Putting hydrological modelling practice to the test

Lieke A. Melsen

Putting hydrological modelling practice to the test

Lieke Anna Melsen

Thesis committee

Promotor
Prof. Dr R. Uijlenhoet
Professor of Hydrology and Quantitative Water Management
Wageningen University & Research

Co-promotors
Dr A.J. Teuling
Associate professor, Hydrology and Quantitative Water Management Group
Wageningen University & Research

Dr M.P. Clark
Senior scientist, Hydrometeorological Applications Program
National Center for Atmospheric Research, Boulder, USA

Other members
Prof. Dr M.F.P. Bierkens, Utrecht University
Prof. Dr G.J.M. De Lannoy, KU Leuven, Belgium
Prof. Dr F. Ludwig, Wageningen University & Research
Dr C. Perrin, IRSTEA, France

Putting hydrological modelling practice to the test

Lieke Anna Melsen

"Theories crumble, but good observations never fade."
- Harlow Shapley

Foreword

Harlow Shapley (1885-1972), the person to which the words on the previous page have been assigned, was an American astronomer, mainly known for his studies on the size of the Milkyway. However, this thesis does not contain anything related to astronomy, nor to the Milkyway. I also have to disappoint everyone who expects a thesis based on my own measurements or observation techniques. For this complete thesis, not a single measurement has been carried out by myself. You might wonder why I decided to use this quote at all. The reason is that it reflects what I have learned during my thesis work. During a discussion for a workshop in Bertinoro, Italy (2016), hydrological models were at some point described as "*a sandwich of theories*". Testing hydrological models is thus testing a set of theories. But, in order to test theories, observations are needed to confirm or reject the theory. If we want to increase our understanding of planet Earth, we have to critically reflect on our theories, and extend or reject them. And that is where more, and new types of observations are needed. It should, however, not be neglected that theories and observations are linked and dependent in many ways, and that there is still a path ahead in developing both new theories and new ways to observe.

Contents

Chapter 1

Introduction

It is tempting to start this introduction with stressing the importance of hydrology, followed by an enumeration of hydrological catastrophes, supported by numbers of casualties or economic losses. The goal of this thesis, however, is not to prove how important hydrology is in every day life, but to evaluate hydrological modelling practice, by applying several models at different scales. In order to understand why models are applied the way they are nowadays, and to understand why I consciously left out the buzzwords 'flood', 'drought', or 'climate change' in the first sentence, we first need to take a small step back in the history of hydrology.

1.1   A short history of hydrology

Water is essential for life. It is therefore not surprising that even for the oldest civilizations, evidence has been found for hydrologic engineering. One of the oldest examples dates back to 3000 years B.C., when Pharao Menes ordered the construction of a dam in the Nile, which diverted the course of the river (*Biswas*, 1970). Around that same time, in several cultures, from Mesopotamia to the Indus valley, water supply and irrigation systems were employed. Hydrologic engineering is thus as old as human civilization. Indeed, understanding the relation between water availability, water management and social structures is considered one of the key themes in archaeological research (*Mithen and Black*, 2011).

The interest in water itself, rather than in its engineering application, started only many centuries later with the Greek philosopher Thales from Milete ($\sim$600 B.C.). Thales tried to explain what he observed around him, and concluded that water is "*the source of everything*" (*Biswas*, 1970; *Fearn*, 2001; *Ten Bos*, 2014). Although Thales leans on mythology in many parts of his work, his manuscripts are often recognized as the beginning of Western science (see e.g. *O'Grady*, 2016), from which it can be concluded that water played an important role in the beginning of science.

Thales' approach stimulated the search for theories to explain the environment. *Brutsaert* (1988) provides a comprehensive overview on the history of the perception of evaporation. Here, however, I will discuss some historical theories on how river runoff is gener-

ated, since this is more related to the topics in this thesis. Plato (∼400 B.C.) formulated two hypotheses to explain the occurrence of rivers and streams. Although he refers to rain in some work as water source, his most known theory is based on '*Tartarus*', an endless underground water basin where all streams and rivers originate from (*Biswas*, 1970). This theory was heavily criticized by his pupil Aristotle, who formulated a theory for rivers and streams which is based on what we currently know as the water cycle, although in his theory water condensates inside the earth like clouds, which cause rivers and streams to flow (*Baker and Horton*, 1936). Over the centuries, many alternative theories have been proposed, e.g. Leonardo da Vinci's water veins which supposedly transported water from the sea through the Earth into the mountains to flow out as rivers (*Pfister and Savenije*, 2006). Aristotle's hydrological theories have not been formally rejected for over 2000 years. Only in the 17th century, when Francis Bacon formalized the empiricist method in science in his work "*Novum Organum*" (1620), theories were tested against observations. Quantitative hydrology developed: hydrological states and fluxes were quantified with the use of observations. Pierre Perrault (∼1650) is considered the first one who demonstrated, based on observations, that precipitation as a water source is, in numbers, sufficient to feed the French river Seine (*Dooge*, 1959; *Gottschalk*, 1992). Perraults' study confirmed the pluvial theory.

The notion of hydrology as a separate branch of science began to grow in the 18th century. Hydrology is, according to the Oxford dictionary, "*the branch of science concerned with the properties of the earth's water, and especially its movement in relation to land*". The earliest notation of this word, according to the Oxford dictionary, stems from 1762. Although the previous sections describe the starting point of hydrology as a science in the old Greek civilization, 1762 was the year that Paul Frisi published his work "*Treatise of Rivers and Torrents*", considered by *Horton* (1931) as the work in which the discussion on the origin of rivers and streams is finally settled in favour of the pluvial explanation. *Horton* (1931) considered hydrology therefore as a new science, at time of his writing "*little more than a birth-certificate*".

Based on this very short and far from complete history of hydrology, it can be concluded that Thales from Milete was the first one to ask questions about the source of water, the quantitative hydrologists in the 17th and 18th centuries were the first ones to answer these questions according to current scientific practice in which theories are tested with observations. The establishment of scientific practice, mainly by induction, has been developed further in the 19th and 20th centuries (*Chalmers*, 1999).

1.2 Science versus engineering

Science is defined as '*knowledge for its own sake*', in case of hydrology focused on increasing our understanding of the hydrological cycle. Engineering, on the contrary, is focussed on solving practical problems. Thales' interests in water itself rather than its application was a breakpoint to distinguish science from engineering. Plato's '*Tartarus*' theory or da Vinci's water veins theory did not directly solve any practical problem but were attempts to understand system Earth, and can therefore clearly be classified as science (although not according to current day scientific practice).

The introduction of quantitative hydrology, where hydrological fluxes are quantified with observations to test theories, seems to bring science and engineering together again. The development of quantitative hydrology lead to the formulation of mathematical rainfall-runoff models to describe hydrological phenomena, for example with the Rational Method proposed by *Mulvany* (1851) to describe the relation between time of concentration and peak flow (*Todini*, 2007). Because mathematical models can have a clear practical application, in contrast to the theories that Aristotle and others had formulated over the last 2,400 years, science started to support engineering: increased understanding of the system could directly be applied to solve practical problems. The Industrial Revolution in the 19th century and the associated urbanization increased the demand for practical knowledge, which pushed science further at the service of engineering (*Eagleson*, 1991).

Defining hydrology as a pure science suffers from several difficulties. One difficulty is of practical nature and inherent to all Earth and Life sciences: fully controlled experiments are difficult to conduct due to the large number of variables and the lack of clear system boundaries (in contrast to e.g. hydraulic experiments in the lab). Hydrological 'experiments' are therefore seldom repeatable. A natural extreme event, for example a flood or a drought, can neither be repeated nor be completely controlled. This leads to uncertainty: there are hidden variables in the complex system Earth, and the researcher has to rely on historical data. Other obstacles to perceive hydrology purely as a science have been brought forward by *Yevjevich* (1968), *Klemeš* (1986a), and *Eagleson* (1991). *Yevjevich* (1968) states that hydrology as a science is often downgraded to an appendage of hydraulic engineering. *Klemeš* (1986a) agrees with this, stating that hydrology is often only considered as a "*dull subject which one had to take*" as part of a study in other sciences such as geology. A clear turning point was advocated by the *National Research Council* (1991) in their book "*Opportunities in the Hydrologic Sciences*", which provides constructive directives to improve the position of hydrologic science, especially compared to hydrologic engineering: "*We cannot build the necessary scientific understanding of hydrology at a global scale from the traditional research and education programs that have been designed to serve the pragmatic need of the engineering community*" (p.1). Also the *Foresight Committee on Hydrological Science in The Netherlands* (2005) presented an exploratory study on how hydrologic sciences should develop in the Netherlands, but unfortunately the *Foresight Committee on Hydrological Science in The Netherlands* (2005) could not resist the temptation to link hydrologic research directly to societal relevance. Despite the pleas from Yevjevich and Klemeš and the massive effort from the *National Research Council* (1991) more than twenty years ago, hydrology programmes are often still linked to environmental sciences or engineering programmes.

The history of hydrology shows that hydrologic engineering is centuries older than hydrology as a science, for the logic reason that civilizations need hydrologic engineering to develop. It is tempting to apply hydrological knowledge to solve practical problems (engineering), rather than to further investigate state-of-the art hydrological knowledge (science). Scientific knowledge does not exclude engineering or vice versa. Hydrology is therefore sometimes described as an 'Engineering science' (*Sivapalan*, 2003), where sci-

ence serves society. This term, however, can be misleading because the type of research and the goal of research are different for engineering and scientific purposes. Current day scientific literature often shows how both approaches are mixed, admittedly not in the last place in the chapters of this thesis, where almost every introduction starts with summing up the practical application of the hydrological knowledge gained.

1.3   The philosophy of hydrological modelling

A mathematical formulation of hydrologic theory, what quantitative hydrologists strived for, can be perceived as a quantitative hydrological model. I adopt the definition of a hydrological model from *Savenije* (2009): "*a hypothesis of how a system works, codified in quantitative terms*". A hydrological model can consist of a single mathematical description of a hydrological phenomenon, or it can consist of a set of equations describing several parts of the hydrological cycle. Models are valuable instruments in modern science, but they give way to several questions concerning the philosophy of science (*Frigg and Hartmann*, 2012), for example in semantics (if a model is an abstract entity, how can it have explanatory power for the physical world? See *Psillos*, 2011), epistemology (what do we learn from models?) and in general philosophy of science (how do models relate to theory?). The answers to these questions, however, depend on whether hydrology is considered as science or as engineering. In the next paragraph, I will give an example of both perceptions.

Oreskes et al. (1994) advocates that models in Earth sciences should only be used for heuristic purposes, "*useful for guiding further study but not susceptible to proof*". Models should be used to test theories, but theories can only be falsified, or at best, confirmed. In contrast, black box models do not provide insight in the system and generally lack explanatory power, but do provide a prediction which can be used for practical application (in other words: engineering). The heuristic approach versus the black-box approach is an example which clearly demonstrates the contrast between the goal of modelling in science and in engineering. In the scientific approach, the focus of the research is on the model itself and a correct understanding and parameterization of the processes, i.e., on being "*right for the right reasons*" (*Kirchner*, 2006), whereas for the engineering approach, the focus is on the model output, which is relevant and should be correct (also if this implies being right for the wrong reasons). Or, as expressed by *Savenije* (2009): "*From the perception of the engineer (...) a model is essentially a tool.*" The difference between a scientific and an engineering approach in hydrological modelling can be diminished or further emphasized, dependent on which scientific philosophy is followed.

Three philosophies of science can be distinguished in which theories are linked to observations (thereby ignoring more overarching views on the process of gaining scientific knowledge, for example by Feyerabend and Kuhn); verificationism, falsificationism, and Bayesianism (*Chalmers*, 1999). In the following paragraphs, a very short summary of each philosophy is given, neglecting important reflections and the development over time. We refer to *Chalmers* (1999) and *Carnap* (1966) for a clear overview of scientific philosophies and the advantages and drawbacks of each philosophy.

In verificationism, also referred to as logical positivism, a theory is formulated, based on

observations (inductivism). Subsequently, the theory is verified against more observations. If the observations do not confirm the theory, the theory is discarded. New observations can also stimulate to further develop, adapt, or extend the theory. If the theory is supported by the observations, the theory becomes "more true". Verificationism requires hypothesis-testing and rejects every statement that cannot be tested. The Vienna Circle (*Wiener Kreiss*), a group of scientists at the beginning of the 20th century, where the first ones to aim to formalize the laws of verificationism, although full agreement was never reached. The pressing question how often a theory has to be confirmed by observations in order to be "true" has never been answered.

Half way the last century the theory of falsification developed in a response to logical positivism. Karl Popper, the main founder of this theory, specifically criticized the assumption of verificationism (and inductivism) that observations are objective and free of theories in itself, and denounced the unsatisfactory feeling that a theory can only be proven to be true when it is supported by an infinite number of observations. Popper adopted the practice of hypothesis-testing from the logical positivists, but added a second requirement: the hypothesis should be falsifiable (*Popper*, 1968). In other words, it should be possible to proof that the hypothesis is *not* true. A scientist should constantly search for proof to reject their theories. Rejected theories are replaced with new theories, which scientists then again try to falsify. There is a clear drive to advance knowledge rather than to confirm knowledge.

Bayes' theorem has become popular with the advance of computational capabilities, but is based on theory of the 18th century mathematician and Presbyterian minister Thomas Bayes. The probability for a given theory to be true is defined based on available knowledge, the so-called prior. Defining the prior is clearly an inductive element of Bayesianism. For every new observation, the likelihood of this observation based on the prior is determined, which then determines the probability of the theory to be true after the observation - the so-called posterior. Adding more observations will constantly influence the posterior. Bayesianism can be perceived as a mathematical formalization of verificationism, because the posterior probability of a theory to be true is a quantification of the degree of belief that a theory is true.

In developing a hydrological model, different steps can be distinguished (*Beven*, 2012). I will demonstrate in which steps the differences between a scientific and an engineering approach and among the three scientific philosophies are most pronounced. Of course philosophies generally do not engage in structural schemes, and I do not consider all the subdivisions within each philosophy. Therefore, this section should be read as a broad directive rather than a complete overview.

1.4 The six steps of hydrological modelling

I. The perceptual model - deciding on the processes

The perceptual model is the conceptualization of the system. A model is always a simplification of reality, but the degree of simplification is a gradual transition. Usually, the conceptualization is determined by the goal of the study. Engineers are not necessarily

interested in the underlying processes of the system, but rather in the practical applicability. Therefore, user-friendliness of the model and the availability of the required input data are considerations. This can be a motivation to simplify the conceptualization or choose for a black box approach. In a scientific approach, the perceptual model should be based on a theory or a set of theories (*Clark et al.*, 2016a), where each theory can be considered as a hypothesis of (a part of) the real world's functioning. These considerations explicitly demonstrate the difference between a scientific and an engineering approach.

II. The conceptual model - deciding on the equations
After the perceptual model is formulated, the equations to describe the processes in the perceptual model should be defined. An engineer can choose for a model which has shown to perform well under the given circumstances, or identify a statistical regression model based on available data. A scientist should develop equations which describe the theories that were defined in the previous step. In an inductive approach, theory follows from observations. Consequently, within verificationism and Bayesianism, statistical regression between some variables to describe the processes in the perceptual model could be applied. Falsificationism, contrastingly, stresses that observations often depend on the background knowledge and available measurement techniques in a scientific branch, and therefore acknowledges the drawbacks of, for example, statistical regression. New equations can be defined based on the conceptual understanding of the system, rather than on observations alone. In this step of model development, verificationism and Bayesianism are closer to engineering, whereas falsificationism diverges from the engineering approach.

III. The procedural model - getting the code to run on a computer
Many (physical) equations do not have an analytical solution, and therefore need a numerical solver. The choice of the numerical solver, however, can have a large impact on the model results (*Clark and Kavetski*, 2010; *Kavetski and Clark*, 2010). An engineer can choose the Euler forward method because it is easy to implement, or the Runge-Kutta scheme because it is more stable; after all the main motivation for an engineer is to have a good model performance. A scientist should be more careful, because the numerical scheme impacts the model output and therefore can lead to a false rejection or confirmation of the tested theory. All the scientific philosophies accept the risk to reject (or adapt the posterior in case of Bayes' theorem) a theory based on model output that was manipulated by the numerical solver. The opposite can also happen, which is even more worrisome: the verificationist risks to confirm a theory which might actually be corrected due to the numerical scheme.

IV. Calibration - getting values for parameters
In the context of hydrological modelling, calibration is defined as "*tuning the parameters to fit the model output with observations*" (*Steele and Werndl*, 2015). From a certain perspective, calibration is by definition engineering, because the model is adapted for a particular situation, which would imply that the model does not describe the system correctly (*Steele and Werndl*, 2015). This is not necessarily true. The equation ($\sim$theory) describing the relation between two variables at a particular scale could be correct, but calibration can be

needed in order to estimate the parameters at the applied scale (the so called 'effective parameters'). Furthermore, the model can contain conceptual parameters which are not directly measurable in the field and therefore need calibration. Calibration of hydrological models is therefore almost inevitable (*Beven and Cloke*, 2012; *Koskela et al.*, 2012). It is, admittedly, difficult to distinguish between wrong equations of the underlying theory, and scaling of parameters.

Another recurring issue in hydrological modelling is that the optimization problem is ill-defined; the information in the data does not provide sufficient information to identify unique parameter values (a so-called ill-posed problem), leading to equifinality (*Beven*, 2006c). A rule of thumb is that discharge time series generally only provide sufficient information to identify 4 or 5 parameters (*Jakeman and Hornberger*, 1993). This makes calibration of a distributed model based on discharge data, where, in principle, each cell could have multiple unique parameter values, highly non-trivial.

The theory of how to calibrate a model is closely related to modelling steps V and VI: evaluation and uncertainty analysis. *Klemeš* (1986b) proposed the differential split sample test as calibration procedure for operational (engineering) models. The same procedure is advocated by *Kirchner* (2006) and *Andréassian et al.* (2009). The differential split sample test can be applicable for engineering, verificationism, and falsificationism (but see also the next section on evaluation). In Bayesian theory, a split sample test is not a prerequisite: it is even inherent to the theory not to apply a split sample, because it is part of the theory to add new information to the model and adapt the model based on the new information (*Steele and Werndl*, 2015). Calibration algorithms have been developed based on Bayes' theorem, for example DiffeRential Evolution Adaptive Metropolis (DREAM[1], *Vrugt et al.*, 2008), but these algorithms are often still applied in a split-sample framework.

A thorough calibration procedure is important for the same reason as mentioned under step III: the underlying theory of a model can be falsely rejected (or accepted in an verificationist framework) if the parameters have been carelessly identified.

V. Evaluation - confronting model output with observations

'Validation' is currently the most widely used term in hydrology when model output data are confronted with independent observations, and it is also the term that is used in this thesis, but it is not undisputed. *Oreskes et al.* (1994), advocating the falsificationist approach, states that models in Earth sciences can never be verified ('establishment of truth') or validated ('establishment of legitimacy'), because of the lack of a closed and controlled system. At its best, models can be *confirmed* by the data. In contrast, model confirmation is a way of model verification for a verificationist or an engineer. The model evaluation procedure clearly drives a wedge between verificationists and falsificationists. Furthermore, falsification theory stimulates a more thorough model evaluation than verifica-

---

[1]It is important to note that DREAM only considers the calibration-step in a Bayesian framework, not the complete theory testing.

tionism does, e.g. regarding internal consistency.

VI. Uncertainty analysis - estimating the uncertainty of the model (output)
For an engineer, uncertainty analysis could be relevant in order to help in decision making, but when it for example comes to hydrologic infrastructure design, the motivation is to be on the safe side, not to thoroughly understand the uncertainty in the model. For scientists, *Beven* (2006a) states that it is simply unscientific nót to do uncertainty analysis, which is a valid point given the difficulties of Earth sciences concerning controlled and repeatable experiments. Hydrological models are prone to uncertainty, for several reasons (*Wagener et al.*, 2004);

1. Measurement, model, and scaling errors in the input data

2. Unknown values of the parameters

3. The model structure does not capture all relevant processes

Uncertainty analysis is needed because it can be a reason to not reject a theory despite a bad model performance, for example when it can be shown that the model is not performing well because of the input data, rather than because of the underlying theory. It can also be a reason to reject a theory despite good model performance, because the parameters show to be highly uncertain. How the uncertainty analysis should be conducted is a matter of dispute or even a matter of taste. In Bayesianism, uncertainty analysis is inherent, because with a Bayesian approach probabilities are assigned, which already indicate a notion of uncertainty. Bayesian approaches like DREAM (*Vrugt et al.*, 2008) are popular, but *Beven* (2006a) argues that many assumptions of a formal Bayesian analysis cannot be proven to be valid in hydrological models (see p.44 where some of these assumptions are discussed and tested). To relax these assumptions, *Beven and Binley* (1992) have proposed the Generalized Likelihood Uncertainty Estimation (GLUE) method, which indirectly accounts for input data and parameter uncertainty.

Based on the different steps in developing a hydrological model, it was shown that the largest differences between a scientific and an engineering approach occur in the perceptualization and conceptualization of the system. It was shown that verificationism sometimes closely resembles the engineering approach, whereas falsificationism clearly calls for a different practice, especially in model evaluation.

