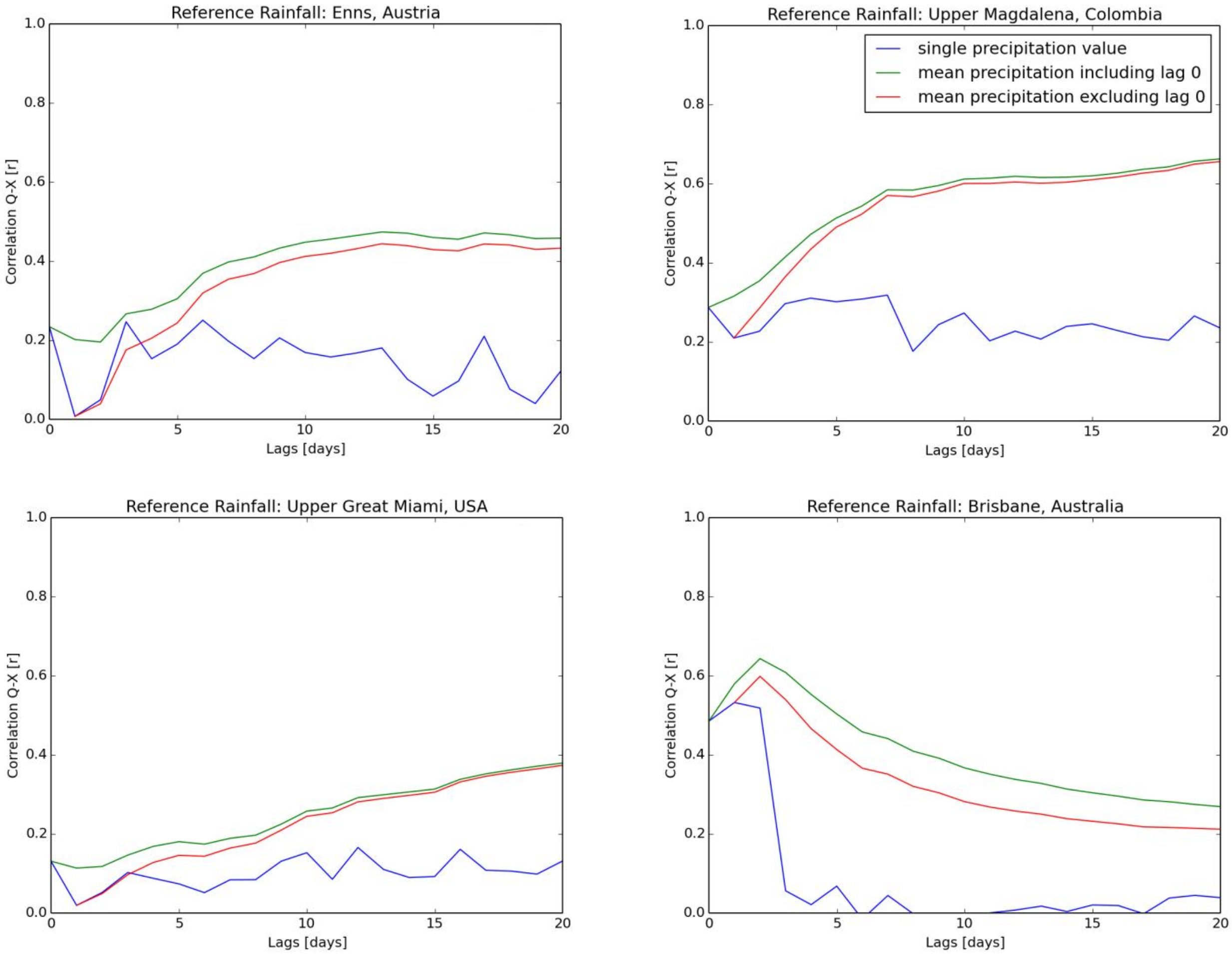

Figure 35. Correlation between discharge minima and precipitation

It was expected that baseflow has a higher influence on the periods in which discharge minima occur. Baseflow is principally influenced by evaporation, which again is driven by atmospheric temperature, for which reason it was tried to establish a relation between discharge minima and temperature. Due to the small variation in temperature series, as was shown in the trend analysis in section 3.2, it was regarded sufficient to use monthly mean temperature values for the analysis. Results of correlation values between monthly discharge minima and mean temperature were chosen to be used as an approximation for minimum data in all basins except the Brisbane basin, although they were only a slightly better approximation than the correlations with precipitation. In the Brisbane basin, due to the very low correlation between discharge minima and temperature data, the analysis of minima was conducted using precipitation data, which was considered to be sufficiently correlated using the mean value of daily precipitation of 2 days prior to the discharge minimum event. The results of correlation between monthly mean temperatures and discharge minima can be seen in table 12.

Basin	Correlation
Enns	0.544
Upper Magdalena	0.526
Upper Great Miami	0.636
Brisbane	0.116

Table 12. Correlation between discharge minima and monthly mean temperature for all basins

As defined in Gardiner (2004) and Sveshnikov (1966), the model parameters *A* and *B* of the Fokker-Planck-Kolmogorov equation are defined as

$$A\left(\underline{q},t\right) = \lim_{\Delta t \to 0} \frac{E\left[\Delta\,\underline{q}\middle|\underline{q}\right]}{\Delta t} \tag{54}$$

$$B\left(\underline{q},t\right) = \lim_{\Delta t \to 0} \frac{E\left[\Delta\underline{q}^2\middle|\underline{q}\right]}{\Delta t} \tag{55}$$

It was inferred that a correlated variable might influence on these relationships, even if the discharge value is infinitely small. For this reason, it was assumed that the independent term of each equation determining the drift and diffusion vectors should be most closely related to the external parameters of the basin.

Furthermore, it was proven in various previous works (Dominguez, 2004; Kozhevnikova et al., 2012; Maldonado, 2009) that changes in the coefficients of variation and skewness are related to changes in the internal parameters of the river basin. Because of this reason, a linear regression analysis using Ordinary Least Squares technique was conducted to estimate the degree to which each of the optimized model parameters correlates with the coefficients of variation and skewness, as well as the rainfall and temperature values, all of which were used of the final month of each translation. Also, the variation and skewness coefficients of the initial month were used as independent variables. For all the translations, the necessary values were combined among all the 4 basins, in order to have a bigger sample, and regression analysis was applied for each variable, mean, maxima and minima, separately.

The results of the regression analysis indicated that none of the parameters could produce a valid correlation with the proposed coefficients and precipitation and temperature values. The only valid tests were those correlating the same coefficients of the initial and final months.

These results led to the conclusion that the optimization of the parameters like they were conducted did not produce a result that models satisfactorily the physical parameters of the process, although in all the translations almost perfect fits could be achieved. Therefore, the optimization process was adapted and a supervised approach was applied.

5.4 Supervised optimization with fixed model parameters

Due to the results of the prior optimizations, it was chosen to determine some of the parameters and assign fixed values.

Leading from equations 54 and 55, the following equation were proposed to calculate fixed values for the parameters k_3 and g_3 for a supervised optimization approach. This is a first attempt to represent these values with the intention of conserving the linear nature of the drift vector and the quadratic one of the diffusion vector.

$$k_3 = \Delta X = X_{final} - X_{initial} \tag{56}$$

$$g_3 = \Delta var[X] = var\left[X_{final}\right] - var[X_{initial}] \tag{57}$$

where X represents precipitation from the time series calculated as presented above. The same equations, replacing precipitation with temperature, were applied to the optimization of discharge minima. With these two parameters taking pre-assigned values, the optimization was conducted only for the remaining 4 parameters.

The optimization was run three times, each time using different initial guesses. These were applied to the optimization of both the explicit and the implicit method. In the previous optimizations, it had shown that the initial guess of k_1 and g_1 with opposing signs could easily cause a failed optimization. Therefore both a run with a positive and one with a negative initial guess were conducted. However, because of stability issues of the algorithm, the negative parameters had to be of a smaller magnitude.