In many hydrological modelling studies science and engineering approaches are mixed. This is worrying, because the goal of both disciplines is different. On the basis of six steps in hydrological modelling, the difference between the scientific and the engineering approach was demonstrated. Without judging or expressing a preference for any of the philosophies of science (yet), which all have their advantages and drawbacks, it is important that a scientist is aware of the applied approach (some implications are nicely discussed in *Nearing et al.*, 2016), and critically reflects on this approach. Considering hydrology as an 'engineering science' does not stimulate thorough and critical analysis of the system under study.
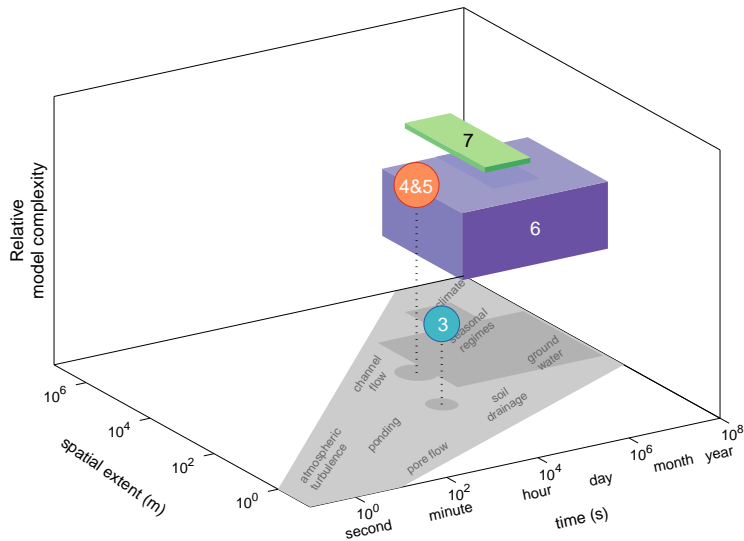
Figure 1.1: Overview of the spatial and temporal scales and the complexity of the employed hydrological models (defined here as the degree to which involved processes are explicitly described) covered in this thesis. The numbers refer to the chapters of this thesis. The x, y-axes are based on Figure 7.1.

1.5   Positioning of the chapters in this thesis

The first three modelling steps (perceptual, conceptual, procedural) as discussed in the previous section in practice often result in simply using an existing model, where the model choice might sometimes be motivated by experience within the research group, available time in a (PhD-)project, habits, or code availability, rather than on the perceptual understanding of the system. The available hydrological models vary widely in complexity; from highly conceptual models describing only the main components of runoff generation (like the unit-hydrograph concept), to fully coupled surface water-ground water models based on physics (*Paniconi and Putti*, 2015). The hydrological cycle comprises many processes that cover a broad range of spatial and temporal scales, as demonstrated in Figure 1.1. Therefore, the spatial and temporal scales covered by hydrological models also varies greatly; from models that only describe the processes at the smallest scale, e.g. pore flow and capillary rise, to models that describe large-scale processes such as long-term climate response or hydrological drought. Summarizing, hydrologic science concerns processes that cover a broad range of spatial and temporal scales and can be approached with models of a wide variety of complexities. The models can be applied to many different basins for which observations are available; all in all, many (subjective) choices for a modeller (as discussed in Chapter 5 and demonstrated in Figure 5.1).

In this thesis, I put hydrological modelling practice to the test and therefore apply widely used hydrological models; in other words, I do not take the royal scientific modelling road either, where I should start with a perceptual understanding of the hydrological system and hypothesize about the dominant processes, but rather use readily available and commonly used hydrological models which substitute modelling step I, II, and III. A clear example of mixing engineering and science approaches, since almost all of the selected models have been developed mainly for projections (as discussed in the next chapter).

Figure 1.1 provides an overview of the spatial and temporal scales covered in each of the chapters of this thesis, and the relative complexity (defined here as the degree to which involved processes are explicitly described) of the models employed. Each chapter focuses on a specific step in the hydrological modelling process.

Chapter 3 discusses the data requirements for different parameter identification methods in order to obtain stable parameters in the calibration procedure *(IV. Calibration)*. In Chapter 4, parameter transferability is investigated and used as indicator for how well the model represents the spatial and temporal variability *(IV. Calibration)*. In Chapter 5, the impact on model output of several modelling decisions during the procedural model and calibration steps is investigated *(III. Procedural model and IV. Calibration)*. Chapter 6 presents an uncertainty analysis, where the three sources of uncertainty are investigated for a large sample of basins *(VI. Uncertainty analysis)*. Chapter 7 is a plea for more thorough model validation, supported by a literature survey *(V. Evaluation)*. Finally, the considerations from this introduction and the five research chapters are combined in the synthesis of Chapter 8.

Chapter 2

Catchments, data, models, and methods

The fist part of this chapter discusses the catchments and the available data that have been used to conduct the studies presented in the following chapters of this thesis. The catchments cover a range of spatial scales as demonstrated in Figure 1.1. The second part of this chapter provides a description of employed hydrological models, which cover a range of complexities (Figure 1.1). The last part of this chapter describes employed sampling and sensitivity analysis methods.

2.1 Study areas and data

The catchments discussed in this section are ordered from small to large spatial extent. Per catchment, the available input and validation data will be discussed.
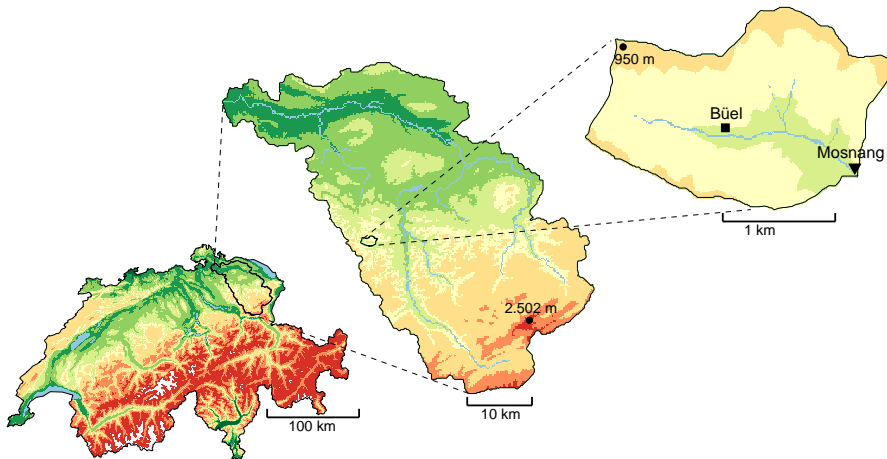

Figure 2.1: Elevation map of Switzerland, the Thur catchment, and the Rietholzbach catchment with the measurement stations indicated.
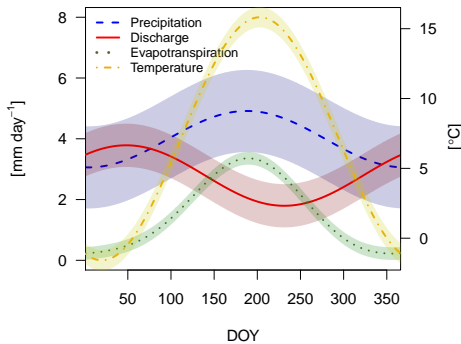
Figure 2.2: The 32 year smoothed average and standard deviation of hydro-meteorological characteristics in the Rietholzbach catchment.

Table 2.1: Climate summary for the Rietholzbach catchment based on 32 years of data (Seneviratne et al., 2012)

Climate Variable	Average
Temperature (2m)	7.1°C
Precipitation	1459 mm yr^{-1}
Discharge	1063 mm yr^{-1}
Actual ET	560 mm yr^{-1}

Rietholzbach catchment

The Rietholzbach catchment, subject of study in Chapter 3, is situated in North-East Switzerland, as shown in Figure 2.1. The catchment is small (3.3 km²), but very well-studied as research catchment of ETH Zürich since 1976 (*Seneviratne et al.*, 2012).

The altitude in the catchment ranges from 950 to 682 m; the average elevation is 795 m. The catchment has an average slope of 8.3° based on a 200×200 m DEM, but a slope of 14.5° according to *Seneviratne et al.* (2012). Land use consists mainly of pastures (71.9%) and forest (25.6%), while smaller portions of the area consist of orchards (1.2%) and settlements (1.3%). Due to the pronounced relief and variation in parent rock types (mainly '*nagelfluh*', limestones, dolomites and conglomerates), a large variety in soil types can be found in the area. The lower part of the catchment contains gleysols and gleyic cambisols (41.6%), while in the higher part of the catchment acidic and calcaric cambisols (40.7%) and regosols (17.6%) can be found. The soil depth is highly variable, and varies from 20 to 50 cm at the steep slopes, to several meters in the valley bottom.

Thirty-two years of evapotranspiration and most hydro-climatological measurements were available for the period 1976–2007 at an hourly time resolution. The hourly resolution is larger than the typical travel time in the stream within the catchment, but considerably smaller than the response time of the system (*Teuling et al.*, 2010). The main discharge gauge is '*Mosnang*' at the outlet of the catchment (see Figure 2.1). This gauge is operated by the Federal Office for the Environment (FOEN, Hydrology Division, Bern, Switzerland), but data were made available by ETH Zürich. Forcing data, like precipitation, temperature, and radiation, were measured at the '*Büel*' station in the centre of the catchment, which is maintained by the ETH Zürich. At the same location, actual evapotranspiration is measured with a weighing lysimeter.

Table 2.1 and Figure 2.2 contain a summary and the average yearly cycle of several meteorological characteristics. Note that the water balance based on measured precipitation, discharge and evapotranspiration does not close. The non-closure of 8% is within the range that can be caused by measurement errors. *Gurtz et al.* (2003) showed with a comparison between the rain gauge and the lysimeter that the largest discrepancies in precipitation measurements exist during winter (15-25% for rainfall, and 55-65% for snowfall), while in other periods both devices agree well. The maximum measurement error in local evapotranspiration by the weighing lysimeter is estimated to be 0.032 mm per measurement, based on the resolution of the electronic measurement device. Additional uncertainty is induced by the design of the lysimeter, e.g. due to a lack of connection to groundwater storage. Instrumental drift has been reported for the lysimeter, but the data has been corrected for drift with linear detrending. The lysimeter is covered with grass, which is the main land use in the Rietholzbach. Therefore it is assumed that the lysimeter is representative for the whole catchment. For the discharge measurements, based on a calibrated stage-discharge relation, errors of around 5 to 10% have been reported (*Seneviratne et al.*, 2012). Overall, measurement errors are relatively small and the data of the Rietholzbach is considered to be of high quality. However, the non-closure of the water balance allows for the possibility that some parts of the data are physically inconsistent: so-called *disinformation* (*Beven and Westerberg*, 2011).

Thur catchment

The Rietholzbach catchment is nested within the Thur catchment (1703 km^2), as shown in Figure 2.1. The Thur catchment, object of study in Chapter 4 and 5, was selected as study area because of the excellent data availability and because of its relevance as a tributary of the river Rhine (*Hurkmans et al.*, 2008). The main river in the catchment (the Thur) has a length of 127 km. The average elevation of the catchment is 765 m a.s.l., the mean slope is 7.9° (based on a 200 × 200 m resolution DEM and slope file). The catchment outlet is situated at Andelfingen at an elevation of 356 m a.s.l. (*Gurtz et al.*, 1999).

The catchment has an alpine/pre-alpine climatic regime, with high temperature variations both in space and time (Figure 2.3). Precipitation varies from 2500 mm yr^{-1} in the mountains to 1000 mm yr^{-1} in the lower areas. Part of the year the catchment is covered with snow. The most striking feature in the Thur catchment is the Säntis, an Alpine peak
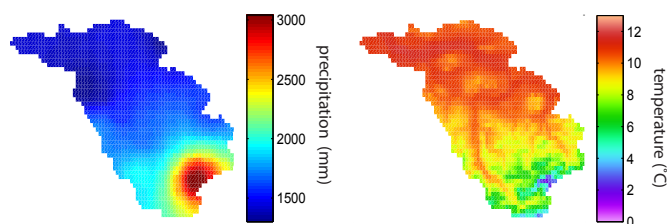

Figure 2.3: Distribution of main climate characteristics in the Thur over the period 1/8/2002 – 31/8/2003. Left: the precipitation sum. Right: the average temperature.

with an altitude of 2502 m. The dominant land use is pasture. The Thur has been subject of many previous studies, see for example *Gurtz et al.* (1999, 2003); *Jasper et al.* (2004); *Abbaspour et al.* (2007); *Yang et al.* (2007). Within the Thur catchment, measurements for nine (nested) sub-catchments are available, which are indicated in Figure 2.4. The smallest gauged sub-catchment is the Rietholzbach catchment (3.3 km², discussed in the previous section), the largest is Halden (1085 km²). Descriptors of the Thur catchment and the nine sub-catchments are summarized in Table 2.2.

For the station at the Thur outlet (Andelfingen) and eight sub-catchments hourly discharge measurements for the period 1974–2012 were made available by the Swiss Federal Office for the Environment (FOEN). All discharge measurements have been obtained using a stage-discharge relation, based on several measurements conducted by FOEN throughout the years, a.o. with an Acoustic Doppler Current Profiler (ADCP). The discharge measurements for the Rietholzbach catchment were made available by ETH Zürich.

Forcing data were made available by the Swiss Federal Office for Meteorology and Climatology (MeteoSwiss). These data have previously been used for numerous applications of hydrological models in the Thur (*Jasper et al.*, 2004; *Abbaspour et al.*, 2007; *Fundel and Zappa*, 2011; *Fundel et al.*, 2013; *Jörg-Hess et al.*, 2015). The data are available in the form required to implement the PREVAH model (*Viviroli et al.*, 2009a,b). Data from nine different meteorological stations in and around the catchment (Güttingen, Hörnli, Reckenholz, Säntis, St.Gallen, Tänikon, Wädenswil, Zürich and Rietholzbach) were available with an hourly time resolution and spatially interpolated with the use of the WINMET tool of the PREVAH modelling system (*Viviroli et al.*, 2009a), using elevation-dependent regression (EDR) and inverse distance weighting (IDW) and combinations of IDW and EDR. The data is available for the period 1981–2004, for which a stable configuration of stations is available. In this thesis, only data for the period May 2002 – August 2003 were
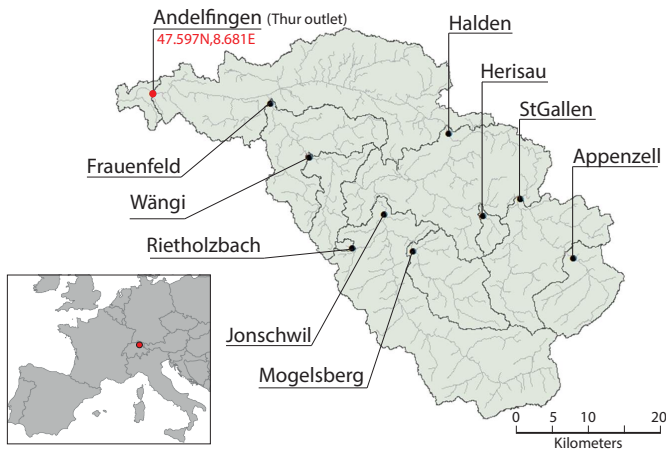

Figure 2.4: The Thur catchment and the nine sub-catchments for which data were available.

Table 2.2: Descriptors of the Thur catchment and the nine sub-catchments

Catchment	Area (km^2)	Mean elev. (m asl)	Mean slope (°)	Dominant land use
Rietholzbach	3.3	795	8.3	Pastures
Herisau	17.8	834	6.8	Pastures
Appenzell	74.2	1255	18.9	Sub-alpine meadow
Wängi	78.9	650	5.6	Pastures
Mogelsberg	88.2	938	11.1	Pastures
Frauenfeld	212	592	4.9	Pastures
St.Gallen	261	1040	12.5	Pastures
Jonschwil	493	1016	13.4	Pastures
Halden	1085	909	10.5	Pastures
Thur	1703	765	8.1	Pastures

used. This period is characterized by extreme high flows and an extreme drought event.

Land use, hydraulic conductivity, elevation, and soil water storage capacity maps, all with a spatial resolution of 200×200 m, were provided by the Swiss Federal Institute for Forest, Snow and Landscape Research (WSL) under license of Swisstopo (JA100118). Also in this case the pre-processing routines created to implement the PREVAH modelling system were used (*Viviroli et al.*, 2009a). Other soil characteristics, such as bulk density, have been obtained from the Harmonized World Soil Database (*FAO et al.*, 2012), which has a spatial resolution of 1×1 km.

605 catchments throughout the contiguous United States

Chapter 6 of this thesis uses a data set of 605 catchments throughout the contiguous United States (CONUS). This dataset is a sub-sample of the set which is presented and described by *Newman et al.* (2015). The catchments cover a wide range of climatic and hydrological conditions. Figure 2.5 illustrates the long-term average observed precipitation, temperature, and discharge for each catchment.

The median size of the catchments in the data set is 361 km^2, but areas range from 4 to 25,800 km^2 (see left panel of Figure 2.6). The mean elevation from the catchments ranges from 14 to 3640 m, with a median elevation of 454 m (middle panel of Figure 2.6). The steepest catchment has a mean slope of 14.3°, the most gentle catchment a slope of 0.05°, and the median slope of the catchments in the data set is 1.5° (right panel of Figure 2.6). Land use in the catchments varies from mixed forest to grassland and from savannah to crop land.

For all catchments, lumped forcing data were available based on the Daymet data set distributed by NASA. Daymet data has a temporal resolution of one day, and a spatial resolution of 1×1 km. The data is available from 1980 to present. The variables in Daymet
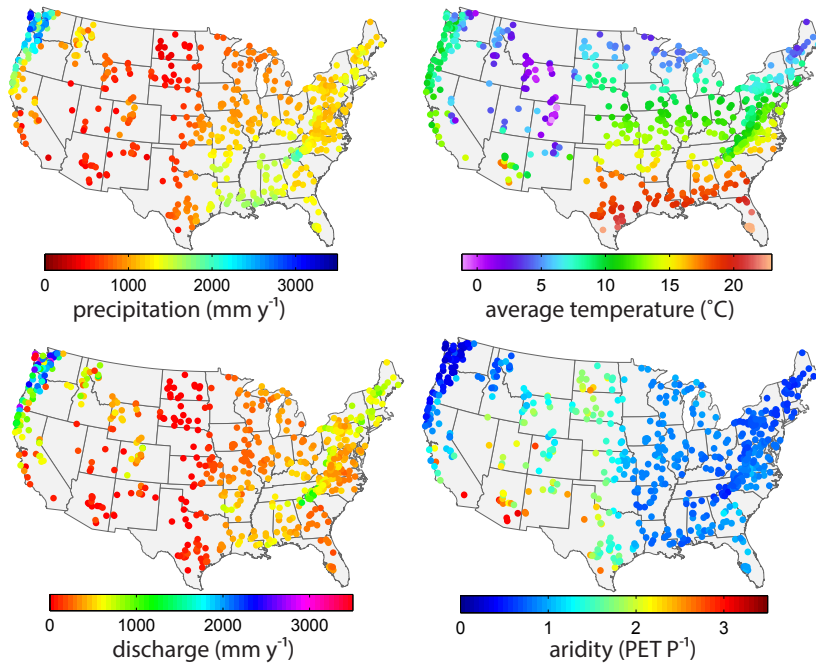
Figure 2.5: Distribution of main climate characteristics in the 605 basins throughout the contiguous United States based on observations over the period 1980–2008. Top left: yearly precipitation sum. Top right: the average temperature. Bottom left: yearly discharge sum. Bottom right: aridity defined as potential evapotranspiration (PET) divided by precipitation (P). PET was determined using the Priestley-Taylor method (Lhomme, 1997).

that were used to force hydrological models are the daily maximum and minimum temperature, precipitation, shortwave downward radiation, day length, and humidity. The Mountain Climate Simulator (MT-CLIM) was used to estimate shortwave radiation and humidity. If potential evapotranspiration (PET) was required as input variable for a hydrological model, PET was obtained using the Priestley-Taylor method. For all 605 catchments, United States Geological Survey (USGS) daily streamflow observations were available over the period 1980 to present.

2.2 Model descriptions

The perceptual, conceptual, and partly the procedural modelling step have been substituted by using readily available hydrological models. Because the goal of this thesis is to put hydrological modelling practice to the test, we made use of commonly used hydrological models. Throughout the research chapters of this thesis, four different rainfall-runoff models have been explored with various complexity: The simple dynamical systems approach (*Kirchner*, 2009), used in Chapter 3, is highly conceptual and clearly chooses a lumped approach for catchment functioning. The Sacramento Soil Moisture Account-

Figure 2.6: Cumulative distribution function (CDF) of catchment size (left), mean elevation (middle) and mean slope (right) for the dataset of 605 catchments.

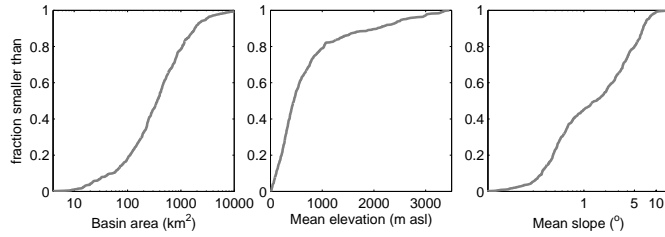ing model (SAC-SMA, *Burnash et al.*, 1973) and the Hydrologiska Byråns Vattenbalansavdelning model (HBV, *Bergström*, 1976), explored in Chapter 6, describe several processes explicitly in more detail, such as soil storage and surface runoff. The Variable Infiltration Capacity model (VIC, *Liang et al.*, 1994), used in Chapter 4–7, also explicitly describes processes such as root water uptake and transpiration coupled to stomatal resistance. The four hydrological models are described in more detail below, ordered from low complexity to high complexity, where we define complexity as the degree to which involved processes are explicitly described

Simple dynamical systems approach

In the 'Catchments as simple dynamical systems'-approach, the main assumption is that discharge solely depends on the amount of water stored in a catchment, and that the function describing this relation is unique. A motivation for this approach is that the physical processes taking place in catchments are highly complex and heterogeneous. In this approach, these processes are considered in a lumped fashion at the catchment scale (*Kirchner*, 2009). This approach is derived from the water balance:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = P - E - Q, \tag{2.1}$$

with $\mathrm{d}S$ being the change in storage in the catchment over time $\mathrm{d}t$, P is precipitation, E evaporation and Q discharge. It is assumed that there is a function f describing the relation between discharge and storage, and that this relation is invertible (no hysteresis):

$$Q = f(S), S = f^{-1}(Q). \tag{2.2}$$

The change in discharge over time can be related to the water balance:

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = \frac{\mathrm{d}Q}{\mathrm{d}S}\frac{\mathrm{d}S}{\mathrm{d}t} = \frac{\mathrm{d}Q}{\mathrm{d}S}(P - E - Q), \tag{2.3}$$

in which $\frac{\mathrm{d}Q}{\mathrm{d}S}$ is the sensitivity of discharge to changes in storage. This is the derivative of the function f describing the relation between storage and discharge. Generally, this derivative would be expressed in terms of storage, but catchment storage is not directly

measurable. Discharge is, and since it is assumed that the relation between storage and discharge is invertible, the derivative can be expressed as a function of discharge:

$$\frac{dQ}{dS} = f'(S) = f'(f^{-1}(Q)) = g(Q). \tag{2.4}$$

The function describing the derivative of discharge to storage is typically referred to as the *sensitivity function* $g(Q)$. While $g(Q)$ is not restricted to a particular form, the common power-law representation was chosen (*Brutsaert and Nieber*, 1977; *Kirchner*, 2009; *Troch et al.*, 2013) in Chapter 3. Applying a power-law results in a model with only two parameters:

$$g(Q) = \frac{dQ}{dS} = \alpha Q^{\beta}. \tag{2.5}$$

The change in discharge (dQ) due to a change in storage (dS) is described by the parameters α and β and the discharge Q. Note that the dimension of α depends on the value of β. There are fundamental differences in the behaviour of the system for $\beta < 1, \beta = 1$ and $\beta > 1$ (*Kirchner*, 2009). For example, $\beta < 1$ implies that there is a residual storage S_0 which remains in the catchment if discharge reduces to zero, for $\beta = 1$ there will be discharge at all storage values, which implies that storage can decline indefinitely, and for $\beta > 1$ S_0 is no longer the lower but the upper bound of storage in the catchment. For a comprehensive discussion on the effect of β, we refer to *Kirchner* (2009).

The 'Catchments as simple dynamical systems'-approach has been explored in the Rietholzbach catchment. The Rietholzbach experiences intermittent snow cover in winter, and therefore snowmelt had to be accounted for. Snowmelt M (mm h^{-1}) is assumed to be dependent on radiation and temperature, following the Restricted Degree-Day Radiation Balance approach by *Kustas et al.* (1994):

$$M = F_1(T_{2m} - T_0) + F_2 R_g. \tag{2.6}$$

F_1 (mm h^{-1} °C^{-1}) and F_2 (mm h^{-1}(W m^{-2})$^{-1}$) are parameters controlling the melt rate, T_{2m} (°C) is the measured temperature at 2 m height, T_0 (°C) is a threshold temperature, and R_g (W m^{-2}) is, in contrast to the net radiation which *Kustas et al.* (1994) proposed, the global radiation. *Teuling et al.* (2010) argued that for small catchments with partial snow cover, like the Rietholzbach, net radiation is highly variable and therefore global radiation was used, which is independent of local surface conditions.

When the snowmelt parameters, and the parameters α and β of the sensitivity function $g(Q)$ are identified, the discharge can be simulated. Due to the strong non-linearity of $g(Q)$, numerical stability improves if changes in log(Q) are simulated, rather than changes in Q (*Kirchner*, 2009). This leads to the following differential equation:

$$\frac{d(\log(Q))}{dt} = \frac{1}{Q}\frac{dQ}{dt} = g(Q) \cdot \left(\frac{P + M - ET}{Q} - 1\right), \tag{2.7}$$

where P is rainfall (mm h^{-1}), M snowmelt (mm h^{-1}) and ET evapotranspiration (mm h^{-1}). Equation 2.7 is solved with a fourth-order Runge-Kutta scheme with variable time step as proposed by *Kirchner* (2009) and adopted by *Teuling et al.* (2010) and *Brauer et al.* (2013).

The SAC-SMA model
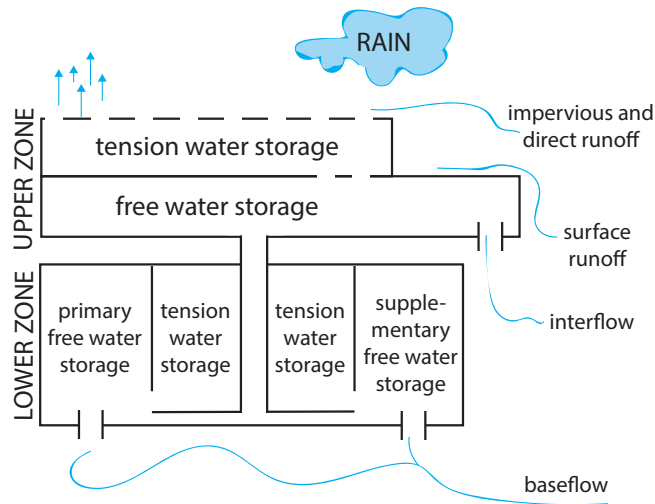
Figure 2.7: Overview of the processes incorporated in the Sacramento Soil Moisture Accounting model (SAC-SMA). Adapted from National Weather Service (2002).

The Sacramento Soil Moisture Accounting model (SAC-SMA, *Burnash et al.*, 1973) was developed by the US National Weather Service, with the goal to provide relatively short-term discharge predictions. The following description of the SAC-SMA model (from now on referred to simply as SAC) is based on the report of the *National Weather Service* (2002). The two basic components of SAC are tension water, water present in the soil but due to absorption to soil particles only removable through evaporation and transpiration, and free water, water that is available for percolation and drainage. Furthermore, SAC divides the soil into an upper and a lower zone. The tension water in the upper zone (represented in the UZTWM parameter) represents the amount of water which can be absorbed by the soil before drainage takes place. The free water in the upper zone (represented in the UZFWM parameter) either moves laterally through the soil as interflow, or drains vertically into the deeper soil. In the lower zone, the tension water (LZTWM) represents the amount of moisture necessary to satisfy the moisture requirements of the soil due to molecular attraction. The free water in the lower zone provides the reservoirs which generate base flow. Free water in the lower zone is divided into a primary type (LZFPM), which represents slow drainage, and a secondary type (LZFSM), which represents fast drainage after a recent rainfall event. The fast drainage can supplement the slow drainage. Direct runoff is generated from impervious areas (which can be parameterized with the PCTIM parameter) and when rainfall intensity exceeds the infiltration rate of the soil or when the soil is saturated. An overview of the processes described in the SAC model is given in Figure 2.7.

The HBV model
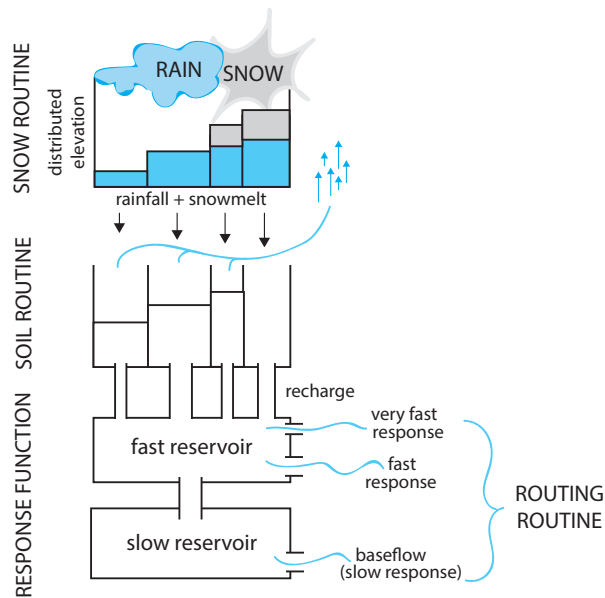
Figure 2.8: Overview of the processes incorporated in the HBV model. Adapted from Uhlenbrook et al. (1999).

The Hydrologiska Byråns Vattenbalansavdelning model (HBV, *Bergström*, 1976) was developed in Sweden with the aim to use the model for hydrological forecasting. Particular attention was given to snow processes due to the relevance of snow in Scandinavian areas. The following model description is mainly based on *Bergström* (1992). The HBV model consists of three main components: snow accumulation and melt (the snow routine), soil moisture (the soil routine), and response and river routing (response function and routing routine). The model is semi-distributed since different elevation and vegetation zones can be defined. Snow accumulation and snowmelt are parameterized with a degree-day expression, based on two parameters, one (Tm) which represents the threshold level above which snowmelt starts and one (Tr) below which precipitation falls as snow (sometimes these two parameters are combined in one parameter named TT). If rain falls on snow, runoff is generated only if the liquid holding capacity of the snow (usually fixed at 10%) is exceeded. The soil routine is controlled by three parameters; FC, BETA, and LP. FC represents the maximum soil moisture storage, BETA is the coefficient of a non-linear function that describes the relation between soil moisture level and recharge; a higher soil moisture level results in a higher recharge to the lower reservoirs. The LP parameter links the soil moisture level to evaporation; if the soil moisture level is lower than LP, actual evapotranspiration is reduced. Recharge from the soil moisture routine is transformed into discharge using two reservoirs: a fast responding reservoir, and

a slow responding reservoir that is fed with percolation water from the fast responding reservoir. The final outflow is smoothed using a triangular weighting function (MAXBAS) to account for river routing. An overview of the processes described in the HBV model is given in Figure 2.8.

The VIC model
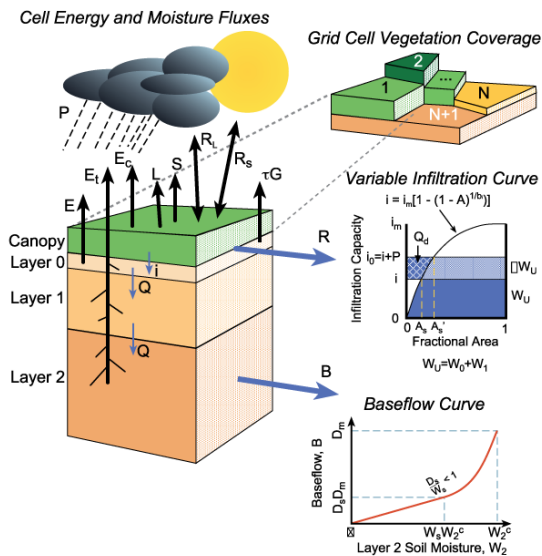
Figure 2.9: Overview of the processes incorporated in the Variable Infiltration Capacity (VIC) model. Source: http://vic.readthedocs.org.

The VIC model (*Liang et al.*, 1994, 1996) was initially developed for large scale applications, to couple climate models to hydrological processes. It is a land-surface model that solves both the water and the energy balance. Sub-grid land use variability is accounted for by providing vegetation tiles that each cover a certain percentage of the total surface area. Three different types of evaporation are considered by the VIC model: Evaporation from the bare soil (E_b), transpiration by the vegetation (T), considered per vegetation tile, and evaporation from interception (E_i). The total evapotranspiration is the area-weighted sum of the three evaporation types. The fraction of land that is not assigned to a particular land use type is considered to be bare soil. Evaporation from bare soil only occurs at the top layer (layer 0). If layer 0 is saturated, bare soil evaporation is at its potential rate. Potential evaporation is obtained with the Penman-Monteith equation. If the top layer is not saturated, an Arno-formulation (*Francini and Pacciani*, 1991), which uses the structure of the Xinanjiang model (*Zhao et al.*, 1980), is used to reduce the evaporation.

For the upper two soil layers, the Xinanjiang formulation (*Zhao et al.*, 1980) is used to describe infiltration. This formulation assumes that the infiltration capacity varies within

an area. Surface runoff occurs when precipitation intensity exceeds the local infiltration capacity of the soil. Moisture transport from layer 0 to layer 1 and from layer 1 to layer 2 is gravity-driven and only dictated by the moisture level of the upper layer. It is assumed that there is no diffusion between the different layers. Layer 2 characterizes the long term soil moisture response, e.g. seasonality. It only responds to short-term rainfall when both top layers are fully saturated. The gravity-driven moisture movement is regulated by the Brooks-Corey relationship:

$$Q_{i,i+1} = K_{\text{sat},i} \left(\frac{W_i - W_{r,i}}{W_i^c - W_{r,i}} \right)^{\text{expt}_i}. \tag{2.8}$$

$Q_{i,i+1}$ is the flow (L T^{-1}) from layer i to layer $i+1$. $K_{\text{sat},i}$ is the saturated hydraulic conductivity of layer i, W_i is the soil moisture content in layer i, W_i^c is the maximum soil moisture content in layer i, $W_{r,i}$ the residual moisture content in layer i. The exponent of the Brooks-Corey relation, expt$_i$, is defined as follows: $\frac{2}{B_p} + 3$, in which B_p is the pore size distribution index. The exponent as a whole is often calibrated.

Base flow is determined based on the moisture level of layer 2. Base flow generation follows the conceptualization of the Arno model (*Francini and Pacciani*, 1991). This formulation consists of a linear part (lower moisture content regions) and a quadratic part (in the higher moisture regions). Base flow is modelled as follows:

$$Q_b = \begin{cases} \frac{d_s d_m}{w_s W_2^c} \cdot W_2 & \text{if} \quad 0 \le W_2 < w_s W_2^c \\[2em] \frac{d_s d_m}{w_s W_2^c} \cdot W_2 + \left(d_m - \frac{d_s d_m}{w_s} \right) \left(\frac{W_2 - w_s W_2^c}{W_2^c - w_s W_2^c} \right)^g \\ \qquad \text{if} \quad W_2 \ge w_s W_2^c \end{cases} \tag{2.9}$$

In this equation, Q_b is the total base flow over the model time step, d_m is the maximum base flow, d_s the fraction of d_m where non-linear base flow begins, w_s is the fraction of soil moisture where non-linear base flow starts. W_2^c is the maximum soil moisture content in layer 2, calculated as a product of porosity and depth. The exponent g is by default set to two (*Liang et al.*, 1996).

Since the grid-size of the VIC model is often larger than the characteristic scale of snow processes, sub-grid variability is accounted for by means of elevation bands. For each grid cell the percentage of area within certain altitude ranges is provided. The snow model is applied for each elevation band and land use type separately; the weighted average provides the output per grid cell. This output consists of the Snow Water Equivalent (SWE) and the snow depth. The snow model is a two-layer accumulation-ablation model, which solves both the energy- and the mass balance. At the top layer of the snow cover the energy exchange takes place. A zero energy flux boundary is assumed at the snow-ground interface.

In order to apply VIC in a distributed fashion, a routing model is required to transport the water between the different grid cells. Therefore, the mizuRoute routine (*Mizukami*

et al., 2016) was implemented. The routing is based on the same concept as the default VIC-routing developed by *Lohmann et al.* (1996), except that in mizuRoute the response is determined per sub-catchment instead of per grid cell.

With the linearised St. Venant equation,

$$\frac{\partial Q}{\partial t} = D\frac{\partial^2 Q}{\partial x^2} - C\frac{\partial Q}{\partial x},$$

(2.10)

water is transported from the boundary of the sub-catchment to the next sub-catchment and finally to the outlet. In Equation 2.10, D (L^2T^{-1}) represents the diffusion coefficient and C ($L\,T^{-1}$) the advection coefficient.

In the default VIC routing of *Lohmann et al.* (1996), water is routed per grid cell and therefore dependent on the spatial resolution of the VIC model. By applying mizuRoute based on pre-defined sub-catchments (~1 km²), the effect of the spatial resolution on the routing process is excluded.

2.3 Parameter sampling strategy

For the VIC model, described in the previous section, a GLUE-based approach (*Beven and Binley*, 1992) was employed for parameter sampling. This approach requires that the model is run with a large number of parameter sets (preferably covering the complete parameter space). Subsequently only the behavioural runs are selected. Different definitions for 'behavioural' can be adopted, dependent on the goal of the study. In Chapter 4 for example, a different definition is used than in Chapter 6, because in the first one mainly a sensitivity analysis is conducted, while in the latter model confirmation is needed to estimate model credibility.

The VIC model has a large number of parameters, divided over three sections: soil parameters, vegetation parameters, and snow parameters. Sampling all parameters completely would be a heavy computational burden (see Figure 2.10a for a 3-parameter mo-
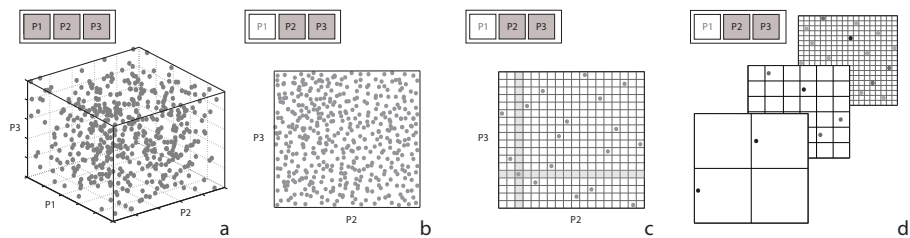
Figure 2.10: Parameter sampling strategy. (a) Example situation when sampling for a model with three parameters. (b) Sensitivity analysis can be conducted to decrease the dimensions of the sampling space. (c) Latin Hypercube sampling is structured and more efficient: one sample in each row and each column, as indicated with the bands. The number of samples has to be determined beforehand. (d) Hierarchical Latin Hypercube sampling allows to extend the sample if necessary, while conserving Latin Hypercube structure.

del example). Therefore, several strategies have been combined to make the sampling more efficient. The number of parameter samples needed to cover the full parameter space can decrease significantly by selecting only the most sensitive parameters (see Figure 2.10b), as described in the next section. The number of parameter sets can be further reduced by choosing an efficient sampling strategy (Figure 2.10c and d).

Sensitivity analysis

To determine which parameters have the largest influence on the modelled discharge, a sensitivity analysis was conducted on a broad selection of parameters in the model (see Table 2.3). The parameter selection was made such that the main hydrological processes were represented and included 28 VIC parameters from the three different sections. Sensitivity analysis was conducted using the Distributed Evaluation of Local Sensitivity Analysis (DELSA) method (*Rakovec et al.*, 2014). DELSA is a hybrid local-global sensitivity analysis method. It evaluates parameter sensitivity based on the gradients of the objective function for each individual parameter at several points throughout the parameter space. Note that this method only provides first-order sensitivities and thus does not account for parameter interaction.

A base set of 100 parameter samples was created. For each parameter k that is accounted for in the analysis, the base set of parameter samples is perturbed. In total, including the base set, this leads to (number of parameters+1) $\times$ 100 parameter samples that need to be evaluated. To save computation time, the sensitivity analysis was conducted on the lumped VIC model for the Thur. To study the effect of spatial scale on sensitivity, also two lumped models for sub-catchments of the Thur have been constructed: The Jonschwil catchment (495 km^2) and the Rietholzbach catchment (3.3 km^2). The Rietholzbach catchment is nested inside the Jonschwil catchment, which is again nested in the Thur catchment (Figure 2.4). The three catchments have comparable land use. Three objective func-
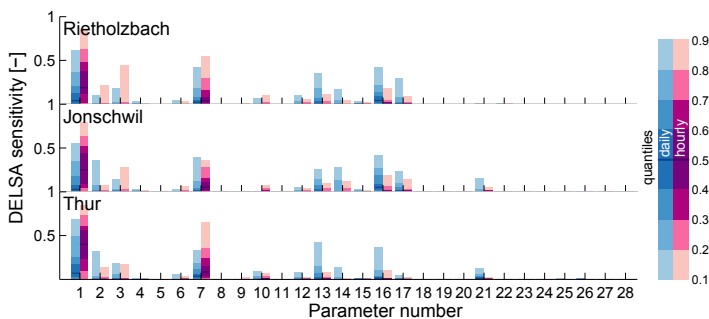

Figure 2.11: DELSA parameter sensitivity (scaled from 0 to 1) for three nested catchments with a different size (Rietholzbach; 3.3 km^2, Jonschwil; 493 km^2, Thur; 1703 km^2). The numbers on the x-axis refer to the parameters in Table 2.3. The sensitivity as shown in this figure is based on NSE(Q) as objective function. Results are shown based on a daily and hourly time interval.

Table 2.3: Description and bounds of VIC parameters that have been considered in the DELSA sensitivity analysis.