Run	k_1	k_2	g_1	g_2
Run 1	0.1	0.001	0.1	0.001
Run 2	-0.01	0.001	-0.01	0.001
Run 3	0	0	0	0

Table 13. Initial guesses used in the 3 runs of supervised optimization

For all the basins except the Brisbane basin, it was possible to obtain an optimization with a mean absolute error below 10 % for at least one of the numerical schemes for most of the translations. In more than 60% of the translations, the explicit scheme, which did not include numerical diffusion, could be optimized this way, which is approximately equal compared to the initial optimization of all model parameters. Therefore, it was decided to use the results of the optimal parameters of the explicit scheme for further analysis.

As it had to be expected, in the Brisbane basin the optimization was not possible for any of the translations, which is most probably due to the Wivenhoe dam that regulates the affluence from the majority of the river basin. Another factor that contributed to this outcome was that the variations in precipitation series in this basin were higher than in the other basins, which therefore made an optimization of the parameters even more difficult, even with different initial guesses. The only method that allowed the optimization of the parameters was a regulation of the fixed parameters k_3 and g_3, where the differences in mean precipitation and its variations were reduced to almost non-existence by division with a high-magnitude regulation coefficient. However, even then was it not always possible to find an appropriate solution and the basin remained that with the least number of successful optimizations.

It could therefore be concluded that the regulatory nature of the dam does not permit a reasonable application of the model in the Brisbane basin. Since the changes in discharge do not respond to the natural variation of the major part of the watershed, but are principally controlled by the artificial conditions imposed by the amount of water passing by the dam, one cannot apply natural operators to the model. Therefore it was not considered reasonable to use the optimization results obtained with regulated parameters, and the Brisbane basin was not used in further analysis.

Regression analysis was applied to the optimized parameters of the explicit scheme of the remaining 3 basins, which indicates a relationship between the parameters k_1, g_1 and g_2 and the coefficients of variation and skewness. Again, also the coefficients of the initial probability distribution play an important role. For the optimization of the translations of mean discharges, it can be seen that all three parameters influence the outcome of the coefficients of variation and skewness. For extreme values, the changes in the coefficients could only be related to the parameter k_1, which suggests that they are less influenced by the fluctuations and more by direct changes to the kernel of the system. In all cases, however, the values of the initial coefficients of variation and skewness play a crucial role. Furthermore, it can be seen that the correlation of the parameter k_1 is stronger with the skewness coefficient than with the coefficient of variation.

Table 14 on the next page shows the result of the regression analysis. The overall correlation value *r* of the test as well as the weights of the parameters and the initial coefficients are shown, but only if they are statistically valid. It can be seen that the influence of the coefficient k_1 is strongest for maxima and weakest for minima, and is always correlated negatively to the respecting coefficients. However, it must be taken into consideration that this parameter is related to precipitation values in the case of means and maxima, and temperature for minima, which most probably influences on the results.

The coefficient k_2 does not have a significant correlation with any coefficient, which indicates that it does not necessarily have to be used. This makes sense, since it is associated to the squared value of discharge to calculate the drift vector, which is of linear nature.

Results of Regression Analysis						
Coefficient of Variation						
Variable	r	k_1	k_2	g_1	g_2	initial Cv
Means	0.996	-0.133		0.007	1.534	0.999
Maxima	0.990	-0.277				0.9848
Minima	0.999	-0.041				0.999
Coefficient of Skewness						
Variable	r	k_1	k_2	g_1	g_2	initial Cs
Means	0.956	-0.503		0.058	11.1	0.903
Maxima	0.981	-0.724				0.874
Minima	0.999	-0.005				1.004

Table 14. Results of multiple regression analysis between the model parameters and the coefficients of variation and skewness

It can be concluded that the parameter k_1 is most definitely related to the internal parameters of the basins, and due to the results of regression analysis of mean values, it is also probable that g_1 and g_2 are likewise, but not to the same extent. Due to the fact that no data related to the internal parameters of the basins was available, the parameters could not be assigned with certainty to any specific one of the internal parameters.

5.5 Application of the model to estimate the change of extreme regimes

In order to simulate the degree of change in the regime of extreme events, the calibrated model was used to predict the future behavior of discharge extremes by changing the parameters of the FPK equation. Two different simulations were conducted, differentiating between changes in the external watershed parameters caused by the alterations triggered for example by global climate change, and those in the internal parameters, which can be related to results of human activities in the river basins, among others. For each simulation, the alterations in probability density was calculated for both maxima and minima, as well as monthly means, in order to establish a relationship as to which amount extreme values experience changes compared to the mean discharge behavior. At the end, both of the two simulations were combined into one simulation. As mentioned before, the changes in discharge regimes in the Brisbane basin do not depend on natural influences, therefore it was not considered useful to include this basin in the evaluation of the model. The schematic outline of the evaluation process is shown in Figure 36.

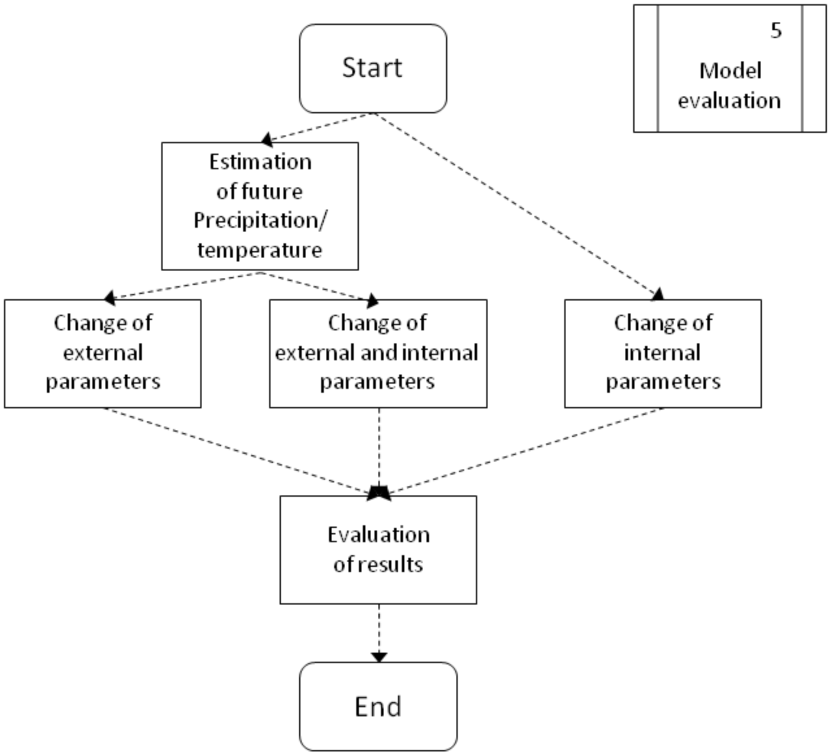
5
Model evaluation
Start
Estimation of future Precipitation/ temperature
Change of external parameters
Change of external and internal parameters
Change of internal parameters
Evaluation of results
End

Figure 36. Schematic outline of model evaluation

5.5.1 Simulation with external parameters

For the simulation of changes in external parameters, observed data could be used. Therefore, the linear trend was calculated for all monthly random variables for precipitation and temperature. With the obtained trend slope, it was possible to estimate future values, representing the possible monthly values for each future year. Future values were estimated for monthly mean precipitation and temperature, as well as monthly reference rainfall in the case of discharge maxima.

For the same variables, an estimation of the change in variability was created. For this purpose, each monthly random variable was divided into 2 subsets, representing the chronologically first and second half of the observed values. For each subset, the variance was calculated, from which a percent wise change could be determined, which was applied to the total variance of the random variable. The percentages of change of these trends generally lay between 2 and 20% of the initial value at the end of a 20-year period for precipitation series, and were higher for some cases in variance. The average absolute value among all calculated trends was of approximately 9%.

For the simulation run, the predicted precipitation and temperature values for 20 years after the last year of observation were used, which was the year 2030 for the Magdalena and Enns basins, and 2032 for the Great Miami and Brisbane basins. The parameters k_3 and g_3 were calculated the same way as in the calibration runs, only applying the newly estimated future precipitation and temperature values instead of the observed ones. The amount of change in probability density was calculated using the Kullback-Leibler divergence criteria. For all of the translations of each random process, the resulting Kullback-Leibler divergence values were averaged to obtain overall values for each basin, which allowed a comparison between them.

The results indicate that the changes have the strongest intensity in the Great Miami basin, and the least in the Upper Magdalena basin. While for the Enns and the Great Miami basins, the change in minima was strongest, for the Magdalena basin it was that of maxima. The averages over the Kullback-Leibler results of all translations of the same variable in each basin are shown in table 15.

Basin	Mean	Maxima	Minima
Magdalena	0.0004	0.0009	0.0003
Enns	0.0008	0.0004	0.0057
Great Miami	0.0018	0.0004	0.0078

Table 15. Averages of Kullback-Leibler divergences between present and future simulations of all PDF translations of each basin, applying changes to external properties

The intensification and weakening of the extreme events could be estimated by the movement of the probability density curve to either direction, or by the way the area contained under the tails of the distributions changed. In most of the cases, the curve was moved to one direction by the changes, but it also occurred that it sharpened or flattened and the tails of the distribution rose or lowered to an approximately equal degree.

In most of the basins, there was an equal amount of months, in which extreme events, both maxima and minima, intensified or weakened. Only for the Upper Great Miami basin, more months showed a weakening of extreme events than intensification. Also, the occurrence is related to the different seasons in some cases, for example in the Enns basin it can be observed that minima weaken in spring and fall and intensify in the drier summer months. Equally, maxima tend to intensify in the spring months. In the Magdalena basin, it can be seen that the maxima tend to intensify in the rainy season between September and November, although the changes are very small compared to those in other basins.

Following, an example of minima is shown that indicates a weakening of these events, following from the curve's movement due to the changes in the external parameters. The probability density functions of the initial month and the one simulated with the optimized model parameters for the final month are shown. The function that resulted by applying the estimated future values of temperature is also added to show the change of the curve in the future scenario. The differences between the present and the future functions are presented at the bottom of the graphic to visualize in which direction the curve moves. From the move of the curve to the right, it may be concluded that minima in August in the Great Miami basin become less intense due to the changes in temperature. It can also be seen that the impact of the changes in external parameters hardly influence on the variation of the curve with only a drift to one side and almost no diffusion.

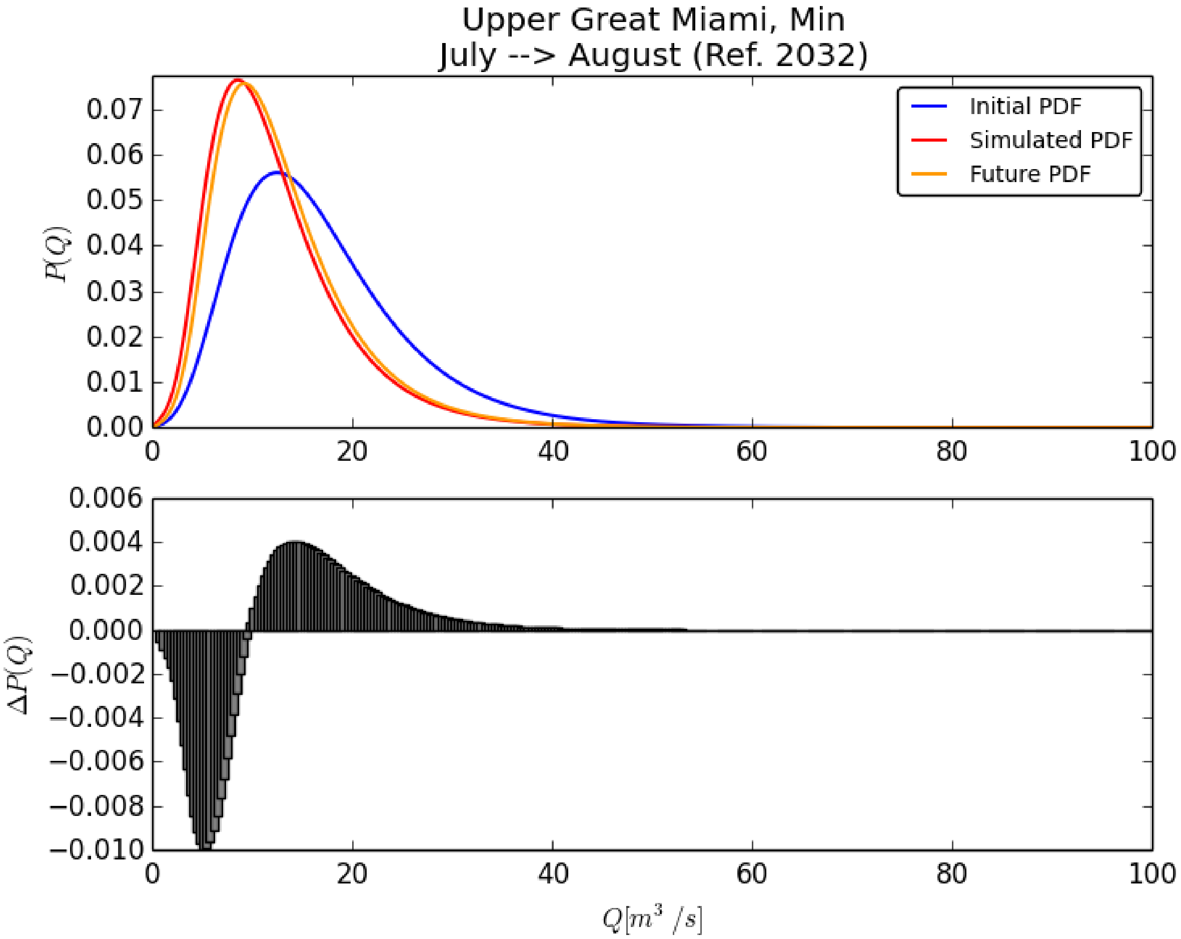

Figure 37. Simulation of the model applying changes to external parameters in the Upper Great Miami basin

5.5.2 Simulation with internal parameters

Due to the non-existence of data concerning the internal parameters of the basins, the parameters for the simulations could not be modified with dependence on existing data. Therefore, the parameters that are related to the internal properties of the system were changed percent wise to an extent that should be similar to the degree of change of the external parameters. Regarding the average percentage of change observed in the trends, 10% of the values were added and subtracted only from k_1, g_1 and g_2. In this investigation, an increase of one parameter and simultaneously a decrease in other parameters was not applied, although this would be an interesting topic for future investigations. The parameter k_2 was not changed, since no relationship to the internal properties of the basins could be established. Also, for k_3 and g_3, the values were not modified compared to the calibration runs. The results of the Kullback-Leibler divergence are shown in table 16.

	Change -10%			**Change +10%**		
Basin	**Mean**	**Maxima**	**Minima**	**Mean**	**Maxima**	**Minima**
Magdalena	0.0030	0.0037	0.0037	0.0041	0.0064	0.0038
Enns	0.0024	0.0004	0.0070	0.0034	0.0019	0.0141
Great Miami	0.0077	0.0051	0.0188	0.0049	0.0062	0.0120

Table 16. Averages of Kullback-Leibler divergences between present and future simulations applying changes to internal properties

The results of the analysis show clearly that the impact of changes in the internal parameters of the system is higher than for the external parameters. Especially for maxima, the change is up to more than 10 times higher in some cases, for example in the Great Miami basin. It can also be noted that with the change of internal parameters, the degree of change in extreme events is almost always similar or higher than that of means.