	Parameter	Units	Lower value	Upper value	Description
Soil parameter file					
1	b_i	-	10^{-5}	0.4	Variable infiltration curve parameter
2	d_s	-	10^{-4}	1	Fraction of d_m where non-linear base flow starts
3	d_m	$mm\,d^{-1}$	1	50	Maximum base flow
4	w_s	-	0.5	1	Fraction of maximum soil moisture where non-linear base flow starts
5	c	-	1	4	Exponent used in the base flow curve
6	$expt_0$	-	4	30	Exponent of the Brooks-Corey drainage equation layer 0
7	$expt_1$	-	4	30	Exponent of the Brooks-Corey drainage equation layer 1
8	$expt_2$	-	4	30	Exponent of the Brooks-Corey drainage equation layer 2
9	$Ksat_0$	$mm\,d^{-1}$	100	1000	Saturated hydrologic conductivity layer 0
10	$Ksat_1$	$mm\,d^{-1}$	100	1000	Saturated hydrologic conductivity layer 1
11	$Ksat_2$	$mm\,d^{-1}$	100	1000	Saturated hydrologic conductivity layer 2
12	$Depth_0$	m	0.01	0.5	Thickness of soil layer 0
13	$^\dagger Depth_1$	m	$Depth_0+0.1$	$Depth_0+4$	Thickness of soil layer 1
14	$Depth_2$	m	0.1	4	Thickness of soil layer 2
15	bulk density$_0$	$kg\,m^{-3}$	1500	2685	Bulk density of soil layer 0
16	bulk density$_1$	$kg\,m^{-3}$	1500	2685	Bulk density of soil layer 1
17	bulk density$_2$	$kg\,m^{-3}$	1500	2685	Bulk density of soil layer 2
18	Wcr-FRACT$_0$	-	0.30	0.47	Fractional soil moisture content at critical point layer 0
19	Wcr-FRACT$_1$	-	0.30	0.47	Fractional soil moisture content at critical point layer 1
20	Wcr-FRACT$_2$	-	0.30	0.47	Fractional soil moisture content at critical point layer 2
21	snow-rough	m	$5 \cdot 10^{-5}$	0.5	Surface roughness of the snow pack
Vegetation parameter file					
22	Root depth 0	m	0.1	3	Root zone thickness layer 0
23	Root depth 1	m	0.1	3	Root zone thickness layer 1
24	Root depth 2	m	0.1	3	Root zone thickness layer 2
Vegetation library file					
25	rmin	$s\,m^{-1}$	30	300	Minimum stomatal resistance of vegetation
26	*LAI	-	0.7	1.3	Leaf Area Index
Global parameter file					
27	T_{min}	$^\circ C$	-1.5	0.0	Minimum temperature at which rain can fall
28	$^\dagger T_{max}$	$^\circ C$	$T_{min}+0.5$	$T_{min}+1.5$	Maximum temperature at which snow can fall

† Value of this parameter must be greater than the related parameter mentioned in the parameter boundaries.
* Implemented as a multiplication factor to the default parameter values.

tions were used to assess the sensitivity of the parameters:

· The Kling-Gupta Efficiency (KGE) (*Gupta et al.*, 2009):

$$\text{KGE(Q)} = 1 - \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2}, \qquad (2.11)$$

where r is the correlation between observed discharge Q_o and modelled discharge Q_m, α is the standard deviation of Q_m divided by the standard deviation of Q_o, and β is the mean of Q_m ($\overline{Q_m}$) divided by the mean of Q_o ($\overline{Q_o}$) .

· The Nash-Sutcliffe Efficiency (NSE) of the discharge to describe the model performance for the higher discharge regions (*Nash and Sutcliffe*, 1970):

$$\text{NSE(Q)} = 1 - \frac{\sum_{t=1}^{T}(Q_o^t - Q_m^t)^2}{\sum_{t=1}^{T}(Q_o^t - \overline{Q}_o)^2} = 2 \cdot \alpha \cdot r - \alpha^2 - \beta_n^2, \qquad (2.12)$$

in which β_n is the bias normalized by the standard deviation.

· The Nash-Sutcliffe Efficiency of the logarithm of the discharge NSE(logQ) to test the model performance for low discharges (*Krause et al.*, 2005).

The analysis showed that parameter sensitivity did not notably change over the assessed scales: the same parameters were found to be most sensitive, but in a slightly different order (see Figure 2.11). There are four parameters which, for all scales and for all objective functions, proved to be highly sensitive: The parameter describing variable infiltration (b_i), the parameter that defines the fraction of d_m where non-linear base flow starts (d_s), the maximum base flow (d_m) and the exponent of the Brooks-Corey relation ($\frac{2}{B_p} + 3$, expt$_1$, see Equation 2.8). Other parameters that showed sensitivity in some cases were the depth and bulk density of soil layer 1, the depth and bulk density of soil layer 2, and the rooting depth of layer 0. The selection of sensitive parameters closely resembles the results of *Demaria et al.* (2007), who applied a sensitivity analysis on VIC over different hydro-climatological regimes. Next to the four identified most sensitive parameters, *Demaria et al.* (2007) found that the depth of soil layer 1 (Depth$_1$) was highly sensitive.

Parameter sampling

Based on the sensitivity analysis and the study of *Demaria et al.* (2007), five parameters in the VIC model were selected for parameter sampling (parameters 1-5 in Table 2.4). Furthermore, the two mizuRoute-parameters were added to the selection because they control the lateral exchange of water between grid cells (parameters 6-7 in Table 2.4). The selected parameters have been sampled using a Latin Hypercube Sample (LHS). This is a variance reduction method which efficiently samples the parameters within each region with equal probability in the parameter distribution (*Vořechovský and Novák*, 2009). To limit calculation time, the LHS should preferably be as small as possible, while still being able to provide insights in e.g. posterior parameter distributions. For a Monte Carlo (MC) sample, it is easy to start with a small sample, and add more samples if this shows to be necessary, e.g. based on the sample variance. For a variance reduction technique such as LHS this is not that straight forward. Therefore, we make use of the Hierarchical Latin Hypercube Sample (HLHS), recently developed by *Vořechovský* (2015). This method allows

us to start with a small LHS and add more samples if necessary, while conserving the LHS-structure. Inherent to this method is that every sample extension is twice as large as the previous sample, which results in a total number of simulations after r extensions:

$$N_{\text{sim},r} = 3^r \cdot N_{\text{start}}, \tag{2.13}$$

with N_{sim} being the total number of simulations, r the number of extensions, and N_{start} the initial number of samples. As a starting sample size 350 was chosen, which is sampled based on a space-filling criterion. For the seven parameters in the HLHS sample a uniform prior is assumed in order the study the full parameter space. The starting sample can be increased by a first extension to 1,050 samples in total, further to 3,150, and even up to 9,450. After each extension, the cumulative distribution function (CDF) of the objective functions (KGE, NSE) is compared with the CDF of the previous extension. A Kolmogorov-Smirnov test is used to test if the CDFs are significantly different. It was found that the CDF estimated from 3,150 samples was not significantly different from the CDF based on 1,050 samples at a 0.05-significance level. Therefore, 3,150 samples was considered sufficient to sample the parameter space. The seven sampled parameters (Table 2.4) have been applied uniformly over the study-catchments (Section 2.1), whereas the other soil- and land use parameters have been applied in a distributed fashion (separate value for each grid cell) based on data provided by the Swiss Federal Institute for Forest, Snow and Landscape Research (WSL) and the Harmonized World Soil Database (*FAO et al.,* 2012).

Table 2.4: Sampled VIC parameters, selected based on sensitivity analysis.

	Par.	Units	Lower value	Upper value	Description
1	b_i	-	10^{-5}	0.4	Variable infiltration curve parameter
2	d_s	-	10^{-4}	1.0	Fraction of d_m where non-linear base flow starts
3	d_m	mm d^{-1}	1.0	50	Maximum base flow
4	expt$_1$	-	4.0	18.0	Exponent of the Brooks-Corey drainage equation layer 1
5	Depth$_1$	m	Depth$_0$+0.1	Depth$_0$+3	Thickness of soil layer 1
6	C	m s^{-1}	0.5	4	Advection coefficient of horizontal routing
7	D	m^2 s^{-1}	200	4000	Diffusion coefficient of horizontal routing

Chapter 3

Parameter identification methods compared for a lumped model

In many rainfall-runoff models at least some calibration of model parameters has to take place. Especially for ungauged or poorly gauged basins this can be problematic, because there is little or no data available for calibration. A possible solution to overcome the problems caused by data scarcity is to set up a measurement campaign for a limited time period. In this study we determine the minimum amount of data, required to determine robust parameter values for a simple model with two parameters. The model is constructed such that the parameters can be determined not only with automatic calibration, but also by recession analysis and a priori from Boussinesq theory. The model has been applied to a research catchment in Switzerland. For automatic calibration and recession analysis one season (five months) is found to be sufficient to give robust parameters for simulation of high flows over the full observation period. For automatic calibration, this should be the season with the highest precipitation, for recession analysis the season with least evapotranspiration. The Boussinesq equation is able to give good parameter estimates for modelling high flows, but detailed in-situ knowledge of the catchment is required. Automatic calibration outperforms recession analysis and Boussinesq theory by far when it comes to parameter estimation with a focus on prediction of low flows.

3.1 Introduction

Modelling the discharge dynamics of catchments is an ever lasting challenge within the science of hydrology. Models are used to extrapolate the available measurements in space, e.g. to ungauged or poorly gauged basins, and time, e.g. to assess future hydrological change (*Beven*, 2012). Because the parameters in hydrological models often represent a different spatial scale than the observation scale, or because conceptual parameters have no directly measurable physical meaning, calibration of hydrological models is almost always inevitable (*Beven*, 2012; *Koskela et al.*, 2012).

Efforts to predict or simulate discharge in a catchment are often hampered by data limitations, especially in ungauged basins (*Kileshye Onema et al.*, 2012). An ungauged basin is a basin with inadequate records of hydrological observations, both in terms of data quality and quantity, for a certain variable of interest (*Sivapalan et al.*, 2003). The Prediction in Ungauged Basins (PUB) initiative was one of the main themes in hydrology during the last decade (*Blöschl et al.*, 2013), although numerous challenges remain. Since estimated model parameters often depend on factors such as the employed time step or the model structure (see for example Chapter 7 and 6), options for regionalization of model parameters to ungauged basins are limited (*Kavetski et al.*, 2011; *Wagener and Wheater*, 2006). *Seibert and Beven* (2009) and *Tada and Beven* (2012) state that field measurements of discharge during a short observation period might be a cost-effective strategy to improve the predictability in ungauged basins. Many studies have been conducted to identify the minimal length of this 'short observation period' in order to obtain reasonable model results, e.g. *Sorooshian* (1983), *Yapo et al.* (1996), *Gan et al.* (1997), *Brath et al.* (2004), *McIntyre and Wheater* (2004), *Perrin et al.* (2007), *Seibert and Beven* (2009), and *Tada and Beven* (2012). The problem of ungauged basins (hence their name) is their lack of validation data (*Sivapalan et al.*, 2003). Therefore these studies, without exception, used heavily gauged basins to test their theories. This is the so-called PUB paradox (*Bonell et al.*, 2006): using well-gauged catchments for testing methods designed for data scarce catchments.

Yapo et al. (1996) found that eight years of daily data are necessary to calibrate parameters which are relatively insensitive to the calibration period. *Brath et al.* (2004) analysed the influence of data availability on model calibration for a spatially-distributed hydrological model and came up with three months as minimum data length before significant deterioration of model performance occurred. *Perrin et al.* (2007) used discontinuous data for calibration, and found that 365 calibration days over a longer period, including wet and dry periods, is sufficient for robust parameter estimation. *Seibert and Beven* (2009) reported 32 daily observations as a plateau, after which additional observations did no longer significantly improve the model performance. *Tada and Beven* (2012) found for two out of three catchments a significant decrease in model performance when data availability was halved from 256 days to 128 days, which implied that there was a characteristic length or time scale in the catchment that should be included in the information. The required data for robust model parameters varied widely between all the studies, from three months to eight years. Climate could play a role, but also catchment and model type are important (*Gan et al.*, 1997). *Yapo et al.* (1996) for example, used

the NWSRFS-SMA model with thirteen optimized parameters, whereas *Tada and Beven* (2012) used TOPMODEL (*Beven and Kirkby*, 1979) with eight optimized parameters. *Perrin et al.* (2007) used a model with four and a model with six parameters, and showed that the more parsimonious model required fewer calibration data to obtain stable and robust parameters.

An 'extreme' example of a parsimonious model, which allows for a minimal number of parameters, is the simple dynamical systems approach as described by *Kirchner* (2009), which is based on a similar framework as the Input-Storage-Output (ISO) function model described by *Lambert* (1972). The fundamental assumption of the approach of *Kirchner* (2009) is that the discharge of the stream depends solely on the amount of water stored in the catchment. This is an important conceptual simplification. *Seibert et al.* (2003), for example, showed that this might only be true for the riparian zone in a hillslope catchment. *Uchida et al.* (2004) found a strong relation between hillslope discharge and upslope subsurface saturated area, suggesting that a unique relation between storage and discharge might be a valid assumption. If a function describing the relation between discharge and storage can be found, the discharge can be modelled using only precipitation and evaporation data.

Each of the studies that investigated the minimal data requirements for parameter identification used automatic calibration for optimization of the parameters. However, in this study, a parsimonious simple dynamical systems model (*Kirchner*, 2009) was constructed such that the parameters could be identified not only with automatic calibration, but also with recession analysis (*Brutsaert and Nieber*, 1977), which is based on the recession periods in the discharge, and a priori based on the physical approach of Boussinesq (*Rupp and Selker*, 2006), a method to identify parameters based on catchment characteristics. Automatic calibration and recession analysis are applied to determine the length of the data series necessary to obtain robust parameters for a small mountainous catchment, and which time of the year suits best if a short measurement campaign would be carried out in a data scarce region. The a-priori estimate of the subsurface parameters with Boussinesq theory provides a null hypothesis against which the other results could be evaluated. The two main objectives in this chapter are:

1. To determine which method (automatic calibration, recession analysis, Boussinesq theory) provides the parameters of the simple dynamical systems model that yield the best discharge simulations for the catchment.

2. To examine which length and timing of discharge data are necessary to obtain reliable parameters and well-estimated discharges, with emphasis on both high and low flow conditions.

This study is conducted in the Rietholzbach basin, a very well-gauged basin and therefore, ironically, in accordance with the PUB-paradox. Section 2.1 describes the catchment and the available data.
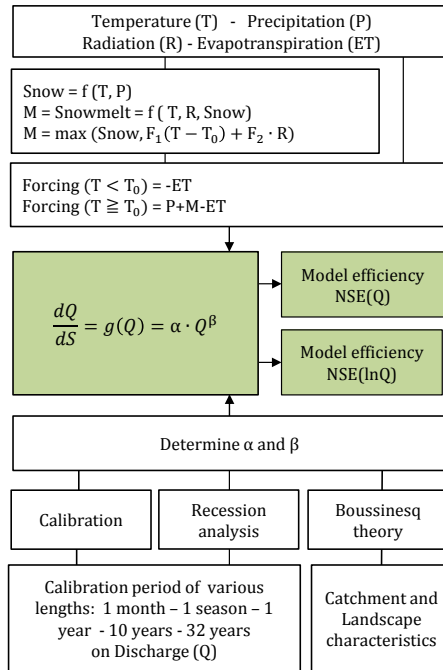
Figure 3.1: Overview of the model built for this study.

3.2 Methods

We define a parsimonious model where the Rietholzbach catchment is considered as a simple dynamical system, following the approach of *Kirchner* (2009). A fundamental underlying assumption of this approach is that the discharge depends solely on the volume of water in storage, and that this relation is invertible. The next section describes the identification procedure for the snow parameters, followed by a description of the three different identification methods for the parameters of the sensitivity function: automatic calibration, recession analysis, and Boussinesq theory.

Model

The model employed in this study is comprehensively described in Section 2.2. In this study the common power-law representation as sensitivity function $g(Q)$ was chosen (*Brutsaert and Nieber*, 1977; *Kirchner*, 2009; *Troch et al.*, 2013), see Equation 2.5 on p.20. While more complex representations of $g(Q)$ might lead to better results, the power-law function has the advantage that theoretical approaches exist which link the power-law coefficients to landscape and soil physical properties, e.g. the Boussinesq equation (*Boussinesq*, 1877; *Rupp and Selker*, 2006), and the simple model set-up allows for extensive optimization over time periods up to 32 years. The power-law function also allows recession analysis as a method to estimate the parameters.

Because the Rietholzbach catchment has intermittent snow cover in winter, snowmelt was accounted for by using a Degree-Day radiation balance approach (*Kustas et al.*, 1994), see Eq. 2.6 on p. 20. For two typical snow months with a clear snowmelt signature visible in the runoff (November 1992 and February 2002) the three snow parameters F_1, F_2 and T_0 were optimized with the DREAM-algorithm (Section 3.2), with the Sum of Squared Errors (SSE)-based likelihood function as implemented in DREAM (*Vrugt et al.*, 2008). The obtained values for the snow parameters in the two calibrations were nearly equal and therefore the averages (F_1=0.18, F_2=0.00050 and T_0=0.99) were used as fixed snow parameters for the remainder of this study. F_2 was found to be the lower boundary value, indicating that snow melt is temperature driven rather than radiation driven in the Rietholzbach catchment. This is confirmed in the study of *Orth and Seneviratne* (2013), who showed that a solely temperature-based model performed well in this catchment.

Model efficiency is expressed in terms of the Nash-Sutcliffe Efficiency NSE (*Nash and Sutcliffe*, 1970), Eq. 2.12. We use the NSE(Q) computed on the basis of observed discharges to put emphasis on prediction of peak flows (due to the large sensitivity of the NSE to high flows *Krause et al.*, 2005), whereas the NSE computed using the logarithmically transformed discharges, NSE(logQ), was used to evaluate model prediction of low flows (*Krause et al.*, 2005). In the remainder of this chapter, statements concerning model performance for low flows refer to NSE(logQ), while statements concerning the model performance for high flows refer to NSE(Q).

An overview of the simple low-dimensional rainfall-runoff model set-up used in this study can be found in Figure 3.1. For automatic calibration and recession analysis, the model has been fed with hourly hydro-climatological data for a monthly time period, a seasonal time period (with summer defined from May until September, and winter from November until March), a yearly time period, a 10-yearly time period (1976–1985, 1986–1995, 1996–2005) and for the full period of 32 years (1976–2007) for which data were available.

Calibration using Bayesian-based DREAM

To determine the optimal parameters of the model with automatic calibration, the Differential Evolution Adaptive Metropolis (DREAM) algorithm (*Vrugt et al.*, 2008) was used. The DREAM algorithm uses a Markov Chain Monte Carlo (MCMC) sampler to explore the posterior distribution of the parameters within a Bayesian framework. The method starts with an initial population of points which strategically covers the response surface, which was constrained such that β lies between 0 and 3 and α between 0 and 2, based on values typically found in natural catchments (*Kirchner*, 2009). The use of more chains with different starting points enables to deal with multiple regions with large attraction. From this initial population, candidate points are generated. Candidate points are accepted or rejected using the Metropolis acceptance probability (*Vrugt et al.*, 2008). In the end, several chains move over the response surface, searching for the global minimum. Ten Markov chains were used. The prior distribution of the parameters, α and β, was assumed to be uniform. The first discharge measurement of the calibration period was used as initial condition. With an SSE-based likelihood function as implemented in DREAM (*Vrugt*

et al., 2008) the optimal parameter set was determined as the set with the lowest SSE. Calibration has been applied directly to the observed data (minimizing the SSE for the discharge Q) and to the log-transformed data (minimizing the SSE for the logarithm of the discharge, logQ).

The DREAM algorithm is especially designed for complex high-dimensional non-linear problems, while the model used in this study is certainly not high-dimensional. At first sight, the DREAM algorithm therefore seems too sophisticated for the problem at hand. However, this algorithm was used because it also provides insight into the uncertainty of the parameters. To show that the response surface of this model is not complex and to show that the parameters can be identified by a simpler algorithm, the local Levenberg-Marquardt (LM) optimization algorithm (*Reed and Marks II*, 1999) was used in addition to DREAM. The classical LM algorithm is a compromise between Newton's method and gradient descent. Although the classical Levenberg-Marquardt algorithm has no parameter constraints, constraints have been implemented here by means of a penalty value if β is lower than 0 or higher than 3, or if α is lower than 0 or higher than 2. First guesses for each parameter were kept constant for each calibration at α=0.01, β=0.8 and Q_0=first discharge measurement. These numbers are based on *Shaw and Riha* (2012) and have been determined without prior knowledge of the parameter values for the catchment used in this study. The sum of squared errors (SSE) was used as objective function, both for calibration directly to the observed data and to the log-transformed data. Finally, optimization with DREAM and the LM algorithm has been conducted for several calibration periods; a monthly time period, a seasonal time period, a yearly time period, a 10-yearly time period and for the full data period (32 years).

Recession analysis

The method to determine α and β, and thus finally the sensitivity function $g(Q)$, from recession analysis (*Brutsaert and Nieber*, 1977) was proposed by *Kirchner* (2009) and is based on discharge observations. Recession analysis can be an interesting alternative to automatic calibration, for example in regions where information on occurrence of precipitation is more accurate than the amounts, which is relevant for poorly gauged basins. Figure 3.2 shows the recession analysis procedure which has been employed with different calibration periods.