A look at the way the curves changed, shows the expected contrary picture between the two types of simulations. If for the scenario with parameters decreasing by 10% events intensify, they tend to weaken for the simulation using increased model parameters and vice versa. Since it is not known, how exactly the parameters are related to the internal parameters of the basin, it is not purposeful to discuss the outcomes to whether extreme events become stronger or weaker in more detail. However, it is important to mention that, contrary to the simulations using changes in the external parameters, in most cases clearer tendencies towards more months showing either intensifying or weakening extremes can be observed. The clearest examples are minima in the Magdalena basin, where the ratio is 2:1, as well as maxima in the Great Miami basin with a ratio of 4:1 and in the Enns basin (3:1). Again, seasonality can be observed in the results.

In the following example of the Enns River, the curve only sharpens and both tails lower, with no apparent movement of the curve to either direction to be seen, which might be interpreted as an indication that the number of events showing a discharge amount similar to the expected value of the distribution become more frequent, and those that are more or less intense, become less frequent.

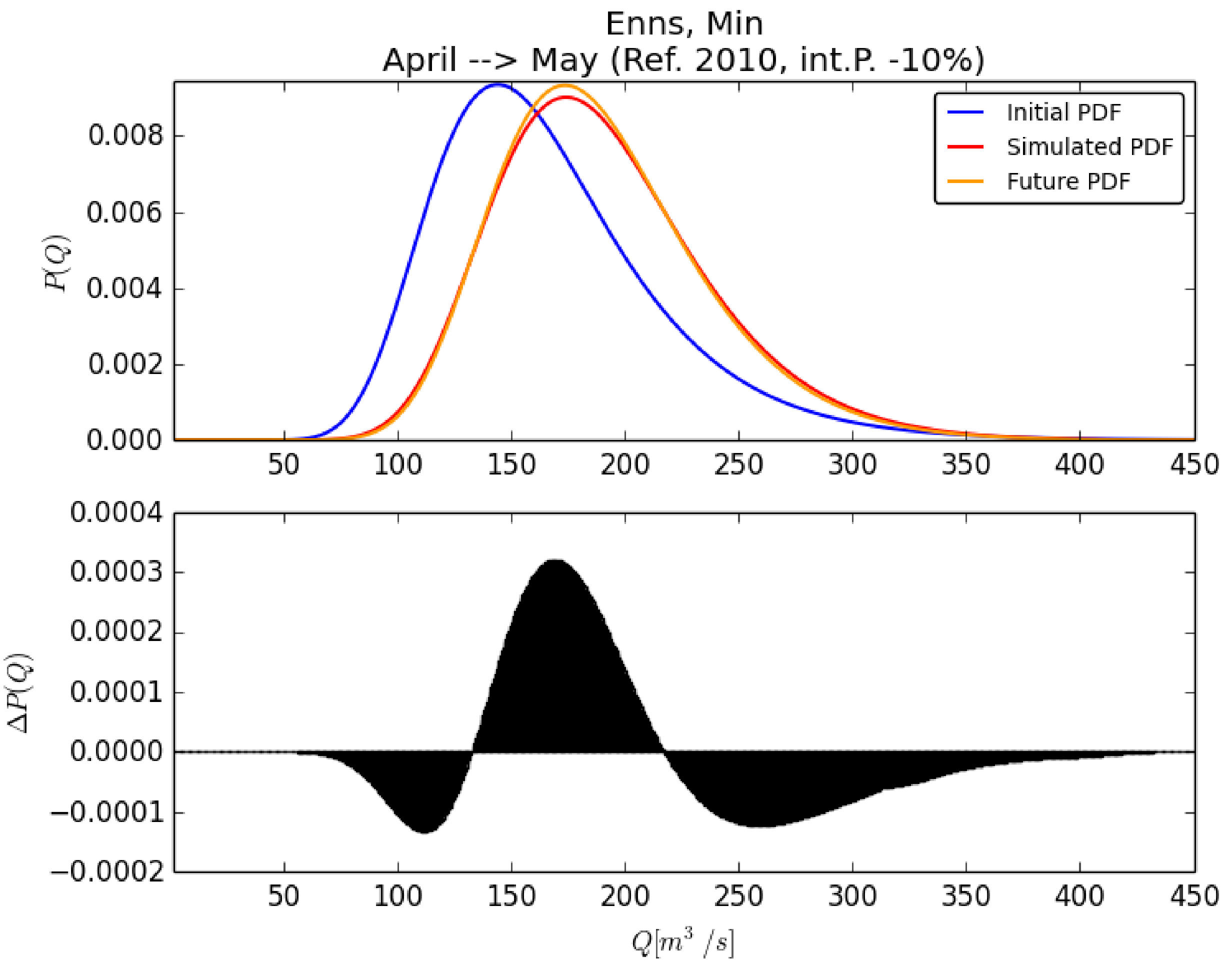

Figure 38. Simulation of the model applying a change of -10% to internal parameters in the Enns basin

The fact that alterations between the curves are stronger applying changes to the internal parameters is shown in the following example (figure 39) of the Upper Magdalena River, for which the differences were hardly noticeable applying changes in the external parameters. Due to the increase of 10% applied to the internal parameters, the curve of the future simulation indicates a higher expected value for the regime of discharge maxima for the month of October. It also allows the conclusion that stronger maxima might become more frequent due to the rise of the upper tail of the distribution. This graphic also shows that an increase of internal parameter values generally causes the curve to a state that indicates more intense maxima and at the same time less intense minima, or expressed in simple words, a move of the curve to the right side. The decrease in parameter values in the majority of the cases shows the opposing impacts.