During recession, i.e. excluding rainfall (P) and snowmelt (M), and during night time (which makes it possible to neglect evapotranspiration), the sensitivity function can be expressed as (see Eq. 2.7):

$$g(Q) = \frac{\mathrm{d}Q}{\mathrm{d}S} = \alpha Q^\beta \approx -\frac{1}{Q}\frac{\mathrm{d}Q}{\mathrm{d}t}|_{P,M,ET<<Q}. \tag{3.1}$$

In order to use Equation 3.1 to determine the sensitivity function parameters, the observed data has to be selected carefully. In this study, two different criteria for data selection have been used:

1. Dry and dark periods, too warm for snow cover to be present.
2. Dry and dark periods, too cold for snow melt to occur.

For an exact description of the criteria per set, see *Teuling et al.* (2010). The strict criteria for data selection only leave part of the data available for recession analysis. From the selected data (shown in the left panel of Figure 3.2), $g(Q)$ is determined (middle panel of Figure 3.2) with Equation 3.1. Because of measurement errors and a possible difference between catchment representation and reality, there is some scatter in the selected data. Therefore, the data has been binned according to the method used by *Teuling et al.* (2010).

If 30 or less data points were selected for the two sets together, α and β were fitted directly on all data points larger than zero (negative data points could exist due to small increases in discharge, e.g. due to the diurnal cycle in evapotranspiration or due to measurement errors), with a minimum of three data points. If more than 30 data points were selected, but only two bins or less, α and β were also determined on all individual data points above zero. If more than 30 data points and more than two bins were available, α and β were fitted on the binned data. These criteria are especially important during recession analysis for a monthly time period, because the number of selected data points per month can be limited. The α (intercept) and β (slope) were determined by plotting the logarithm of $g(Q)$ versus the logarithm of Q and fitting a straight line $\log(\alpha) + \beta \cdot \log(Q)$ (right panel in Figure 3.2), which is the classical approach of recession analysis as described by *Brutsaert and Nieber* (1977) and referred to as 'logRA' in the sequel of this Chapter. In addition, α and β were determined on linear axes for both $g(Q)$ and Q, fitting a power law. This method is further referred to as 'linRA'.

Boussinesq equation for sloping aquifers

Boussinesq (1877) derived an equation describing groundwater flow in an unconfined aquifer overlying an impermeable layer. An important assumption of this approach is the Dupuit-Forchheimer approximation, assuming that groundwater moves parallel to the impermeable layer in an unconfined aquifer and that groundwater discharge is proportional to the thickness of the saturated aquifer (*Troch et al.*, 2013). With this assumption,
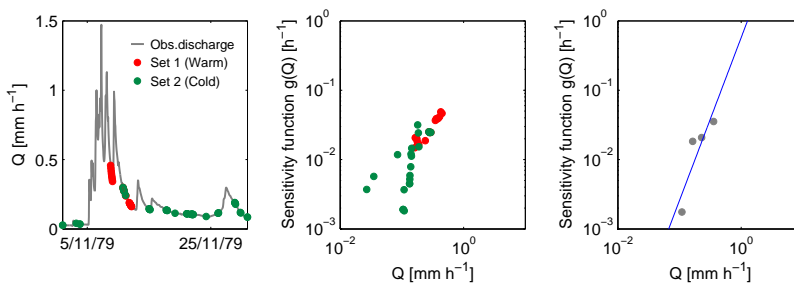
Figure 3.2: The process of recession analysis for a randomly selected month. Left panel: data are selected according to different criteria. Middle panel: sensitivity function $g(Q)$ is determined for each of the selected data points, based on Equation 3.1. Right panel: the points are binned, and a line is fitted through the data, α is represented by the intercept of the line, β by the slope of the line.

Boussinesq (1877) adapted Darcy's law and combined it with the continuity equation. An important property of the Boussinesq equation (or its linearisations) is that analytical solutions of the equation can be expressed in the form of Equation 2.5 (*Rupp and Selker*, 2006), which gives the opportunity to estimate the two simple dynamical system parameters in this study with the use of Boussinesq theory. The advantage of using a physically-based method such as the Boussinesq equation is that it does not need time consuming and costly discharge measurements. Disadvantage is that effective values for subsurface parameters at the catchment scale have to be estimated. It represents the whole catchment as a single hillslope draining to a channel, as shown in Figure 3.3, not taking into account heterogeneity in the topography or subsurface of the catchment. It is therefore a strong simplification of reality, especially in a catchment like the Rietholzbach with complex topography.

Nowadays, many different forms of the Boussinesq equation are in use, see for example the overview given in *Rupp and Selker* (2006). For the Boussinesq equation adapted for sloping aquifers, relevant in this study, *Rupp and Selker* (2006) used numerical simulations to empirically derive an analytical solution to the non-linear Boussinesq equation for the late part of the recession. Assumptions of the applied Boussinesq equation are that no water flows through the bottom and water divide boundaries (see Figure 3.3), and that the water level in the stream is constant. *Rupp and Selker* (2006) present the following equations for α and β in terms of physical parameters:

$$\alpha = \frac{(n+1)^2}{(n+0.01)\varphi A} \cdot \left[\frac{2k_D L \sin\phi}{(n+1)D^n}\right]^{\frac{1}{n+1}} \cdot (10^{-3} A)^{\beta}, \qquad (3.2)$$

$$\beta = \frac{n}{n+1}. \qquad (3.3)$$

Figure 3.3: Water table profile for the applied Boussinesq model, which is adapted for sloping aquifers. Zero flux boundaries at the right hand side and at the bottom are assumed. Parameters D, ϕ, and L have to be estimated from a Digital Elevation Model (DEM).

Table 3.1: Parameter values for the Boussinesq equation

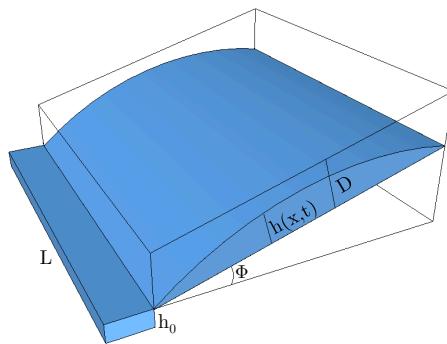|	Unit	Value	Description
k	$m\,h^{-1}$	range:	Saturated hydraulic conductivity
		$6.5{\cdot}10^{-5}$	Rietholzbach average (*König and Lang*, 1994)
		0.1170	Nagelfluh (*König and Lang*, 1994)
		0.0360	Lowest estimate (*Lehner and Seneviratne*, 2010)
		1.8	Highest estimate (*Lehner and Seneviratne*, 2010)
L	m	2500	Channel length
φ	-	0.1;0.2;0.25;0.3;0.35	Drainable porosity
A	m^2	$3.31 \cdot 10^6$	Catchment area
ϕ	°	14.5	Average slope
D	m	1	Average depth of the soil
n	-	0;1;10;20;30;40;50;60;70	Saturated hydraulic conductivity profile parameter

With φ (-) the drainable porosity, A (m²) the catchment area, k_D (m h⁻¹) the saturated hydraulic conductivity at depth D, L (m) the length of the aquifer, ϕ (-) the slope of the aquifer, and D (m) the depth of the aquifer. The n-parameter determines the saturated hydraulic conductivity profile, with the following equation (*Rupp and Selker*, 2006):

$$k(z) = k_D(z/D)^n.\tag{3.4}$$

For n=0, no variation with depth is assumed, n=1 indicates a linear profile with saturated hydraulic conductivity decreasing with depth, $n > 1$ leads to power law profiles. Note that for n=0, β=0 (thus leading to a linear reservoir). Experience from earlier studies (e.g. *Kirchner*, 2009; *Teuling et al.*, 2010; *Brauer et al.*, 2013) showed that β is hardly ever equal to zero, so a range of possible values for n was used. The k_D in Equation 3.2 and 3.4 differs from the average $\bar{k}$ in the area (see e.g. *König and Lang*, 1994). Integration of Equation 3.4 leads to the following relationship:

$$k_D = (n+1) \cdot \bar{k}.\tag{3.5}$$

The parameter values that have been used for Equation (3.2) and (3.3) are given in Table 3.1. For the saturated hydraulic conductivity $\bar{k}$, the drainable porosity φ, and the saturated hydraulic conductivity profile n, exact values were unknown, hence a range of values was used (Table 3.1). This lead to 180 different parameter sets. For each set the model efficiencies, NSE(Q) and NSE(logQ), over the full validation period were determined.

3.3 Results and discussion

First the results from the automatic calibration procedure are discussed, followed by the results obtained with recession analysis and Boussinesq theory. Finally, a comparison between the three methods is made. The discussed model efficiencies are based on a common validation period covering the full data length, which was 32 years, so that all reported NSE values refer to the same 32 year period and can be compared directly. No

split sample validation was employed. Hence, there is at least one month overlap between the calibration and the validation period. The overlap between calibration and validation period increases further with longer calibration periods: for the calibration period of 32 years, the calibration and validation period are identical. Since the analysis focusses on much shorter periods than the full validation period, the effects of this approach on the results are believed to be small.

Calibration with DREAM

Both with the DREAM algorithm (*Vrugt et al.*, 2008) with an SSE-based likelihood function, and with the Levenberg-Marquardt (LM) algorithm with the SSE as objective function, α and β were calibrated. The model was calibrated directly on the data, maximizing the NSE(Q), and on a log-transformation of the data, maximizing NSE(logQ). As explained in the Methods section, DREAM is a heavy algorithm for such a simple problem and this is recognized in the runtime of the process. For optimization with one month of calibration data, the fastest convergence was reached after 152 seconds, while the longest optimization took over 30 hours (median: 29 minutes). On the contrary, with the LM algorithm the fastest optimization on one month of data took less than a second, while the slowest optimization took 123 seconds. The number of forward runs is also much lower for the latter; on average 15.6 evaluations for the monthly calibration, while DREAM uses 4000 function evaluations. Calibration results for both algorithms are comparable. Rounded to one decimal, both algorithms found equal values in 84% of the cases based on monthly data (log monthly: 86%), increasing to 97% (log: 97%) and 100% (log: 100%)

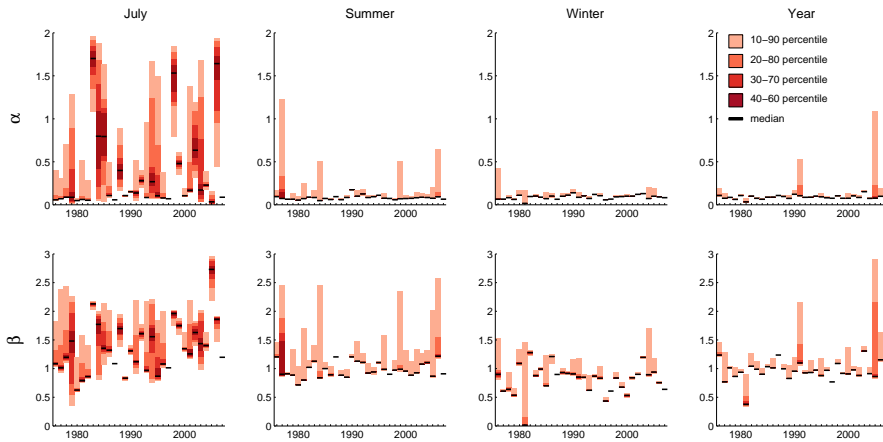*Figure 3.4: Box-plots showing the distribution of the parameter sample obtained with DREAM, with each panel showing a different calibration period: one month (only July is shown), a summer season, a winter season, and a year. Each box-plot represents one optimization, with 32 optimizations in total per calibration period, corresponding to 32 years of available data (for the period 1976-2007).*

for year and 10 year calibration periods respectively. It is thus shown that optimization of the parameters in the simple model can be achieved with a simple local optimization algorithm. The discussion of the results from now on will only consider the results obtained with DREAM.

With the Gelman and Rubin criterion $\hat{R}$ (*Gelman and Rubin*, 1992), implemented in DREAM, it was shown that for some calibration months the Markov Chains were not yet completely converged (i.e., the $\hat{R}$ was larger than 1.2). These were also the months with the longest run time for the DREAM algorithm. For calibration directly on the data, in 15% of the calibration-months the chains had not completely converged for either α, β or both. For calibration on the log-transformed data this was the case for 8% of the months. Nonconvergence was not restricted to one particular month or period but occurred throughout the year. Examination of the paths of the Markov Chains learned that a strong correlation between α and β was found for these particular months, which makes exact identification of the parameters impossible. For calibration periods longer than one month, the chains always converged for both parameters.

The DREAM algorithm is a MCMC sampler, and therefore provides insight in how well the parameters are defined. Figure 3.4 shows the distribution of the complete posterior for each optimization that was conducted on the discharge, for different calibration periods. The figure shows that both α and β are not clearly identifiable with a calibration period of one month (an example for July is shown). If the calibration period is increased to one season or one year, the parameter identifiability already strongly improved. Model efficiencies for the *calibration* period were in general high (not shown). The median of the NSE(Q) for different calibration periods varied from 0.71 to 0.75, with maxima reaching up to a NSE(Q) of 0.99 for calibration on one month of data. The model efficiency for the log-transformed data during the calibration period varied between a NSE(logQ) of 0.76 and 0.80 for different calibration periods, with again a maximum of 0.99 for calibration on one month of data. The model efficiencies obtained for the calibration period shows that the model can perform well in this catchment, but validation of the model is important to determine the applicability for poorly gauged basins with a short observation record. The results of the optimization for the 32-year-*validation* period are presented in Figure 3.5, the upper row for calibration directly on the data, and the third row for calibration on the log-transformed data.

Calibration on the original data
The optimal parameter values obtained based on 32 years of data were $\alpha = 0.083$ and β=0.943 (NSE(Q)=0.71, NSE(logQ)=0.18). Figure 3.5 (upper row) shows that reasonable model results (NSE(Q) >0.6) can be obtained as soon as one season of data or longer is available for calibration. The median model efficiency obtained with one season of data is NSE(Q)=0.67. This increased only slightly (NSE(Q)=0.69) when one year of data was used, and was the highest for 32 years of data (NSE(Q)=0.71). Figure 3.6 shows the distribution of model efficiencies for the summer and winter season separately. The figure shows that the summer season (median NSE(Q)=0.69) as calibration period leads to better results than the winter season (median NSE(Q)=0.62). The parameters obtained
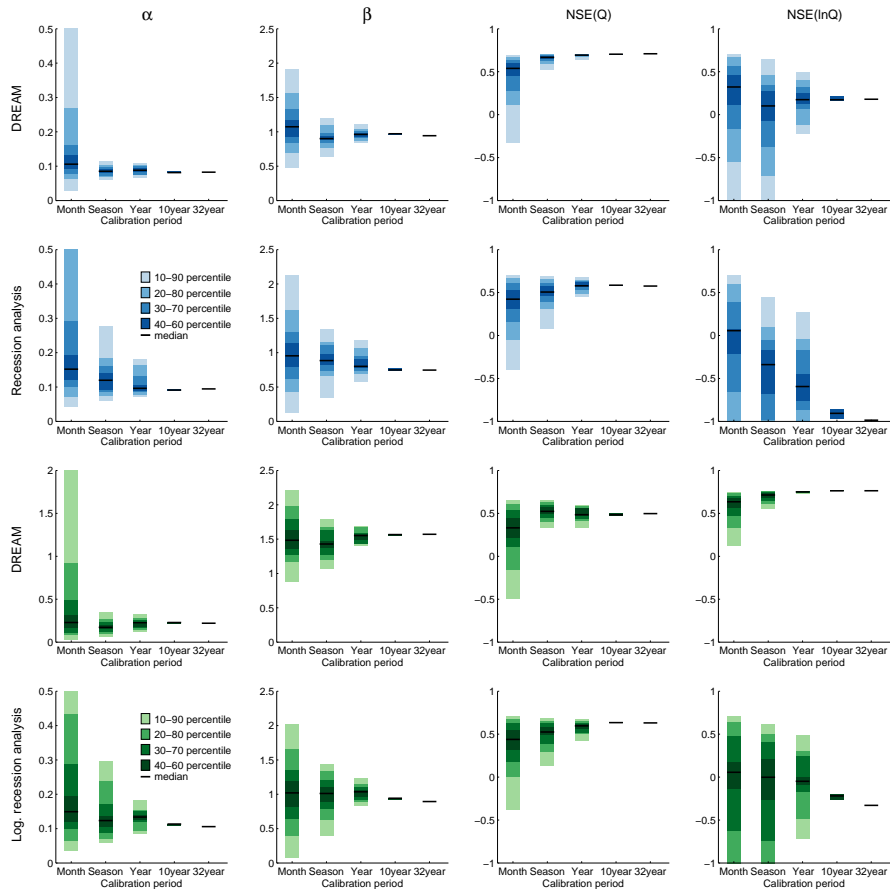
Figure 3.5: Box-plots showing the distribution of the optimal parameters (α, β) and the model efficiencies (NSE(Q), NSE(logQ)) for the validation period of 32 years for different calibration periods and different optimization procedures (dependent on the calibration period; calibration on one month of data is conducted 384 times, while calibration on one year of data was conducted 32 times). The blue box-plots (upper two lines) show results for calibration directly on the data (automatic calibration) or fitting on linear axes (recession analysis), the green box-plots (lower two lines) show results for calibration on the log-transformed data (automatic calibration) or fitting on logarithmic axes (recession analysis).
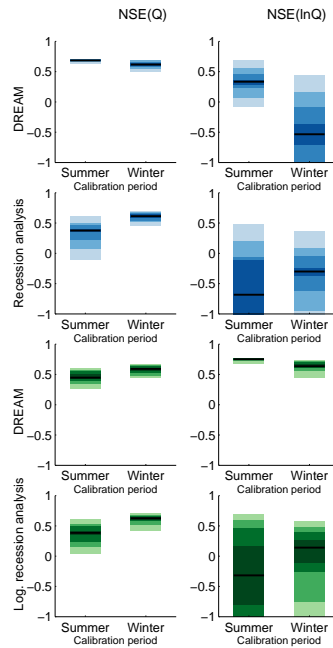
Figure 3.6: Box-plots showing the distribution of NSE(Q) and NSE(logQ) obtained with models that were calibrated on the summer (May until September) and the winter (November until March) season. The blue box-plots (upper two lines) show the results for automatic calibration on the original data and linear recession analysis, the green box-plots (lower two lines) show the results for automatic calibration on log-transformed data and logarithmic recession analysis. Legend according to Figure 3.5.

based on one summer of data perform equally well as parameters that were obtained from calibration on one year of data. Although the parameters obtained with automatic calibration directly on the data seem to be able to reasonably capture high flows in the model, they are clearly not able to describe low flows, given the low NSE(logQ).

Calibration on the log-transformed data
Calibration on the log-transformed data leads to α=0.220 and β=1.571 based on 32 years of data (NSE(Q)=0.50, NSE(logQ)=0.76). Optimization on the log-transformed data improved the capability of the model to predict low flows, but at the same time its capability of describing high flows decreased. The third row of Figure 3.5 shows that again, one season of data seems sufficient for reasonable model results (NSE(logQ)=0.71), but the model efficiency improved slightly when one year of data were used (NSE(logQ)=0.75). Figure 3.6 shows that for calibration on log-transformed data, the summer season is more appropriate (median NSE(logQ)=0.75) than the winter season (median NSE(logQ)=0.64). The same was shown in the previous section. Again, the parameters obtained based on

one summer of log-transformed data perform equally well as parameters that were ob-tained from calibration on one year of log-transformed data.

A possible explanation for the fact that the summer season performs better for calibra-tion than the winter season, is that in the Rietholzbach catchment the summer season is wetter (more precipitation) than the winter season (see Figure 2.2). Calibration is, amongst others, based on precipitation input data. The best result is apparently obtained during high precipitation (summer). Another explanation could be the absence of uncertainty related to snow processes in this period.

The above results show that, with the applied calibration strategy, the model is not able to describe both the high and the low flows with a single set of parameters. This seems to imply that the model structure is inappropriate, and probably a too simplified represen-tation of reality. On the other hand, the response surface (Figure 3.7) shows correlation between the two parameters of the model, suggesting that the model is already over-parameterized *Fowler et al.* (2016) demonstrate that a different calibration strategy (e.g. based on a Pareto-front) can improve the model performance for different flow condi-tions using a single parameter sets. Before any conclusions can be drawn about the ap-propriateness of the model structure, such a strategy should be tested. However, with the same data set, namely with the length of one season, the optimal parameter set for high flows (direct calibration on the data) and the optimal parameter set for low flows (cali-bration on the log-transformed data) can be determined. Therefore, for the application in ungauged basins, a measurement campaign with the length of one season seems able to provide a good first estimate for the system parameters. This should, dependent on the climate, preferably be the summer season.

Predictive uncertainty

In this section the predictive uncertainty is estimated based on the MCMC approach of DREAM. Stochastic interpretation of results obtained with the SSE-based likelihood func-tion assumes independent and identically distributed (i.i.d.) errors with a normal distri-bution. Diagnostics on the residuals showed significant (at the 95% confidence level) au-tocorrelation for the first five time steps. The assumption of an identical distribution of the residuals is not met because the residuals tend to increase with increasing discharge (i.e. the residuals show heteroscedasticity) . Because of the violation of the assumptions (as discussed in Section 1.4 under uncertainty analysis), the predictive uncertainty de-rived from the MCMC results will be biased.

To estimate the predictive uncertainty of the optimized parameters, the full parameter sample obtained with DREAM after convergence according to the Gelman and Rubin criterion (*Gelman and Rubin*, 1992). The calibration-months where no convergence was reached have been excluded from this analysis, hence predictive uncertainty for calibra-tion at a one-month period is underestimated.

Figure 3.8 focusses on the predictive uncertainty of simulations for an extreme flood event (1994) and a severe summer drought (2003). The figure shows that the discharge range

around the flood peak is large if only one month - only January is shown - is used for calibration. The range becomes smaller when the calibration period is increased to one summer season, although all the parameters in the converged posterior obtained with calibration on a summer season miss the first peak of the flood. Remarkably enough the predictive uncertainty does not clearly decrease when one winter season or a year of data was used for calibration.

For predicting a drought event (lower panels in Figure 3.8), it is again visible that the width of the discharge range declines when the calibration period is extended from a monthly calibration period to a summer season calibration period. The predictive uncertainty band for calibration on the winter season and a year of data is broad; even broader than the predictive uncertainty band for calibration on January only. This is consistent with the results shown in Figure 3.5 and Figure 3.6 for NSE(logQ); for automatic calibration directly on the data, the median NSE(logQ) for calibration on a monthly time period is higher than the NSE(logQ) obtained with a calibration period of a winter season and a year. The predictive uncertainty for parameters obtained with calibration on the log-transformed data lead to similar results (not shown), although in this case the drought is projected much better than the flood. Again, the discharge range for calibration on the summer season is clearly smaller than the discharge range obtained with calibration on a winter season or one year of data.

Recession analysis

Figure 3.5 (second and fourth row) shows the results for the recession analysis, both when the equation was fitted on linear axes (linRA) and on logarithmic axes (logRA). The pa-
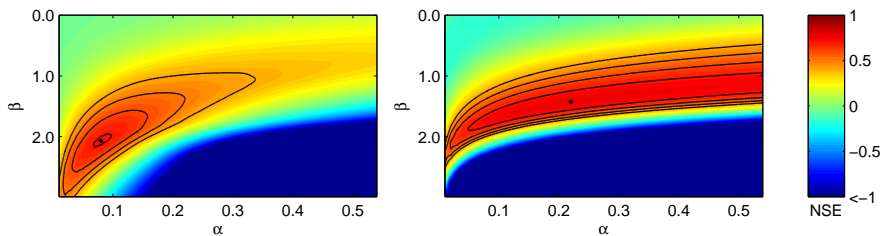
Figure 3.7: Fraction of the response surface for the two model parameters α and β based on 32 years of data (note α is only shown from 0 to 0.5). The left panel shows the response surface for direct calibration on the data (NSE(Q), the right panel for calibration on the log-transformed data, NSE(logQ). Contours are shown for NSE=0.4, 0.5, 0.6 and 0.7. The point indicates the highest model efficiency as found in the part of the response surface that is shown. In the right panel, a ridge with high model efficiencies is found, while in the left panel the area with the highest model efficiencies is concentrated around α=0.1. The figures also show the correlation that exists between both parameters.
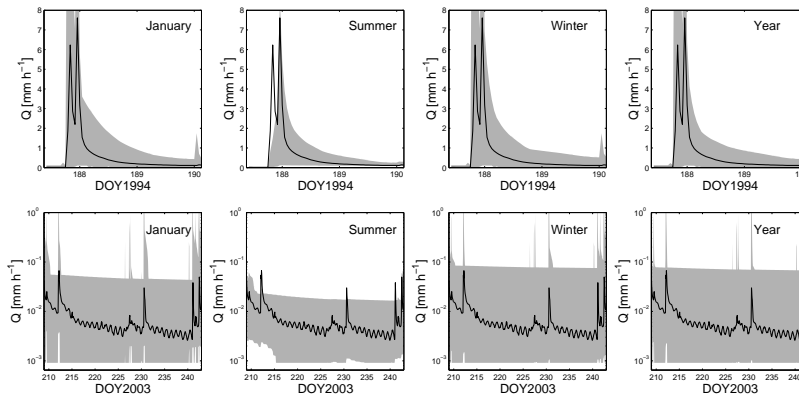
Figure 3.8: The black line shows the measured discharge for a flood event in 1994 (upper panels) and a drought event in 2003 (lower panels). The grey area shows the predictive uncertainty, based on the converged part of the posterior for all 32 calibrations that have been executed per calibration period. The panels show different calibration periods, from left to right: Month (only January is shown), summer season, winter season, year. The parameters in the posterior were obtained with calibration directly on the data, which can explain the poor performance for the drought event.

rameter values obtained with recession analysis on the full period of 32 years were α=0.094, β=0.747 (NSE(Q)=0.58, NSE(logQ)=-0.99) for fitting on linear axes, and α=0.106, β=0.894 (NSE(Q)=0.63, NSE(logQ)=-0.33) for fitting on logarithmic axes.

Figure 3.5 shows that the two recession analysis methods do not show many differences, although the logRA method seems to perform slightly better, especially in case of low flows. The model performance for high flows reaches acceptable values as soon as one year of data is available for calibration for each of the recession analysis methods, with a median model performance of NSE(Q)=0.58 for linRA and median NSE(Q)=0.60 for logRA. Figure 3.6, however, shows that model performance based on a calibration period of one winter is even higher (NSE(Q)=0.62 and NSE(Q)=0.63 for linRA and logRA respectively). This is in contrast with the automatic calibration results, where summer season was found to be more effective.

Recession analysis is mainly based on discharge data, and shows to perform best during high discharge (winter). It has been shown before that calibration for a wet catchment is easier than for a dry catchment (*Perrin et al.*, 2007; *Yapo et al.*, 1996; *Sorooshian*, 1983), although 'wet' can be interpreted both as high precipitation (for automatic calibration) or high discharge (for recession analysis). Another explanation why recession analysis performed better during winter is that recession analysis uses only part of the data supplied, namely the data that met the criteria as described in Section 3.2. For the summer season,

on average only 1.4% of the data was left for analysis, for the winter season this number was more than five times as large: 8.0% on average. This puts more focus on high flows and can also explain the poor performance of recession analysis in simulating low flows. The third and probably most important explanation for the better performance of recession analysis on winter data above summer data is the effect of evapotranspiration. Due to the diurnal cycle in ET, the recession can show small increases during night time (*Teuling et al.*, 2010). During this small part of the recession, $-\frac{dQ}{dt}$ becomes negative. These negative values were taken into consideration when the data points were binned, and can therefore suppress the value of the bin. This effect is larger in summer than in winter. Especially the fact that the performance of the model for calibration on the winter season is better than the performance of the model calibrated on one year of data, seems to imply that evapotranspiration is an important factor.

In addition to fitting the line through the (binned) data points (see Figure 3.2 right panel), a 68% confidence interval for the fit was determined, in order to compare the parameters obtained from recession analysis with the parameters obtained from automatic calibration. For each recession analysis parameter set, it was checked whether the obtained parameters from automatic calibration with the same calibration period fell within the 68% confidence bounds. This has only been done for the automatic calibration parameters obtained from calibration directly on the data and not for automatic calibration on the log-transformed data, because the recession analysis performance is poor in this case. For the α and β obtained with DREAM based on one month of data, only 10.4% and 12.0% for linRA and logRA fell, respectively, within the 68% confidence bound. For summer season, this was 9.4% for both linRA and logRA, decreasing to 0% for winter season calibration and longer calibration periods. Recession analysis and automatic calibration thus lead to substantially different parameter values. Most of the parameter sets obtained with automatic calibration did not fall within one standard deviation from the parameter sets obtained with recession analysis.

Recession analysis is especially relevant in areas where precipitation measurements are inaccurate. In these areas, a seasonal data set (five months, preferably a winter season with relatively low evapotranspiration) for recession analysis can provide a first estimate of the model parameters. The logarithmic recession analysis (logRA) method is advised above the linear recession analysis (linRA) method because it consistently leads to better results. Somewhat counter-intuitive, recession analysis with the simple dynamical systems model seems inappropriate if the focus of the study is on the prediction of low flows.

Boussinesq analysis

For calculation of the model parameters α and β according to the Boussinesq theory, a range of values was used for the saturated hydraulic conductivity, the porosity, and the shape parameter of the saturated hydraulic conductivity profile (see Table 3.1), because the exact value for each of these parameters was unknown.

If no validation data were available and if no further knowledge on the in-situ param-

Table 3.2: The highest NSE(Q) and NSE(logQ) per shape parameter value (n) for the Boussinesq model, with the accompanying porosity, saturated hydraulic conductivity and α-value

	β	NSE(Q)	φ	k	α	NSE(logQ)	φ	k	α
n=0	0	0.49	0.3	0.0360	0.017	-0.15	0.1	$6.5 \cdot 10^{-5}$	$9.2 \cdot 10^{-5}$
n=1	0.50	0.60	0.2	1.8	0.032	0.48	0.2	0.1170	0.008
n=10	0.91	0.69	0.1	0.0360	0.092	0.65	0.2	$6.5 \cdot 10^{-5}$	0.026
n=20	0.95	0.71	0.25	0.1170	0.081	0.64	0.35	$6.5 \cdot 10^{-5}$	0.040
n=30	0.97	0.71	0.35	0.1170	0.086	0.42	0.35	$6.5 \cdot 10^{-5}$	0.068
n=40	0.98	0.70	0.35	$6.5 \cdot 10^{-5}$	0.096	0.14	0.35	$6.5 \cdot 10^{-5}$	0.096
n=50	0.98	0.61	0.35	$6.5 \cdot 10^{-5}$	0.124	-0.14	0.35	$6.5 \cdot 10^{-5}$	0.124
n=60	0.98	0.44	0.35	$6.5 \cdot 10^{-5}$	0.152	-0.42	0.35	$6.5 \cdot 10^{-5}$	0.152
n=70	0.99	0.32	0.35	$6.5 \cdot 10^{-5}$	0.181	-0.69	0.35	$6.5 \cdot 10^{-5}$	0.181

eters was available, the estimated ranges for the parameters based on the Boussinesq equation would be very broad. The lowest α found was $2.6 \cdot 10^{-5}$, the highest α equalled 2.542 (both extremes were found for n=0, so a constant saturated hydraulic conductivity profile). The lowest β was zero, the highest β 0.99. In total, 180 different α's were obtained. The broad range of parameter values that has been found do not have any additional value as a priori estimation, hence either more knowledge on the in-situ parameters has to be available, or data to validate the model.

Table 3.2 shows the highest NSE(Q) and NSE(logQ) values obtained per shape parameter n, also giving the porosity and saturated hydraulic conductivity belonging to this model efficiency. The porosity and saturated hydraulic conductivity are for most NSE(Q) and NSE(logQ) values about 0.35 and $6.5 \cdot 10^{-5}$ m h^{-1}, respectively, but remarkably enough these numbers are different for those n-values that score the highest efficiencies. Next to that, there seems to be a relation between the objective function and the saturated hydraulic conductivity corresponding to the highest model efficiencies; the NSE(logQ) criterion produces the highest model efficiencies only with low saturated hydraulic conductivities. It also seems as if the optimal saturated hydraulic conductivity decreases with increasing n for the NSE(Q). It is therefore difficult to determine what the optimal value is of the porosity and the saturated hydraulic conductivity.

There were two n-values with the highest model efficiency for NSE(Q): n=20 and n=30, which implies values of β of respectively 0.95 and 0.97. This approaches the value of β obtained with automatic calibration, which was 0.94. The α for the discussed models varies from 0.081 (n=20) to 0.086 (n=30). The α obtained with automatic calibration on 32 years of data lies exactly within this range (α=0.083). So, based on NSE(Q), the saturated hydraulic conductivity profile in the Rietholzbach is described with an n-value between 20 and 30 in Eq.3.4, a drainable porosity between 0.25 and 0.35 and a saturated hydraulic conductivity of 0.1170 m h^{-1}. These numbers are different if NSE(logQ) is used as reference. For focus on low flows, NSE(logQ) values were highest for n=10 and n=20. This lead to a β of respectively 0.91 and 0.95. The α obtained with the Boussinesq approach for these models varied from 0.026 (n=10) to 0.040 (n=20). This is almost one

order of magnitude lower than the α obtained with calibration on the logarithm of 32 years of data, which was 0.220. This difference is probably caused by the difference in β, which was 1.571 after automatic calibration, but between 0.91 and 0.95 for the Boussinesq approach. With the employed Boussinesq method, it is not possible to obtain a β larger than one (see Equation 3.3). It therefore seems that the Boussinesq method of *Rupp and Selker* (2006) is appropriate to give an estimate of parameters for the higher discharge ranges, but not for the lower discharge ranges, even though this method was developed for the late part of the recession (see Section 3.2).

Comparison of the three methods

Three different methods to determine the parameters of the simple dynamical systems model have been employed: automatic calibration, recession analysis, and Boussinesq theorem. Figure 3.9 shows the distribution of the model efficiency NSE(Q) for the models obtained with automatic calibration using the DREAM algorithm, recession analysis fitted on logarithmic axes (logRA) and the Boussinesq theory based on the different input ranges for parameters with unknown values (see Table 3.1). For each calibration period except for the winter season, automatic calibration scores the highest model efficiencies. Recession analysis seems a useful alternative for automatic calibration if precipitation data are inaccurate. The highest model efficiency that was found was NSE(Q)=0.71 for calibration on 32 years of data, but with Boussinesq theory equally high model efficiencies were found (see Table 3.2). The Boussinesq theory therefore seems applicable to obtain a-priori estimates for model parameters, but the distribution of model efficiencies is broad due to uncertainty in several catchment properties.

When focussing on low flows and thus comparing NSE(logQ) among the different methods, the picture looks totally different (not shown). The median NSE(logQ) for all recession analysis calibration periods except two lies below zero, and the same holds for the median NSE(logQ) from the Boussinesq theory. On the contrary, calibration with DREAM on log-transformed data scores up to NSE(logQ)=0.76 for 32 years of data. For a study with a focus on low flows, recession analysis and Boussinesq theory are therefore not suitable and automatic calibration is inevitable.

3.4 Implications and limitations of this case study

Gan et al. (1997) reported the following four factors that are important during a calibration procedure: The optimization algorithm, the objective function, calibration data, and the model structure. Although *Gan et al.* (1997) focused only on automatic calibration, in this study the 'optimization algorithm' was applied in a broader context: Three different methods were employed to identify system parameters. The main focus was on the calibration period, which was varied in length to study the effect of data availability on calibration performance. The objective function, calibration data, and the model structure are discussed below.

Objective function

During calibration, one objective function was used, the SSE, but this has been applied directly on the data and on the log-transformed data. Therefore, two different efficien-
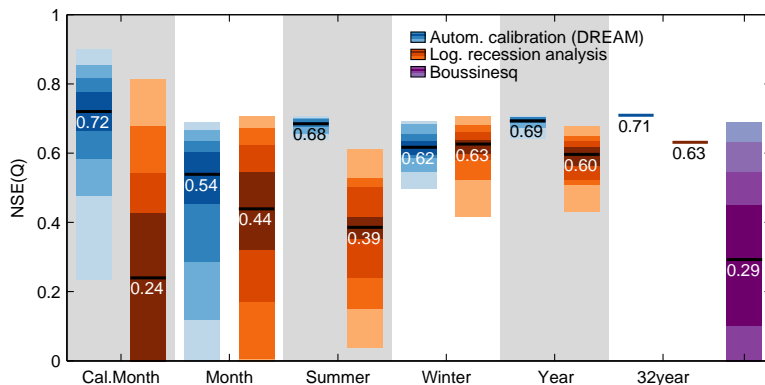
Figure 3.9: Model performance obtained with three different parameter identification methods: automatic calibration with DREAM, recession analysis fitted on logarithmic axes, and Boussinesq theory. The first two box-plots on the left show the model performance for the calibration period of one month. All other box-plots show the distribution of the model efficiency for the full validation period of 32 years for different calibration lengths. The numbers in the chart indicate the median model efficiency NSE(Q). Legend according to Figure 3.5.

cies were calculated: NSE(Q) and NSE(logQ). *Sorooshian* (1983) showed that the obtained optimal parameters can vary for different objective functions. This was confirmed in the current study. The optimal parameters obtained with the SSE directly on the data were different from the parameters obtained with the SSE criterion on log-transformed data. *Kirchner* (2009) and *Teuling et al.* (2010) therefore used different parameter sets for different values of β, because of the physically different behaviour of the catchment if $\beta < 1$, β=1 and $\beta > 1$ (see Section 3.2). This study showed that the parameter sets for both criteria can be determined from the same data set, by calibrating either directly on the data or on the log-transformed data. The objective function should thus reflect the goal of the study. Other objective functions might have lead to different parameter values, but the objective function (as long as it is based on the discharge) does not influence the data needs for robust calibration.

Calibration data

Thirty-two years of hourly observations were used for calibration and validation. Important was that actual evapotranspiration data were available from lysimeter measurements. In most other catchments, and especially in ungauged catchments, actual evapotranspiration data are not available. In these basins the potential evapotranspiration has to be estimated first, e.g. based on the Priestley-Taylor or Penman-Monteith models, and subsequently the actual evapotranspiration has to be estimated from the potential evapotranspiration. This increases the uncertainty and the number of parameters in the model, and hence might increase the data requirements for calibration. On the other hand, given the constant development in satellite observations, evapotranspiration data might

be available for many regions in the near future.

Another characteristic of the calibration data was the hourly resolution of the data. In comparative research (e.g. *Sorooshian* 1983; *Perrin et al.* 2007; *Seibert and Beven* 2009) data with a daily resolution was applied. A coarser time resolution of the data might lead to different results as far as required data length is concerned, but this was not studied in this research. The required time resolution depends on the goal of the study, but considering the measurement equipment available nowadays, an hourly measurement frequency during a short measurement campaign is certainly feasible.

Last but not least, the quality of the data determines successful parameter identification. Because the measured water balance of the Rietholzbach does not close (see Table 2.1), the data series contain disinformative data (*Beven and Westerberg*, 2011). By not taking into consideration these disinformative data, the required calibration data length might be overestimated. Although beyond the scope of this study, it would be an interesting exercise to identify and remove data inconsistencies beforehand (see e.g. *Kauffeldt et al.*, 2013), followed by a similar exercise as performed in this Chapter. Despite the presence of disinformative data, the available data for the Rietholzbach are considered of high quality (*Seneviratne et al.*, 2012). More or larger measurement errors and a lower accuracy of the data, which is not uncommon within the field of hydrology, might increase the data needs for parameter identification.

Model structure

The dynamical systems model that has been applied in this study has a very simple structure. The model assumes no hysteresis, and is therefore mainly applicable in catchments with limited or no hysteresis (e.g. the Severn and Wye catchment in Wales, which were employed by *Kirchner* 2009). In the Rietholzbach catchment, hysteresis does exist, but still the model performs well (*Teuling et al.*, 2010). This implies that the simple model does not give a physically realistic representation of all the separate processes in the catchment and is a strong simplification of reality. This introduces model structural uncertainty in the results. Model structural uncertainty might be lower for more physically based models, but this kind of models will again introduce more uncertainty due to the large amount of required input data or the large number of parameters. Nevertheless, the employed model structure has much in common with the subsurface representation in more complex models. Therefore it is believed that the results presented in this chapter can be extended to more complex models. In addition, the comparison between the three different parameter identification methods is expected to be independent from the model structure and thus broadly interpretable.

3.5 Conclusion

The minimal data requirement for automatic calibration of the two parameters in the simple dynamical systems model (*Kirchner*, 2009) for the Rietholzbach catchment in Switzerland to reach reasonable model efficiencies is one season of five months. This should preferably be the summer season (May until September, with median NSE(Q)=0.69 and NSE(logQ)=0.74). For focus on high flows, direct calibration on the data with the SSE

as objective function will provide the optimal system parameters. If the focus is on low flows, calibration should take place on the log-transformed data. Longer calibration periods only lead to marginal improvements.

For recession analysis, also one season of available data is found as threshold for reasonable model results, although in this case the winter season (November until March, with a median NSE(Q)=0.63) outperforms the summer season. The logarithmic recession analysis (logRA) leads to higher model efficiencies than linear recession analysis (linRA) and is therefore advised. Longer calibration periods for logarithmic recession analysis do not lead to higher model efficiencies. Recession analysis seemed inappropriate to identify the parameters necessary to simulate low flows in the right way with this model, and should therefore only be used for focus on high flows.

With Boussinesq theory an a-priori estimate of the parameters could be obtained. There were many uncertainties in the physical characteristics of the catchment, necessary to substitute in the corresponding expressions of α and β. This uncertainty lead to an extremely broad estimation range of the model parameters and thus model efficiencies. In order to make the a priori estimates of added value, either more knowledge on in-situ parameters, or validation data to determine the in-situ parameters indirectly, should be available. It was shown that the employed Boussinesq model adapted for sloping aquifers (*Rupp and Selker*, 2006) is able to give good estimates of the model parameters, although a large restriction is that β cannot exceed one and the model therefore focusses more on the higher discharge ranges.

Comparing the three employed parameter identification methods, it can be concluded that automatic calibration leads to the best model results for high flows. Recession analysis only scores slightly lower with the same calibration length and is therefore a good alternative for automatic calibration when precipitation measurements are inaccurate. The identified parameter values with both methods, however, differ significantly. Also for low flows, automatic calibration scores the highest model efficiencies. Boussinesq theory and recession analysis are unable to identify the parameters to model low flows with reasonable model efficiencies. For this application, automatic calibration is inevitable.