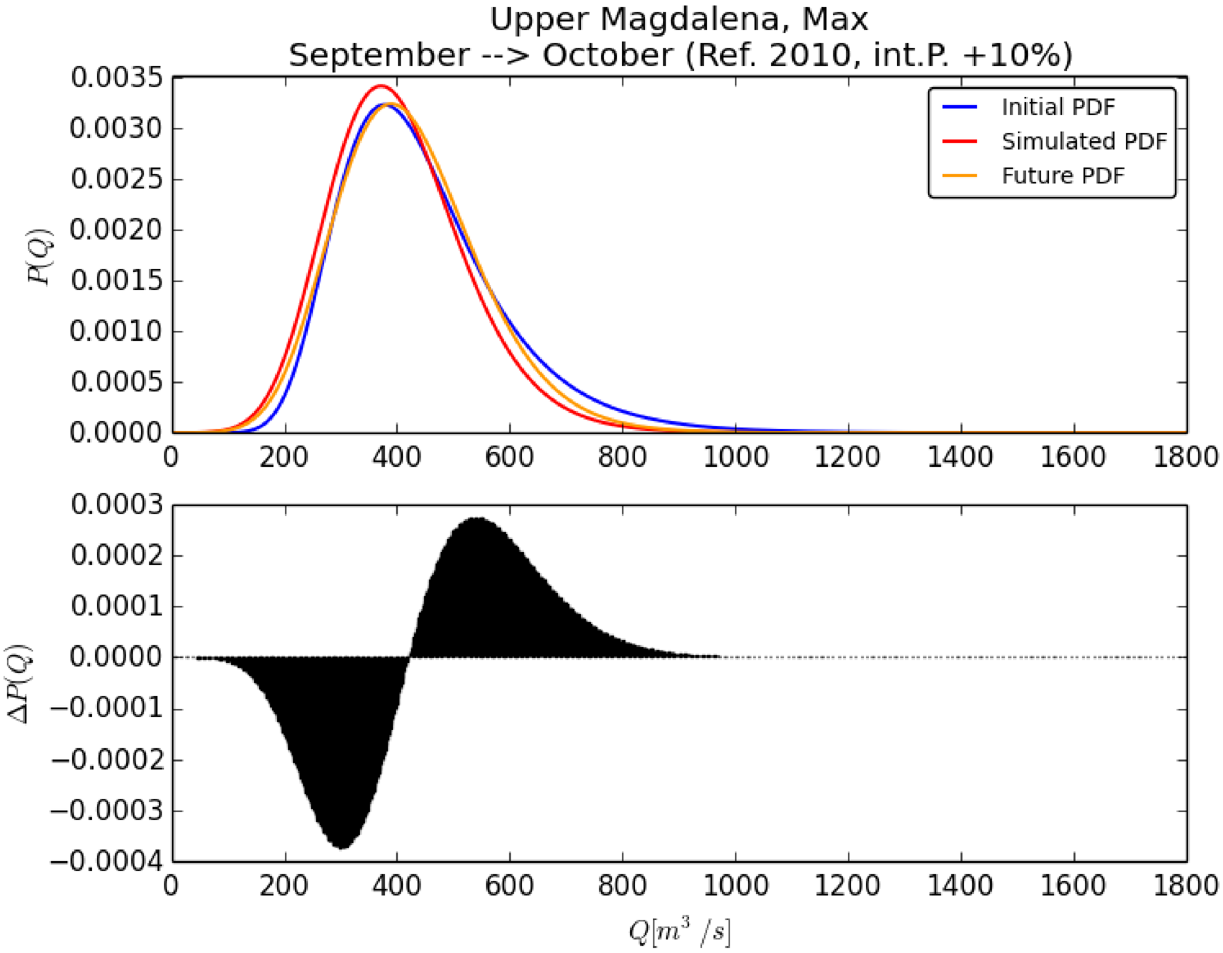

Figure 39. Simulation of the model applying a change of +10% to internal parameters in the Upper Magdalena basin

5.5.3 Simulation with external and internal parameters

The most probable situation for a simulation of a future scenario is that both external and internal parameters of the basin change. For this reason, both before described simulations were combined into one to predict a future situation in all 3 watersheds. Both 10% increase and decrease of the internal parameters were combined with the predicted precipitation and temperature values as they were used in 5.5.1.

	E: 20yr. trend, I:Change -10%			E: 20yr. trend, I:Change +10%		
Basin	**Mean**	**Maxima**	**Minima**	**Mean**	**Maxima**	**Minima**
Magdalena	0.0033	0.0036	0.0042	0.0038	0.0052	0.0046
Enns	0.0030	0.0019	0.0074	0.0036	0.0026	0.0193
Great Miami	0.0038	0.0053	0.0183	0.0090	0.0064	0.0156

Table 17. Averages of Kullback-Leibler divergences between present and future simulations applying changes to external and internal properties

In general, the results are similar in magnitude to those of the simulation with only internal parameters, in some cases however, such as the maxima in the Enns basin, the values increased significantly.

Again, in all basins it can be seen that the ratio of months with intensifying extremes to those with weakening extremes is not equal, as it was for the simulations using only internal parameters.

In some cases, it is shown that with the change of parameters chosen in this simulation, the regime of monthly flows changes. In the example of the Upper Magdalena River, an increase of monthly mean flows and probably maxima might be expected for the month of August, using the future precipitation values and a 10% increase in the parameters describing the internal properties of the basin.

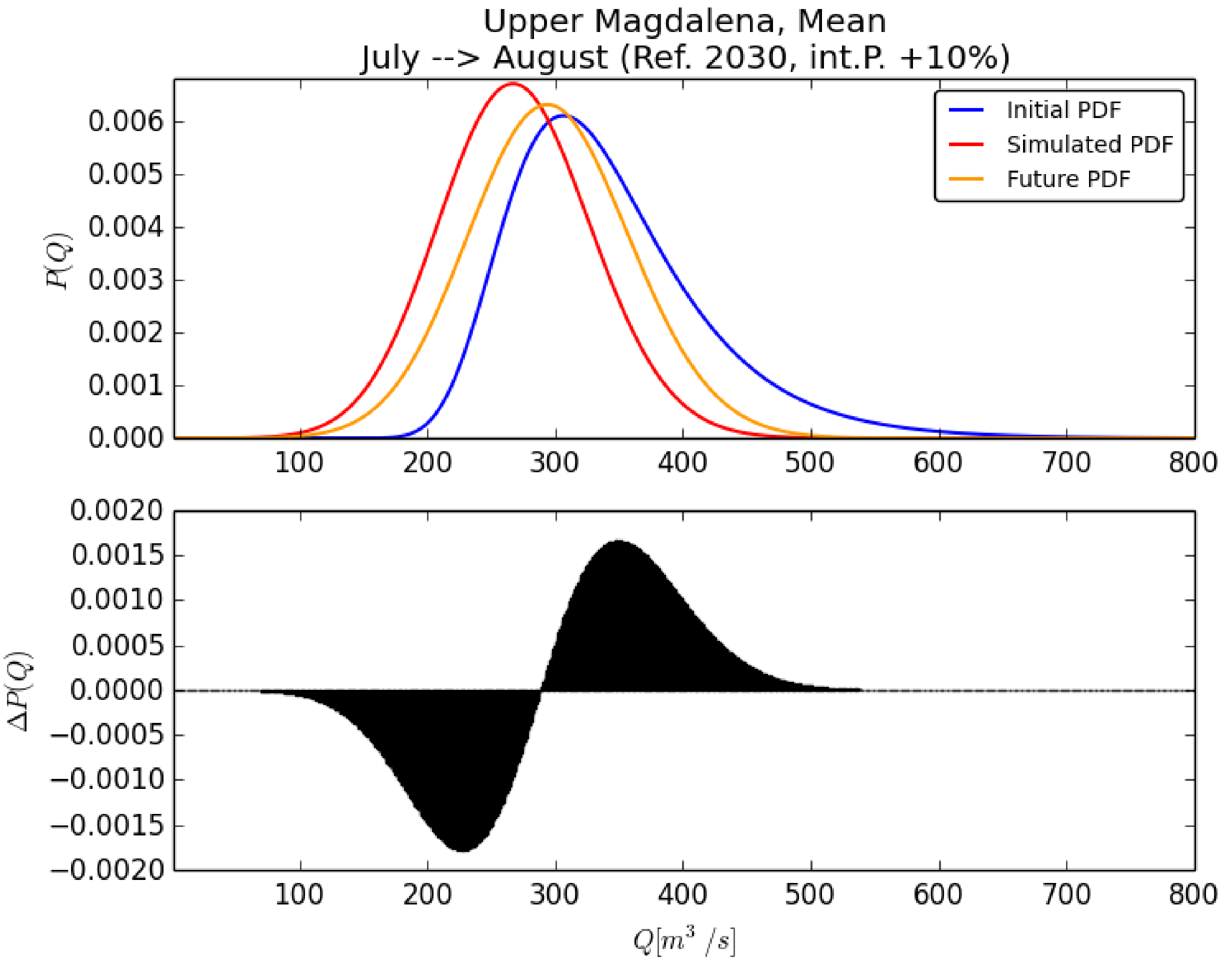

Figure 40. Simulation of the model applying changes to external and +10% to internal parameters, in the Upper Magdalena basin

As it had to be expected, the strongest changes could be achieved by applying changes to all model parameters, although they are not so much different from the changes only to the parameters representing internal properties of the system, both in magnitude and in patterns of occurrence. Furthermore, it is shown that changes to the external properties of the watersheds do not influence as strongly on the changes of hydrological regimes. This finding is applicable especially for changes in patterns of extreme events.