Chapter 4

Parameter transferability across spatial and temporal resolutions

The transfer of parameter sets over different temporal and spatial resolutions is common practice in many large-domain hydrological modelling studies. The degree to which parameters are transferable across temporal and spatial resolutions is an indicator for how well spatial and temporal variability are represented in the models. A large degree of transferability may well indicate a poor representation of such variability in the employed models. To investigate parameter transferability over resolution in time and space we have set-up a study in which the Variable Infiltration Capacity (VIC) model for the Thur basin in Switzerland was run with four different spatial resolutions (1×1 km, 5×5 km, 10×10 km, lumped) and evaluated for three relevant temporal resolutions (hour, day, month), both applied with uniform and distributed forcing. The model was run 3,150 times using a Hierarchical Latin Hypercube Sample and the best 1% of the runs was selected as behavioural. The overlap in behavioural sets for different spatial and temporal resolutions was used as indicator for parameter transferability. A key result from this study is that the overlap in parameter sets for different spatial resolutions was much larger than for different temporal resolutions, also when the forcing was applied in a distributed fashion. This result suggests that it is easier to transfer parameters across different spatial resolutions than across different temporal resolutions. However, the result also indicates a substantial underestimation in the spatial variability represented in the hydrological simulations, suggesting that the high spatial transferability may occur because the current generation of large-domain models have an inadequate representation of spatial variability and hydrologic connectivity. The results presented in this Chapter provide a strong motivation to further investigate and substantially improve the representation of spatial and temporal variability in large-domain hydrological models.

4.1   Introduction

In Section 1.1, it was described how hydrological modelling started with the empiricist approach in the 17th century. According to *Todini* (2007), the history of modern hydrological modelling dates back to halfway the nineteenth century, starting with empirical models to predict peak flows. For a long time, hydrological models were developed only at the catchment scale, evolving from empirically based to more physically based. Computational power and increased data availability have driven the development of increasingly complex and distributed hydrological models (*Boyle et al.*, 2001; *Liu and Gupta*, 2007). Distributed hydrological models can incorporate spatially varying parameters, including those reflecting land use and soil characteristics (*Carpenter and Georgakakos*, 2006), and spatially variable forcing. In 1989 the first Global Hydrological Model (GHM) was presented (*Vörösmarty et al.*, 1989; *Sood and Smakhtin*, 2015). Continuing improvements in computational power and data availability provides new opportunities for GHMs, for example expressed in the recent ambition to develop global models with a resolution in the order of ∼1 km and higher, the so-called hyper-resolution (*Wood et al.*, 2011; *Bierkens et al.*, 2015; *Bierkens*, 2015).

The increased complexity of hydrological models and the increased application domain has resulted in more complex and time consuming optimization procedures for the model parameters. Although recent developments in e.g. satellites and remote sensing can provide spatially distributed data to construct and force models, discharge observations (or other model output variables) are still required to calibrate and validate model output. Both to decrease calculation time of the optimization procedure, and to be able to apply the model in ungauged or poorly gauged basins and areas, many studies have focused on the transferability of parameter values over time, space, and spatial and temporal resolution (for example *Wagener and Wheater*, 2006; *Duan et al.*, 2006; *Troy et al.*, 2008; *Samaniego et al.*, 2010; *Rosero et al.*, 2010; *Kumar et al.*, 2013; *Bennett et al.*, 2016).

Sometimes it is assumed that parameters are directly transferable, for example by calibrating on a coarser time step than the time step at which the model output will eventually be analysed (e.g. *Liu et al.*, 2013; *Costa-Cabral et al.*, 2013). *Troy et al.* (2008) rightly question what the effect is of calibrating at one time step and transferring the parameters to another time step. Their results suggest that the time step had only limited impact on the calibrated parameters and thus on the monthly runoff ratio. On the other hand, *Haddeland et al.* (2006) found that modelled moisture fluxes are sensitive to the model time step. Several studies (e.g. *Littlewood and Croke*, 2013; *Kavetski et al.*, 2011; *Wang et al.*, 2009; *Littlewood and Croke*, 2008) have found that parameter values are closely related to the employed time step of the model. *Chaney et al.* (2015) investigated to what extent monthly runoff observations could reduce the uncertainty around the flow duration curve of daily modelled runoff. They found a decrease in the uncertainty around the flow duration curve when the monthly discharge observations were used, but the magnitude of the reduction was dependent on climate type. *Ficchì et al.* (2016) conducted a thorough analysis on the effect of temporal resolution on the projection of flood events, where it was shown that the specific characteristics of the flood event determined the sensitivity

for the temporal resolution.

Less intuitive and less common is to transfer parameters across different spatial grid resolutions. *Haddeland et al.* (2002) showed that the output of the Variable Infiltration Capacity (VIC) model was significantly different when the parameters of the model were kept constant for several spatial resolutions. For the same model, *Liang et al.* (2004) showed that model parameters calibrated at a coarse grid resolution could be applied to finer resolutions to obtain comparable results. *Troy et al.* (2008) on the contrary, found that calibrating the VIC model on a coarse resolution significantly affected the model performance when applied to finer resolutions. *Finnerty et al.* (1997) investigated parameter transferability over both space and time for the Sacramento model, and showed that it can lead to considerable volume errors.

Although the ambition of GHMs is to move towards hyper-resolution ($\sim$1 km and higher), more physically-based catchment models have already been applied at spatial resolutions in the order of 100 meters. Also for these models at this scale, the effect of spatial resolution has been investigated (e.g. *Vivoni et al.*, 2005; *Sulis et al.*, 2011; *Shrestha et al.*, 2015). Even for fully coupled surface-groundwater land-surface models, the effect of spatial resolution on hydrologic fluxes was found to be considerable (*Shrestha et al.,* 2015).

The impact of transferring parameters across spatial and/or temporal resolutions on model performance is thus ambiguous, but relevant in the light of hydrological model development, especially for GHMs which are at the upper boundary of computational power and data availability. Calibration on a coarse temporal or spatial resolution and subsequently transferring to a higher resolution could potentially reduce computation time, and it is therefore relevant to investigate the opportunities. But parameter transferability across spatial and temporal resolutions is also interesting for another reason: it is an indicator for the degree to which spatial and temporal variability are represented in the model. Ideally, in a model that describes all relevant hydrological processes correctly, parameters should to a large extent be transferable over time because longer time steps are simply an integration of the shorter time steps (although long-term cycles, e.g. seasonal or decadal cycles, should still be accounted for). On the other hand, parameters should not or hardly be transferable over space, because the physical characteristics which they represent are different from place to place. Investigating parameter transferability across spatial and temporal resolutions can thus provide insight in the model's representation of spatial and temporal variability.

In this study, we employ the Variable Infiltration Capacity (VIC) model (*Liang et al.*, 1994), which has also been applied at the global scale (*Nijssen et al.*, 2001; *Bierkens et al.*, 2015), to study parameter transferability across temporal and spatial resolutions, accounting for the difference between uniform and distributed forcing. A description of the VIC model and the routing procedure can be found in Section 2.2. This study is conducted on a well-gauged mesoscale catchment in Switzerland (the Thur basin, 1703 km^2, see Section 2.1) on spatial resolutions that are relevant for hyper-resolution studies (1$\times$1 km, 5$\times$5 km and 10$\times$10 km, as well as a lumped model which represents the 0.5° grid used in many

global studies). We use the most common temporal resolutions for which discharge data are available (hourly, daily, monthly). We ran the models both with distributed forcing (different forcing input for each grid cell) and with uniform forcing (same forcing input for each grid cell), where the latter is in line with many of the datasets currently used for forcing global hydrological models (e.g. WATCH forcing data, 0.5°).

Several studies already investigated scale effects in the VIC model, for instance *Haddeland et al.* (2002); *Liang et al.* (2004); *Haddeland et al.* (2006); *Troy et al.* (2008); *Wenger et al.* (2010); *Wen et al.* (2012). Novel in this study is that we choose a probabilistic rather than a deterministic approach: essentially we employ a GLUE-based approach (*Beven and Binley*, 1992, 2014) in which we implicitly account for parameter uncertainty. We quantify parameter transferability by evaluating the overlap in behavioural sets for different temporal and spatial resolutions. To determine the behavioural sets, we make use of three different objective functions focusing on high flows, average conditions, and low flows. Novel is also that we test the effect of forcing on the results, and that we use several sub-basins to explain the results. Our case study provides a benchmark for parameter transferability for models applied at larger scales, dealing with the same spatial and temporal resolutions as employed here. The results of our study also provide an indication of the current status of spatial and temporal representation in the VIC model, being representative for a larger group of land-surface models.

4.2   Experimental set-up

We have constructed four VIC models with different spatial resolutions: 1×1 km, 5×5 km, 10×10 km, as well as a lumped model. These models have been run with both uniform and distributed forcing. Since for the lumped model there is no difference between uniform and distributed forcing, this leads to a total of seven different model set-ups. Because the runtime of the model combined with all the post-processing is rather long (on average 2.5 hours for the 1×1 km model on a standard PC), an efficient sampling strategy was designed as discussed in Section 2.3. The sampled parameters were applied uniformly over the catchment, whereas all other soil- and land use parameters have been applied in a distributed fashion. After running the models with the parameter sample, the output was evaluated and the best 1% of the runs was defined as behavioural. The overlap in behavioural sets was used as an indicator for parameter transferability.

Spatial model resolution

Four VIC implementations with different spatial resolutions (0.0109° roughly corresponding to 1×1 km, 0.0558° ≈5×5 km, 0.1100° ≈10×10 km, as well as a lumped model) were constructed. The 1×1 km model represents the so-called hyper-resolution. Several studies already explore GHMs at this resolution, e.g. *Sutanudjaja et al.* (2014) for the Rhine-Meuse basin. The model with the 10×10 km resolution can be characterized as 'regional'. The 5×5 km model is in between the hyper-resolution scale and the regional scale. The lumped model, which represents an area of 1703 km², is in the order of magnitude of grid cells with a 0.5° resolution, which represents the original scale for which VIC was developed. Figure 4.1 gives an overview of the cell size of the four models. The sampled parameters (see Section 2.3) have been applied uniformly over the catchment, all other

parameters have been applied in a distributed manner.

Temporal model resolution

The models are run at an hourly time step, implying that they solve both the energy and the water balance. The hourly output of the routing model is aggregated to daily and monthly time steps for further evaluation, see Figure 4.1.

Simulation period

The models are run for a period of 1 year and four months. The first three months are used as spin-up period and not used for further analysis. Tests with the same parameter set and different initial conditions revealed that three months are sufficient to eliminate the effect of initial conditions (see Figure 4.2). The initial soil moisture content of the model before spin-up was fixed at θ=0.9 because we found that the model reaches equilibrium faster when starting from a wet state. The models have not been subjected to a validation procedure on another time period, because in this particular application the goal was not to identify the best performing model, but to investigate the role of temporal and spatial resolution on parameter transferability.

The analysed period is 1 August 2002 – 31 August 2003 (see Figure 4.2). This period is characterised by three very high peaks (August, September 2002) as well as the severe 2003 drought (June, July, August 2003). The 2002 peaks (see e.g. *Schmocker-Fackel and Naef*, 2010) have an estimated return period of 15 to 20 years. The peaks were caused by a larger system that also caused the heavy floods in the Elbe and the Danube (*Becker and Grünewald*, 2003). In contrast, the 2003 summer was extremely warm and dry in Western and Central Europe (*Miralles et al.*, 2014), with Switzerland being among the hottest and
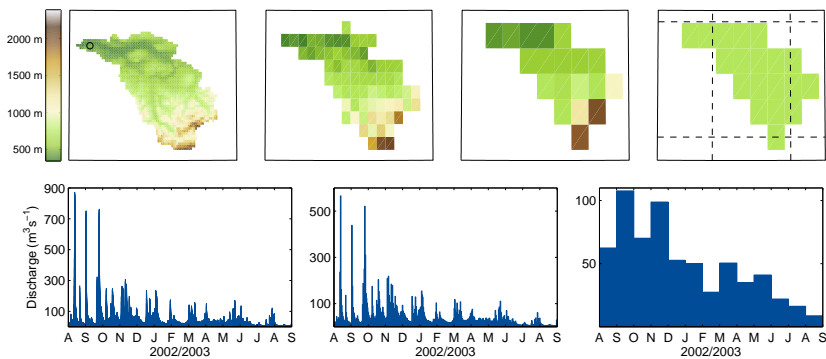
Figure 4.1: *Overview of the spatial and temporal resolutions employed in this study. Top from left to right: DEM grid cells for 1×1 km, 5×5 km, 10×10 km resolution and the lumped model. The circle in the left panel shows the location of the Thur outlet where the discharge is measured. The dotted lines in the right panel indicate a 0.5° grid. Bottom: The three temporal resolutions, observed discharge at an hourly, daily and monthly time step.*

driest regions (*Andersen et al.*, 2005; *Rebetez et al.*, 2006; *Zappa and Kan*, 2007; *Seneviratne et al.*, 2012). With these two extremes the selected period covers a large part of the flow duration curve, both in the high and the low flow regions (right panel in Figure 4.2).

Determination of behavioural sets and parameter transferability

After running the VIC model with 3,150 parameter sets, a selection is made of the best parameter sets, the so-called behavioural runs (*Beven and Binley*, 1992). For each model run, three objective functions were evaluated: the KGE(Q), the NSE(Q), and the NSE(logQ), defined in Eq. 2.11 and 2.12 on p.28. The objective functions are calculated for the seven different VIC set-ups and based on hourly, daily and monthly time steps. The best 1% (which is different for different objective functions and different model set-ups) of the 3,150 runs (32 members) are selected as behavioural. We value all 32 parameter sets equally plausible and do not assign weights to the best performing sets within the behavioural selection, to account for uncertainty in the observations. Inherent to our approach, selecting a certain percentage of runs rather than applying a threshold level based on an objective function, is that the selected runs do not necessarily comply to an acceptable model performance. We expect that this neither positively nor negatively influences our results concerning parameter transferability, because we essentially conduct a sensitivity analysis.

We define parameter transferability $\underleftrightarrow{\theta}$ as the percentage agreement in selected behavioural sets:

$$\underleftrightarrow{\theta} = 100 \cdot \#(A_{S_i,T_j} \cap B_{S_k,T_l})/n, \tag{4.1}$$

in which A_{S_i,T_j} is the set of selected behavioural members for spatial resolution S_i and temporal resolution T_j, and B_{S_k,T_l} are the selected members for spatial resolution S_k and temporal resolution T_l. The n is the total number of selected members (in this case 32). Equation 4.1 expresses $\underleftrightarrow{\theta}$ as a percentage; if $\underleftrightarrow{\theta}$=100, this indicates that for two
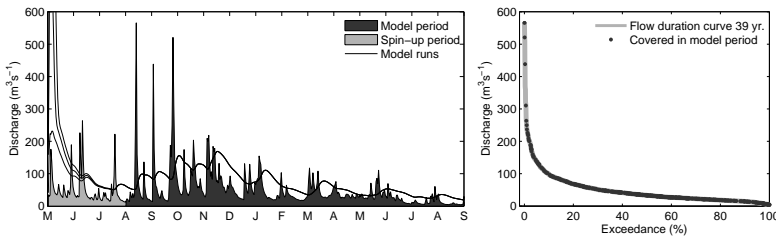
Figure 4.2: Daily discharge characteristics for the Thur basin. Left panel: the daily discharge in the Thur for the selected model period. The black lines show three model runs with the same parameter set but with different initial conditions (θ=0.5,0.7,0.9). Right panel: part of the flow duration curve covered within the model period. The flow duration curve is based on 39 years of daily discharge observations in the Thur for the period 1974–2012.
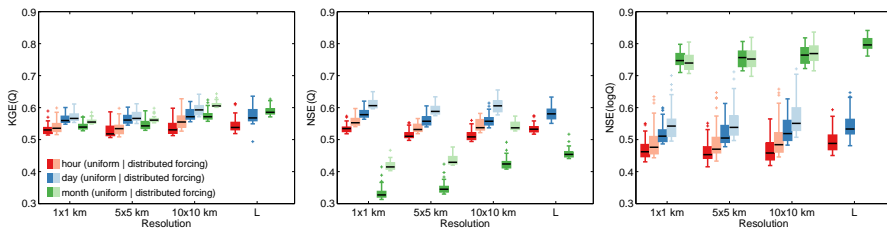
Figure 4.3: Model performance of the behavioural sets for different temporal resolutions and different spatial resolutions. The left panel shows the KGE(Q), the middle panel the NSE(Q) and the right panel the NSE(logQ). Per objective function the most behavioural sets were selected. The box shows the 25–75% quantile.

different resolutions (either spatial, temporal or both) exactly the same parameter sets were selected as behavioural.

4.3 Results

First, the impact of temporal and spatial resolution on model performance is discussed for both uniform and distributed forcing, followed by a discussion of the impact of the temporal and spatial resolution on parameter distribution. The parameter transferability across temporal and spatial resolution is assessed by determining the overlap in behavioural sets as defined by Equation 4.1. After that, parameter transferability over temporal and spatial resolution combined is assessed. Finally, we investigate parameter transferability over the sub-basins of the Thur.

Impact of temporal and spatial resolution on performance and parameter distribution
Figure 4.3 shows the model performance of the behavioural sets for the different spatial and temporal resolutions and the different objective functions, both for uniform and distributed forcing. We will first discuss the results for the uniform forcing.

With uniform forcing, the lumped model outperforms the distributed models for all three objective functions and time steps. The monthly time step shows for all three objective functions an increasing model performance with decreasing spatial resolution. It is remarkable that the model with the monthly time step outperforms the models with daily and hourly time step when the NSE(logQ) was used as objective function, while with the NSE(Q) as objective function exactly the opposite is the case. It is important to notice here that the monthly model results are simply an aggregation from the hourly model results which might imply that the higher score on the monthly time step is the result of errors which compensate for each other, and that the model performance scores for the monthly time step are based on a considerable lower number of points. The KGE(Q) as objective function does not lead to a remarkably different model performance for the monthly time step. From the figure it seems that both the spatial and temporal resolution have impact on the model performance. This is confirmed with a statistical test.
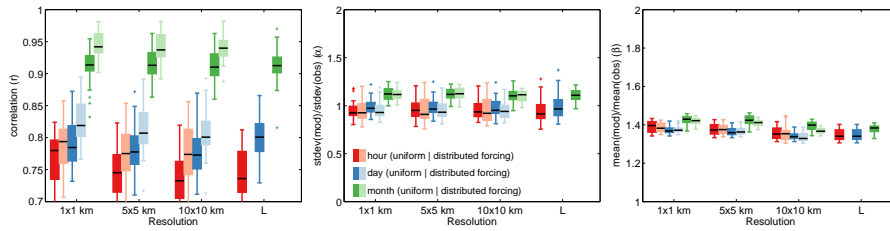
Figure 4.4: The model performance for the three separate components of the Kling-Gupta Efficiency of the behavioural sets for different temporal and spatial resolutions. The left panel shows the correlation r, the middle panel the standard deviation of the model output divided by the standard deviation of the observations (α), and the right panel shows the mean of the model output divided by the mean of the observations (β).

Analysis of Variance (see description on p.86) with two factors (temporal resolution; spatial resolution), with three, respectively four levels (hourly, daily, monthly; 1×1 km, 5×5 km, 10×10 km and lumped) shows that both the spatial and the temporal resolution have significant ($p<0.05$) impact on all three objective functions.

Distributed forcing leads in all cases except one (1×1 km, monthly, NSE(logQ)) to an improved model performance compared to uniform forcing. Note that for the lumped model uniform and distributed forcing are the same - it should therefore be remarked that while with uniform forcing the lumped model outperforms the other model set-ups, for the distributed forcing the 10×10 km model outperforms the other spatial resolutions (except for NSE(logQ)). An ANOVA analysis confirmed that also for distributed forcing, both spatial and temporal resolution have significant ($p<0.05$) impact on the model performance for all three objective functions.

Figure 4.4 shows the distribution of the behavioural sets for the three separate components of the KGE(Q). Regarding the correlation r, the monthly time step scores higher than the daily and hourly time step. On the other hand, the hourly and daily time steps score higher with respect to the β (closer towards 1). Although Figure 4.3 gives the impression that the model performance in terms of KGE(Q) is relatively insensitive to temporal and spatial resolution, Figure 4.4 reveals this is actually the result of compensations from the three different components of the KGE(Q): The monthly time step has a higher correlation, while the daily and hourly time steps have a lower bias (higher β).