This leads to the conclusion that future scenarios, as far as they could be derived from existing alterations in time series which are probably caused by climate change, play a role in the alterations of the regimes of hydrometeorological extremes, although only a minor one. As the results indicate, changes in the regimes of extreme events are caused by a greater extent by changes in internal factors of the watershed, in which the activity of humans influences significantly. The global climate change would probably have to increase significantly compared to the current developments to have a comparable impact on the behavior of discharge extreme events. These conclusions are in accordance with those of the IPCC (IPCC, 2013).

As stated previously and in Bordi et al. (2009), the differentiation between trends and long-term periodicities of natural phenomena is difficult and cannot be known for data describing a short range of time, such as in the present study. Therefore, the method of applying future changes derived from past trends is only one of many possible scenarios that can be applied to the data and can therefore not be seen as a certain assumption. However, it was preferred to the strategy of deriving future hydrometeorological information from global circulation models due to its big uncertainties related to downscaling the data to fit the needs in a small area like the test basins (Buytaert et al., 2009; Shackley et al., 1998).

SECTION 6

CONCLUSIONS

This work was a first attempt to stochastically model the regimes of extreme events in hydrometeorological time series and possible mechanisms of alterations, which concentrated principally on discharge data. It provided insight into many of its details, but at the same time opened new possibilities for further research on the topic.

As in many other projects, data collection occupied a large part of the time spent for preliminary works. Due to the research during this stage, a comprehensive collection of data sources could be established. With this collection, a solid data base for research was created that consists of quality controlled information of a wide range of hydrometeorological variables that go back in time as far as the 1700s in some cases and are in most cases are updated to almost the current date, if data is still collected. The data coverage encompasses the whole range of the globe for 2 of the main variables of this research, precipitation and temperature. The collection of discharge data proved to be more difficult due to the lack of a freely accessible global database. However, a collection of different national data providers could be compiled, from which the coverage of data could be assured for large parts of the American continent, as well as Australia and some European countries, for which data can be retrieved at least in a semi-automated way. With the creation of interfaces for all the different data providers, a tool was created that enables the user to rapidly gather the necessary information and save it in a consistent data format, that permits easy further processing in different kinds of analysis.

Another achievement of preliminary work was the creation of a collection of programming tools, where approximately 50 functions were compiled for basic hydrological and statistical analysis. This collection of tools formed the foundation of functionality that permitted a rapid analysis of watershed data.

The conducted worldwide trend analysis indicated that a change in global climate is occurring and that for all of the studied variables, significant trends can be found, both in mean value and extreme value series. Some patterns of global change could be described, such as positive precipitation trends in the northern and negative ones in the southern hemisphere, as well as differences in the number of trends found at different time resolutions. In any case, changes in the patterns of hydrometeorological extreme events could not be proven by this trend analysis for the long term, mainly due to the short period of data observation available.

For this reason, a stochastic model was proposed that has its foundation in the Langevin equation, which has already been applied successfully in a wide range of hydrological applications. To solve this equation, a numerical solution of the Fokker-Planck-Kolmogorov equation was proposed, an equation that controls the evolution of the system's probability density function in time. For this reason, a stable algorithm was developed that permits the simulation of many different hydrological problems with the one-dimensional FPK equation, not only the study of extreme events. The algorithm is written in a way that it is easily expandable or applicable to other methods and is based on the implementation technique proposed by Dominguez and Rivera (2010). The way of implementation of the equation was therefore not a novelty, but improvements could be made, especially the implementation in Python code, which offers many new possibilities due to the large number of modules Python makes available, for example the possibility to implement optimization algorithms for the parameters of the FPK equation. Also, a considerable improvement in computing time was achieved due to the slicing technique in Numpy. This way, the code could be accelerated by approximately 100 times and created a foundation for the implementation of the multidimensional FPK equation in future investigations, which requires a lot more computing power.

Additionally, a reliable optimization procedure was developed that was able to optimize the complete set of parameters of the PDF translations for all variables in all basins. In any case, the results of this procedure could not be given a clear physical interpretation. The optimization with some fixed parameters did not work as well, but still provided acceptable results for more than 80% of the translations. It was also found that the optimization is extremely sensitive to the initial guesses that are provided for each PDF translation. It remains to be studied if a possibility exists to successfully optimize the application of "best initial guesses", which can be calculated from the model's parameters and offer the operator a higher possibility to find the set of optimized parameters more easily. The fact that for each translation, at least one set of optimal parameters can be found that permits a nearly perfect fit was shown in this work.

Also, it could be seen that wrong initial guesses can cause instability in the algorithm, especially if the initial diffusion vector includes predominantly negative values. The modeling of negative diffusion, which corresponds with a sharpening in the curve, could be achieved during optimization, as shown in the results of unsupervised optimization. Additionally, it was shown that in hydrological systems, which are clearly regulated, specifically with the example of the Wivenhoe dam in the Brisbane basin, the model cannot be applied successfully. In this case, the system does not respond to the natural variations of the system anymore, which is one of the principles of the model.

The proposed methodology of inverse modeling proved to be a powerful tool to effectively implement the studied problem. This way, the task of modeling a complicated physical process could be achieved without initially knowing its detailed structure. After the results of this investigation, it can be said that this modeling technique can most likely be applied to numerous other problems of hydrological modeling, for example in the assessment of hydrological risks, the study of hydrometeorological variability or further topics related to global climate change.

Finally, this work proposes a rigorous approach to estimate the degree to which the characteristics of a system change due to the alterations of its internal and external parameters. This way, evidence that global climate change impacts on the regimes of extreme events in discharge series was found, but it was also shown that changes in the internal parameters influence the system to a higher degree. Changes to internal parameters were further found to have a bigger impact on extreme events than on monthly means, which is in accordance with the finding that model parameters describing the fluctuations do not have a strong correlation with the changes in extreme events, but rather those describing the kernel of the system. These results and the fact that a large number of changes in internal parameters are caused by human activity reiterates the fact that we humans have it in our own hands to control the future of events like floods or water shortages due to high or low river flows by understanding better the internal mechanisms of river basins and taking the right steps to control the effects that these changes have on the occurrence of extreme events.

To understand completely all the mechanisms driving the alterations in the behavior of extreme events, further investigations are necessary that focus more profoundly on this topic. Especially works where the drift and diffusion parameters are linked to the basin’s land use and coverage parameters and to geomorphometric characteristics should be encouraged, since no clear allocation of the model parameters to specific internal parameters of the basins could be made.

REFERENCES

Abghari, H., Tabari, H., Hosseinzadeh Talaee, P., 2012. River flow trends in the west of Iran during the past 40 years: Impact of precipitation variability. Glob. Planet. Change.

Aguilar, E., Peterson, T.C., Obando, P.R., Frutos, R., Retana, J.A., Solera, M., Soley, J., Garcia, I.G., Araujo, R.M., Santos, A.R., 2005. Changes in precipitation and temperature extremes in Central America and northern South America, 1961–2003. J. Geophys. Res. Atmospheres 1984–2012 110.

Barros, V., Castañeda, M.E., Doyle, M., 2000. Recent Precipitation Trends in Southern South America East of the Andes: An Indication of Climatic Variability, in: Smolka, D.P., Volkheimer, P.D.W. (Eds.), Southern Hemisphere Paleo- and Neoclimates. Springer Berlin Heidelberg, pp. 187–206.

Bordi, I., Fraedrich, K., Sutera, A., 2009. Observed drought and wetness trends in Europe: an update. Hydrol. Earth Syst. Sci. 13.

Buytaert, W., Célleri, R., Timbe, L., 2009. Predicting climate change impacts on water resources in the tropical Andes: Effects of GCM uncertainty. Geophys. Res. Lett. 36, L07406. doi:10.1029/2008GL037048

Chaitin, G.J., 1969. On the simplicity and speed of programs for computing infinite sets of natural numbers. J. ACM JACM 16, 407–422.

Coles, S., 2001. An introduction to statistical modeling of extreme values. Springer.

Colomer, M., Montori, A., García, E., Fondevilla, C., 2014. Using a bioinspired model to determine the extinction risk of Calotriton asper populations as a result of an increase in extreme rainfall in a scenario of climatic change. Ecol. Model. 281, 1–14.

Dai, A., Qian, T., Trenberth, K.E., Milliman, J.D., 2009. Changes in continental freshwater discharge from 1948 to 2004. J. Clim. 22, 2773–2792.

Del Río, S., Herrero, L., Pinto-Gomes, C., Penas, A., 2011. Spatial analysis of mean temperature trends in Spain over the period 1961–2006. Glob. Planet. Change 78, 65–75.

Delgado, J.M., Apel, H., Merz, B., 2010. Flood trends and variability in the Mekong river. Hydrol. Earth Syst. Sci. 14, 407–418.

Denisov, S.I., Horsthemke, W., Hänggi, P., 2009. Generalized Fokker-Planck equation: Derivation and exact solutions. Eur. Phys. J. B 68, 567–575.

Dolgonosov, B.M., Korchagin, K.A., 2007. A nonlinear stochastic model describing the formation of daily and mean monthly water flow in river basins. Water Resour. 34, 624–634.

Dominguez, E., 2004. Stochastic forecasting of streamflow to Colombian hydropower reservoirs (PhD Thesis). Russian State Hydrometeorological University, St. Petersburg.

Dominguez, E., Rivera, H., 2010. A Fokker-Planck-Kolmogorov equation approach for the monthly affluence forecast of Betania hydropower reservoir. J. Hydroinformatics 12, 486–501.

Druzhinin, V., Sikan, A., 2001. Statistical methods for the treatment of hydrometeorological information. AM Vladimirov Ed.

European Parliament, 2007. Directive 2007/60/EC of the European Parliament and of the Council of 23 October 2007 on the assessment and management of flood risks.

Falvey, M., Garreaud, R.D., 2009. Regional cooling in a warming world: Recent temperature trends in the southeast Pacific and along the west coast of subtropical South America (1979–2006). J. Geophys. Res. Atmospheres 1984–2012 114.

Feldman, R.M., Valdez-Flores, C., 2009. Applied Probability and Stochastic Processes. Springer.

Frolov, A.V., 2006. Dynamic-stochastic modeling of long-term variations in river runoff. Water Resour. 33, 483–493. doi:10.1134/S0097807806050022

Gardiner, C.W., 2004. Handbook of stochastic methods: for physics, chemistry & the natural sciences.

Gell-Mann, M., 1995. The Quark and the Jaguar: Adventures in the Simple and the Complex. Macmillan.

Helsel, D.R., Hirsch, R.M., 2002. Statistical methods in water resources. US Geological survey.

Hirsch, R.M., Ryberg, K.R., 2012. Has the magnitude of floods across the USA changed with global CO2 levels? Hydrol. Sci. J. 57, 1–9.

Hoffman, J.D., Frankel, S., 2001. Numerical Methods for Engineers and Scientists, Second Edition,. CRC Press.

Hu, Y., Maskey, S., Uhlenbrook, S., 2012. Trends in temperature and rainfall extremes in the Yellow River source region, China. Clim. Change 110, 403–429.

IPCC, 2013. Climate Change 2013: The Physical Science Basis. Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.

Kolmogorov, A., 1931. Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. Math. Ann. 104, 415–458. doi:10.1007/BF01457949

Kolmogorov, A.N., 1965. Three approaches to the quantitative definition ofinformation’. Probl. Inf. Transm. 1, 1–7.

Koutsoyiannis, D., 2008. Probability and statistics for geophysical processes. National Technical University of Athens.

Koutsoyiannis, D., Yao, H., Georgakakos, A., 2008. Medium-range flow prediction for the Nile: a comparison of stochastic and deterministic methods/Prévision du débit du Nil à moyen terme: une comparaison de méthodes stochastiques et déterministes. Hydrol. Sci. J. 53, 142–164.

Kovalenko, V., 1986. Hydrometrical assessment of streamflow with an stochastic approach. Leningr. Politekh. Inst. 61.

Kovalenko, V., 2012. Method of characteristic applied in Fractionaly infinite hydrology. St. Petersburg RGGMU 134.

Kovalenko, V., Viktorova, N.V., Gaidukova, E., 1993. Modelling of hydrological processes. Guid. St. Petersburg 255.

Kozhevnikova, I., Shveikina, V., Domínguez, E., 2012. Modelling fluctuations of Caspian Sea levels using a mixed probability distribution. J. Flood Risk Manag. 5, 3–13.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Stat. 79–86.

Kundzewicz, Z.W., Pińskwar, I., Brakenridge, G.R., 2013. Large floods in Europe, 1985–2009. Hydrol. Sci. J. 58, 1–7.

Kunkel, K.E., Andsager, K., Easterling, D.R., 2010. Long-term trends in extreme precipitation events over the conterminous United States and Canada.

Lindström, G., Bergström, S., 2004. Runoff trends in Sweden 1807–2002/Tendances de l’écoulement en Suède entre 1807 et 2002. Hydrol. Sci. J. 49, 69–83.

Maldonado, C.E., 2009. Complejidad: revolución científica y teoría. Editor. Univ. Rosario.

Mass, C., Skalenakis, A., Warner, M., 2011. Extreme Precipitation over the West Coast of North America: Is There a Trend? J. Hydrometeorol. 12, 310–318.

matplotlib: python plotting [WWW Document], 2014. URL http://matplotlib.org/ (accessed 4.11.14).

Min, S.-K., Zhang, X., Zwiers, F.W., Hegerl, G.C., 2011. Human contribution to more-intense precipitation extremes. Nature 470, 378–381.

Moreno, J., 2011. El mecanismo de reforzamiento hidrológico de los procesos de calentamiento global - Caso de estudio Colombia - (Master Thesis). Pontificia Universidad Javeriana, Bogotá.

Morin, E., 2011. To know what we cannot know: Global mapping of minimal detectable absolute trends in annual precipitation. Water Resour. Res. 47.

Naidenov, V.I., Podsechin, V.P., 1992. A nonlinear mechanism of water level fluctuations of inland reservoirs. Water Resour 6, 5–11.

Naidenov, V.I., Shveikina, V.I., 2002. A nonlinear model of level variations in the Caspian Sea. Water Resour. 29, 160–167.

Naidenov, V.I., Shveikina, V.I., 2005. Hydrological Theory of Global Warming of the Earth's Climate. Russ. Meteorol. Hydrol. 46–56.

Nicholson, S.E., Nash, D.J., Chase, B.M., Grab, S.W., Shanahan, T.M., Verschuren, D., Asrat, A., Lézine, A.-M., Umer, M., 2013. Temperature variability over Africa during the last 2000 years. The Holocene.

Nikouei, A., Ward, F.A., 2013. Pricing irrigation water for drought adaptation in Iran. J. Hydrol. 503, 29–46.

NumPy — Numpy.org [WWW Document], 2014. URL http://www.numpy.org/ (accessed 4.11.14).

Nyeko-Ogiramoi, P., Willems, P., Ngirane-Katashaya, G., 2013. Trend and variability in observed hydrometeorological extremes in the Lake Victoria basin. J. Hydrol.

pandas: Python Data Analysis Library [WWW Document], 2014. URL http://pandas.pydata.org/ (accessed 4.11.14).

Performance Python: Solving The 2D Diffusion Equation With numpy | t-square [WWW Document], 2012. URL http://www.timteatro.net/2010/10/29/performance-python-solving-the-2d-diffusion-equation-with-numpy/ (accessed 5.26.12).

PyPI - the Python Package Index [WWW Document], 2014. URL https://pypi.python.org/pypi (accessed 4.25.14).

Python.org [WWW Document], 2014. URL https://www.python.org/ (accessed 4.11.14).

scipy.optimize.anneal — SciPy v0.14.0 Reference Guide [WWW Document], 2014. URL http://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.anneal.html#scipy.optimize.anneal (accessed 1.27.14).

scipy.optimize.curve_fit — SciPy v0.13.0 Reference Guide [WWW Document], 2014. URL http://docs.scipy.org/doc/scipy-0.13.0/reference/generated/scipy.optimize.curve_fit.html (accessed 4.25.14).

scipy.optimize.minimize — SciPy v0.14.0 Reference Guide [WWW Document], 2014. URL http://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.minimize.html#scipy.optimize.minimize (accessed 1.27.14).

SciPy.org [WWW Document], 2014. URL http://www.scipy.org/ (accessed 4.11.14).

Shackley, S., Young, P., Parkinson, S., Wynne, B., 1998. Uncertainty, complexity and concepts of good science in climate change modelling: Are GCMs the best tools? Clim. Change 38, 159–205.

Smith, J.A., Baeck, M.L., Morrison, J.E., Sturdevant-Rees, P., Turner-Gillespie, D.F., Bates, P.D., 2002. The regional hydrology of extreme floods in an urbanizing drainage basin. J. Hydrometeorol. 3.

Solomonoff, R.J., 1964. A formal theory of inductive inference. Part I. Inf. Control 7, 1–22.

Sonali, P., Nagesh Kumar, D., 2012. Review of trend detection methods and their application to detect temperature changes in India. J. Hydrol.

Spekkers, M.H., Kok, M., Clemens, F., Ten Veldhuis, J.A.E., 2013. A statistical analysis of insurance damage claims related to rainfall extremes. Hydrol. Earth Syst. Sci. 17.

Statistics (scipy.stats) — SciPy v0.14.0 Reference Guide [WWW Document], 2014. URL http://docs.scipy.org/doc/scipy/reference/tutorial/stats.html (accessed 4.25.14).

Sveshnikov, A.A., 1966. Applied methods of the theory of random functions. Pergamon.

SWAT | Soil and Water Assessment Tool [WWW Document], 2012. URL http://swat.tamu.edu/ (accessed 4.10.14).

UNESCO, 2014. The United Nations World Water Development Report 2014 - Water and Energy. Paris.

Vargas, W.M., Minetti, J.L., Poblete, A.G., 2002. Low-frequency oscillations in climatic and hydrological variables in southern South America’s tropical-subtropical regions. Theor. Appl. Climatol. 72, 29–40.

Wang, G., Xia, J., Chen, J., 2009. Quantification of effects of climate variations and human activities on runoff by a monthly water balance model: A case study of the Chaobai River basin in northern China. Water Resour. Res. 45.

Weickert, J., Romeny, B.T.H., Viergever, M.A., 1998. Efficient and reliable schemes for nonlinear diffusion filtering. Image Process. IEEE Trans. On 7, 398–410.

Woo, M., Thorne, R., Szeto, K., Yang, D., 2008. Streamflow hydrology in the boreal region under the influences of climate and human interference. Philos. Trans. R. Soc. B Biol. Sci. 363, 2249–2258.

Xu, Z., Liu, Z., Fu, G., Chen, Y., 2010. Trends of major hydroclimatic variables in the Tarim River basin during the past 50 years. J. Arid Environ. 74, 256–267.

Yue, S., Pilon, P., Cavadias, G., 2002. Power of the Mann–Kendall and Spearman’s rho tests for detecting monotonic trends in hydrological series. J. Hydrol. 259, 254–271.

Annex A: List of data sources

Hydrological Data:

Worldwide Databases:

- Global Historical Climatology Network (GHCN)
  http://www.ncdc.noaa.gov/oa/climate/ghcn-daily/
  ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/

- Climate Data Online
  http://www.ncdc.noaa.gov/cdo-web/

- Global Runoff Data Centre (GRDC)
  http://www.bafg.de/GRDC/EN/Home/homepage_node.html

National Databases

- Argentina: Sub secretary of Hydrological Resources (Subsecretaría de Recursos Hídricos): BDHI database
  http://bdhi.hidricosargentina.gov.ar/sitioweb/frmInicial.aspx

- Australia: Bureau of Meteorology
  http://www.bom.gov.au/water/hrs/#panel=data-explorer

- Australia: Government of Queensland Water Monitoring Portal
  http://watermonitoring.derm.qld.gov.au/host.htm

- Austria: Ministry of Life: eHyd database
  http://ehyd.gv.at/

- Brazil: National Agency for Water (Agência Nacional de Águas, ANA): Hidroweb database
  http://hidroweb.ana.gov.br/

- Canada: Environment Canada (HYDAT database)
  ftp://arccf10.tor.ec.gc.ca/wsc/software/HYDAT/

- Mexico: National Commission for Water (Comisión Nacional del Agua, CONAGUA)
  ftp://ftp.conagua.gob.mx/

- South Africa: Department of Water Affairs
  http://www.dwaf.gov.za/Hydrology/hymain.aspx

- United Kingdom: Centre of Ecology and Hydrology
  http://www.ceh.ac.uk/data/nrfa/data/search.html

- United States: United States Geological Survey (USGS): National Water Information System
  http://waterdata.usgs.gov/nwis/

Geospatial Data:

- ArcSWAT
  http://swat.tamu.edu/

- European Environmental Agency – Ecrins dataset
  http://www.eea.europa.eu/data-and-maps/data/european-catchments-and-rivers-network

- NASA World Satellite Web Map Service
  http://wms.jpl.nasa.gov/wms.cgi?request=GetTileService

- NOAA United States river dataset
  http://www.nws.noaa.gov/geodata/catalog/hydro/metadata/riversub.htm

- USGS Hydrological Units dataset
  http://water.usgs.gov/GIS/metadata/usgswrd/XML/huc250k.xml

- USGS HydroSHEDS project
  http://hydrosheds.cr.usgs.gov/index.php

- World Country Boundary Shapefile
  http://geocommons.com/overlays/33578.html

Annex B: List of created Python functions

This annex contains the most important functions created for the study. Some of the functions contain subfunctions which are not listed below.

Module hydroscripts.py

Data preparation

- Prepare data as daily data matrix (stochastic process form)
- Create monthly and annual data matrices from daily matrix
- Create monthly time series from daily time series
- Create monthly data matrix of monthly maxima or minima from daily matrix

Statistical tests

- Linear regression test
- Kolmogorov goodness of fit test
- Kullback-Leibler divergence criterion
- Autocorrelation function
- Test of statistical independence (Streak test) of a random variable
- Test of homogeneity of a random variable
- Correlation moment of a random process
- Cross correlation function of a random process
- PDF function fit to a random variable
- Determine best PDF fit to a random variable
- Mann Kendall trend test
- Evaluate reference rainfall lag time for monthly discharge extremes
- Create matrix of reference rainfall for monthly discharge extremes

Module dataops.py:

- Download data of GHCN stations
- Prepare dataset of GHCN stations
- Prepare dataset of USGS stations
- Prepare dataset of GRDC stations
- Create list of eHyd stations from station metadata
- Prepare dataset of stations from eHyd (Austria)

- Prepare dataset of stations from BDHI (Argentina)
- Prepare dataset of stations from ANA (Brazil)
- Prepare dataset of stations from CONAGUA (Mexico)
- Prepare dataset of stations from BOM (Australia)
- Extract data from HYDAT database and prepare dataset of stations (Canada)
- Calculate percentage of missing data in time series

Module fpk.py:

- Explicit FPK (adaptation of existing code)
- Implicit FPK (adaptation of existing code)
- Run FPK with parameters
- Optimize FPK parameters
- FPK optimization for all random processes of all basins
- Supervised FPK optimization for all random processes of all basins
- Evaluation of FPK with parameter alterations for all basins