Figure 4.5 shows the parameter distribution of the seven sampled parameters (Table 2.4), and shows how the distribution varies as a result of temporal and spatial resolution, both for distributed and uniform forcing. The distribution of the behavioural parameter sets for the daily and hourly time steps are very much alike for all parameters, but the distribution for the monthly time step is in some cases broader, which implies that the parameters are less clearly defined. The parameter showing the clearest effect of tempo-
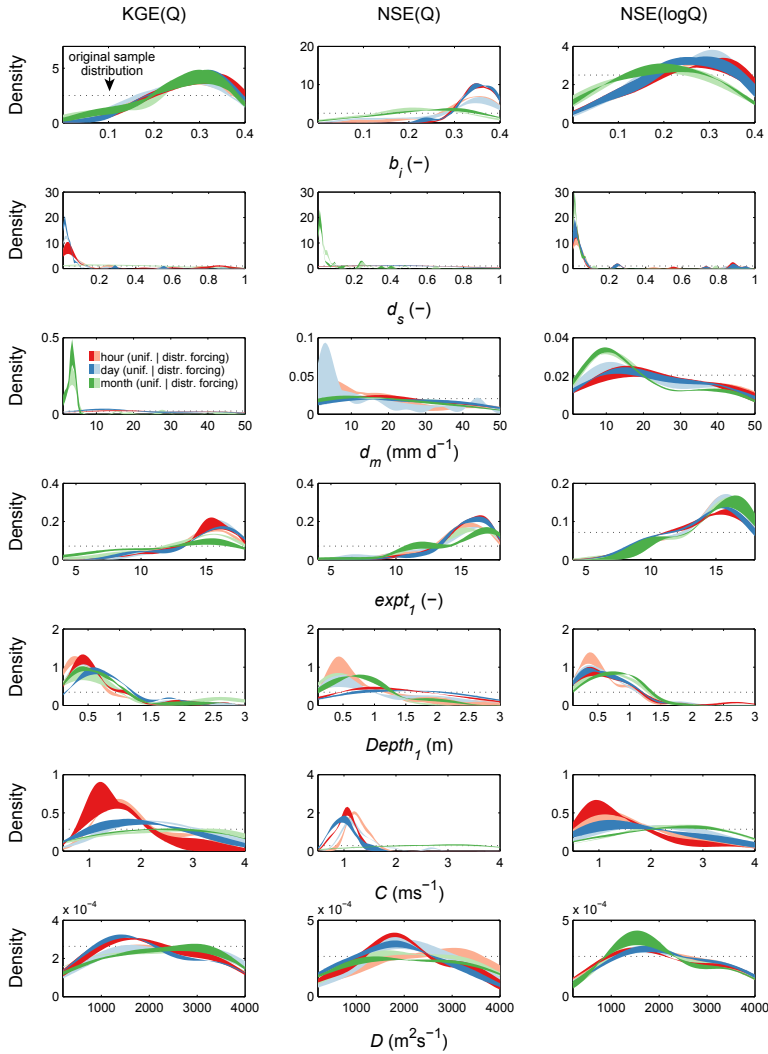
Figure 4.5: Distribution of the sampled parameters for the behavioural sets, fitted with a kernel-density. The width of the line indicates the variation in distribution between the different spatial resolutions. The left column is based on KGE(Q), the middle column on NSE(Q) and the right column on NSE(logQ).
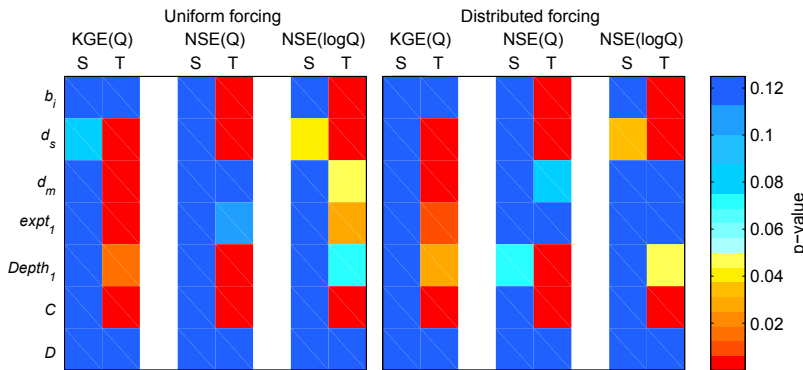
Figure 4.6: The effect of spatial and temporal resolution on parameter distribution. The p-value indicates the significance of the impact of spatial resolution (S) and temporal resolution (T) on the parameter values of the behavioural sets, evaluated for the three objective functions using Analysis of Variance (ANOVA).

ral resolution is the advection coefficient C (Figure 4.5). The C parameter, the velocity component in the routing, becomes less well defined with increasing time step, which is intuitive because timing becomes less relevant for longer time intervals.

The difference in the parameter distribution when comparing distributed and uniform forcing is limited. The clearest difference can be found for the d_m parameter with the NSE(Q) as objective function. This parameter describes the maximum base flow, and can potentially impact short term processes for which distributed forcing seems important, like surface runoff. However, there are other parameters, such as the b_i parameter, which are more directly linked to infiltration and surface runoff processes and do not show a clear difference in parameter distribution between distributed and uniform forcing.

With an ANOVA analysis, the significance of temporal and spatial resolution on the parameter distribution of the behavioural sets was tested. Figure 4.6 shows that the significance of spatial and temporal resolutions on the parameter distribution depends on which objective function was used to determine the behavioural sets. Uniform and distributed forcing show comparable patterns. In general, the temporal resolution has more impact on the parameter distribution (at least four parameters are significantly affected by temporal resolution) than the spatial resolution (only one parameter for one objective function experiences significant impact of the spatial resolution). Only two parameters are significantly impacted by the temporal resolution for all three objective functions: d_s and C.

Parameter transferability
The main research question in this chapter is to what extent parameters are transferable across temporal and spatial resolutions, and we will use that as indicator for the repre-

sentation of spatial and temporal variability in the model. We have defined parameter transferability $\underleftrightarrow{\theta}$ as the percentage agreement in identified behavioural sets (Equation 4.1). Table 4.1 and Table 4.2 give an overview of $\underleftrightarrow{\theta}$ for different temporal and spatial resolutions, respectively, both for uniform and distributed forcing. Table 4.1 shows that the $\underleftrightarrow{\theta}$ is generally high for different spatial resolutions, which suggests that the parameters are to a large extent transferable across spatial scales. In contrast, Table 4.2 shows that parameters are hardly transferable over the temporal scale. The selected sets for hourly and daily time steps largely agree, but the selected sets on a monthly time step are clearly different. Surprisingly, this is also strongly related to the objective function. The selection based on the NSE(logQ) is less sensitive to temporal resolution than those based on the KGE(Q) or the NSE(Q). A possible explanation is that the NSE(logQ) tends to put more focus on lower discharges with a longer time scale, with less focus on the short term flashy response of a catchment. Parameter transferability over space is in general slightly lower when distributed forcing is used compared to uniform forcing. On the other hand, parameter transferability over time is slightly higher for distributed forcing. Decreased sensitivity for the temporal resolution and increased sensitivity for the spatial resolution can indicate an improved physical representation with distributed forcing compared to uniform forcing, as one would expect.

Table 4.1 and 4.2 list the parameter transferability over only one dimension (either spatial resolution or temporal resolution). We also investigated the combined effect of transferring parameters over both the spatial and the temporal resolution. Figure 4.7 shows the relative impact of temporal and spatial resolution on parameter transferability based on KGE(Q) for uniform forcing. To illustrate the relative impact of changes in spatial and temporal resolution, we fitted a linear surface through the data points from our study (R^2=0.68). The figure clearly shows that temporal resolution has a stronger impact on parameter transferability than spatial resolution. The linear regression equation that describes the surface in Figure 4.7 is given below:

$$\underleftrightarrow{\theta}_{KGE(Q)} = 83.3 - 12.6 \cdot \frac{T_j}{T_l} - 3.0 \cdot \frac{S_i}{S_k}, \tag{4.2}$$

in which $\frac{T_j}{T_l}$ is the ratio in temporal resolution between the two model set-ups over which parameters are transferred and $\frac{S_i}{S_k}$ is the ratio in spatial resolution (L/L) between the two model set-ups. The effect of temporal resolution on parameter transferability is stronger (slope of 12.6) than the effect of spatial resolution (slope of 3.0). Parameter transferability decreases when the ratio between the original and the intended spatial and temporal resolution increases. The surfaces based on NSE(Q) (R^2=0.60) and NSE(logQ) (R^2=0.75) show a similar behaviour:

$$\underleftrightarrow{\theta}_{NSE(Q)} = 88.6 - 12.8 \cdot \frac{T_j}{T_l} - 2.8 \cdot \frac{S_i}{S_k}, \tag{4.3}$$

$$\underleftrightarrow{\theta}_{NSE(logQ)} = 92.9 - 7.4 \cdot \frac{T_j}{T_l} - 3.6 \cdot \frac{S_i}{S_k}. \tag{4.4}$$

When we fit a surface through the points obtained for the models run with distributed forcing, the linear regression equations (R^2=0.66, 0.67, 0.88 respectively) look as follows:

$$\underset{\longleftrightarrow}{\theta}_{KGE(Q)} = 80.3 - 11.4 \cdot \frac{T_j}{T_l} - 2.6 \cdot \frac{S_i}{S_k}. \tag{4.5}$$

$$\underset{\longleftrightarrow}{\theta}_{NSE(Q)} = 75.3 - 10.3 \cdot \frac{T_j}{T_l} - 4.3 \cdot \frac{S_i}{S_k}, \tag{4.6}$$

$$\underset{\longleftrightarrow}{\theta}_{NSE(logQ)} = 91.3 - 5.4 \cdot \frac{T_j}{T_l} - 2.8 \cdot \frac{S_i}{S_k}. \tag{4.7}$$

Also for the models with distributed forcing, the slope for the temporal resolution is steeper than the slope for spatial resolution, implying that parameter transferability is more sensitive to temporal than to spatial resolution. Compared to uniform forcing, the slope for temporal resolution, and hence the impact of temporal resolution on transferability, is less steep for distributed forcing, while the slope for spatial resolution is on average comparable for both forcing types.
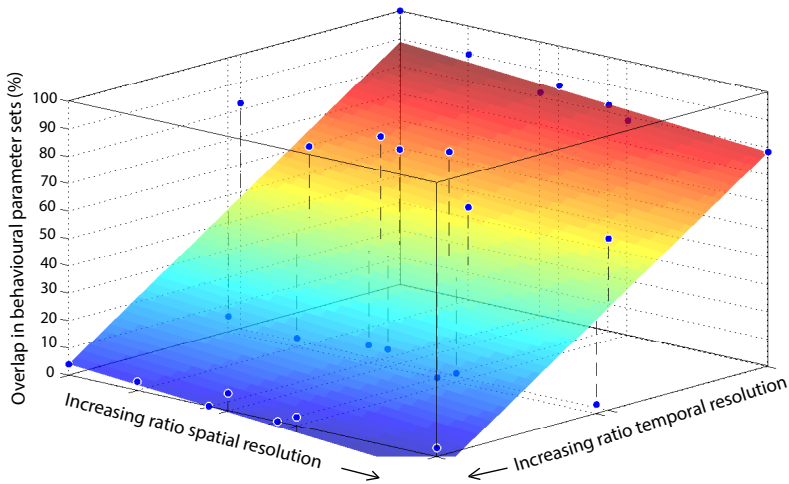
Figure 4.7: Parameter transferability as a function of ratio in temporal and spatial resolution. Ratio of temporal resolutions is defined as follows: transfer from hourly to daily time step is a ratio of 24, whereas transfer from hourly to monthly is a ratio of 732 (732 hours in one month of 30.5 days). The ratio of spatial resolutions is defined as the square root of the number of cells that would fit in the other cell: from 1×1 km resolution to 5×5 km resolution is a ratio of $\sqrt{25}$=5. The behavioural sets were determined based on the KGE(Q). The linear surface (R^2 = 0.68) was fitted to illustrate the relative impact of changes in spatial and temporal resolution.

Table 4.1: *Transferability of parameters across spatial resolution, expressed as percentage agreement in detected behavioural runs for different spatial resolutions (in km) at different temporal resolutions. The cell colours correspond to the percentage agreement (darker = higher transferability).*

	Uniform forcing (% agreement)			Distributed forcing (% agreement)		
			HOUR			
	KGE(Q)	NSE(Q)	NSE(logQ)	KGE(Q)	NSE(Q)	NSE(logQ)
1×1 vs 5×5	78	84	91	88	75	84
1×1 vs 10×10	72	81	81	78	56	78
5×5 vs 10×10	94	94	91	88	81	94
1×1 vs lumped	78	88	91			
5×5 vs lumped	91	84	94			
10×10 vs lumped	88	81	88			
			DAY			
	KGE(Q)	NSE(Q)	NSE(logQ)	KGE(Q)	NSE(Q)	NSE(logQ)
1×1 vs 5×5	94	84	84	91	84	91
1×1 vs 10×10	84	69	69	78	69	81
5×5 vs 10×10	91	84	84	89	84	91
1×1 vs lumped	91	81	88			
5×5 vs lumped	91	88	94			
10×10 vs lumped	84	84	81			
			MONTH			
	KGE(Q)	NSE(Q)	NSE(logQ)	KGE(Q)	NSE(Q)	NSE(logQ)
1×1 vs 5×5	75	88	88	84	84	91
1×1 vs 10×10	66	84	81	66	78	84
5×5 vs 10×10	88	91	94	78	88	94
1×1 vs lumped	78	72	94			
5×5 vs lumped	78	75	88			
10×10 vs lumped	78	78	88			

Table 4.2: *Transferability of parameters across temporal resolution, expressed as percentage agreement in detected behavioural runs for different temporal resolutions at different spatial resolutions. The cell colours correspond to the percentage agreement (darker = higher transferability).*

	Uniform forcing (% agreement)			Distributed forcing (% agreement)		
			1×1 km			
	KGE(Q)	NSE(Q)	NSE(logQ)	KGE(Q)	NSE(Q)	NSE(logQ)
hour vs day	56	81	81	69	63	75
hour vs month	3	6	34	6	9	47
day vs month	3	6	47	6	13	63
			5×5 km			
	KGE(Q)	NSE(Q)	NSE(logQ)	KGE(Q)	NSE(Q)	NSE(logQ)
hour vs day	66	88	81	69	69	81
hour vs month	3	6	38	9	6	53
day vs month	3	6	47	9	6	66
			10×10 km			
	KGE(Q)	NSE(Q)	NSE(logQ)	KGE(Q)	NSE(Q)	NSE(logQ)
hour vs day	63	75	78	59	72	78
hour vs month	3	3	44	13	6	59
day vs month	0	6	63	13	6	75
			lumped			
	KGE(Q)	NSE(Q)	NSE(logQ)	KGE(Q)	NSE(Q)	NSE(logQ)
hour vs day	66	84	81			
hour vs month	3	0	44			
day vs month	3	3	53			
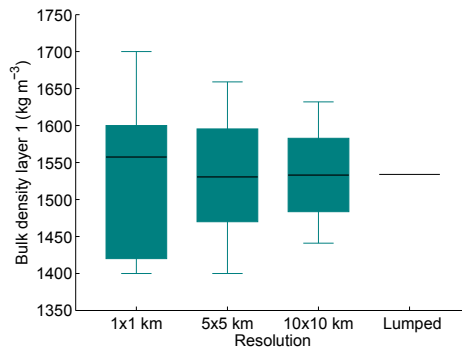
Figure 4.8: Distribution of bulk density over the grid cells for the four different spatial resolutions.

Spatially distributed parameters

The advantage of distributed hydrological models over lumped models is that distributed models can incorporate spatially varying parameters, including those reflecting land use and soil characteristics (*Carpenter and Georgakakos*, 2006), and spatially varying forcing. Figure 4.8 for example, shows how the spatial variation in bulk density decreases with increasing resolution. In this study, as in most large-domain studies with distributed models, the most sensitive parameters (i.e. the once that were calibrated) have, however, been applied uniformly over the grid cells. The main motivation for this practice is that the problem is ill-posed (too many parameters have to be identified with too little information), in addition to computational time. This implies that the advantage of a distributed model remains unused for the parameters that impact the model output most. To test the spatial distribution of the most sensitive parameters for the Thur basin, we have investigated parameter transferability between the Thur basin and the nine sub-basins for which discharge data were available (see Section 2.1 and Figure 2.4). Table 4.3 gives an overview of transferability for a selected number of spatial and temporal resolutions. The table shows that parameter transferability from the Thur to the sub-basins is notably low. An extreme example is the St.Gallen catchment, which has maximum one behavioural parameter set in common with the Thur basin. Table 4.3 therefore shows that the spatial variation in the calibrated parameters is underestimated in the current model set-up.

4.4 Discussion

Model performance

It seems counter-intuitive that the model performance is significantly affected by both the temporal and spatial resolution, while the parameter distribution is mainly impacted by the temporal resolution. This can be explained, however. Model performance can still be significantly impacted by temporal and spatial resolution, even if the same parameters are selected for different spatial resolutions. This implies that the model performance is mainly limited by the model structure or set-up, and much less by the parame-

Table 4.3: Transferability of parameters from the Thur to the nine sub-basins, expressed as percentage agreement (%) in detected behavioural runs. The forcing was applied uniformly and the KGE(Q) was used as objective function. The cell colours correspond to the percentage agreement (darker = higher transferability).

Catchment (size)	1×1 km			5×5 km	10×10 km
	hour	day	month	hour	hour
Rietholzbach (3.3 km^2)	19	0	0	25	19
Herisau (17.8 km^2)	16	6	0	16	16
Appenzell (74.2 km^2)	28	25	9	28	16
Wängi (78.9 km^2)	9	56	31	34	50
Mogelsberg (88.2 km^2)	28	38	66	19	28
Frauenfeld (212 km^2)	3	3	75	3	0
St.Gallen (261 km^2)	3	0	0	3	0
Jonschwil (493 km^2)	6	0	0	6	0
Halden (1085 km^2)	19	9	0	18	13

ter values. This is confirmed by comparing the uniform and distributed forcing. Although the distribution of the behavioural parameters was not very different for the two forcing types, the model performance for distributed forcing was in almost all cases better than the model performance for the uniform forcing.

Liang et al. (2004) defined a so-called 'critical resolution', beyond which a finer spatial resolution would not lead to any improvement in the model performance. In the study of *Liang et al.* (2004) this critical resolution for the VIC model was found to be $1/8°\,(\approx 12.5 \times 12.5$ km). All spatial resolutions applied in this study but the lumped one are below this critical resolution. The results in this study are therefore consistent with the results from *Liang et al.* (2004), because we did not find any improvement in model performance with increasing spatial resolution, neither for the uniform nor for the distributed forcing. Rather, we find the contrary; for the uniform forcing the lumped model outperformed the higher resolution models, and for the distributed forcing the 10×10 km outperformed the other models. If something like a critical resolution exists, it is probably related to the processes represented in the model. Contradictory to our findings are the results of *Zappa* (2002), who found that a critical spatial resolution in the Thur region is in the order of 500×500 m using the PREVAH model, because of the complex topography and snow processes in the catchment. This can either imply that the sub-grid variability parameterization in VIC is effective, or that not all relevant hydrological processes are included in the VIC model. In order to check this last suggestion, future research on parameter transferability should consider more hydrological fluxes and states besides the discharge, e.g. evapotranspiration.

The high sensitivity for temporal resolution
The conclusion that parameters cannot be transferred across temporal resolution seems

to contradict the results of *Troy et al.* (2008). The large difference is that *Troy et al.* (2008) only used sub-daily time steps (1, 3, 6, 12 hours), whereas we did find agreement between the hourly and daily time step. Therefore, our results are not necessarily contradictory. *Troy et al.* (2008) chose the sub-daily time steps in order to investigate if time could be saved in the calibration process by calibrating on a coarser time step. Unfortunately, the reality is that in most large-domain studies models are calibrated with monthly discharge observations (see Chapter 7) rather than with sub-daily observations. Our results suggest that models which were calibrated or validated at a monthly time step cannot be interpreted at the daily or hourly time step. *Chaney et al.* (2015) showed that monthly discharge observations could decrease the uncertainty around the daily flow duration curve. The decrease in uncertainty by adding monthly discharge information differed for different climates. The Thur basin, with a wet continental climate, would experience a high reduction in uncertainty. This means that our results, which show that with monthly data it is impossible to determine the optimal parameter set for the hourly or daily time step, would even be stronger for dry climates (*Chaney et al.*, 2015).

Kavetski et al. (2011) showed that parameter values can significantly change by changing the temporal resolution (this is also demonstrated in Chapter 7). They found that the sensitivity of a parameter to temporal resolution could be related to the model structure; the parameters from simpler model structures were more sensitive to temporal resolution than the parameters from more complex models.

Figure 4.9 shows that the conclusions we draw from Table 4.1 and Table 4.2 are not only valid for the best 1% of runs selected as behavioural. Figure 4.9 gives an overview for two selected cases, which show that model performance deteriorates when parameters are transferred over time, also for the best 10% up to higher thresholds, whereas the impact of spatial resolution on model performance deterioration is limited.

Models versus nature: Do the current generation of models adequately represent spatial variability?
The results of this chapter show that parameter transferability is more sensitive to temporal than to spatial resolution. A key question is to what extent this result stems from the model representation of spatial variability. Spatial variability can be reflected in three domains of the model: the routing, the forcing, and the soil- and land use parameters.

In this study we excluded the effect of routing by using a high-resolution drainage network based on sub-basins with a size of ~1 km^2, independent of the resolution of the hydrological model. We think that the effect of spatial resolution can be increased by adapting the routing scheme accordingly. Drainage network resolution may affect the projected hydrograph, for example with changes in the stream network and the channel slope. However, this effect should then be assigned to the routing model, and not to the runoff generation model (the hydrological model). For clarity, we decided to exclude the effect of spatial resolution on routing in this study.

We investigated the effect of forcing by comparing the results for distributed and uni-
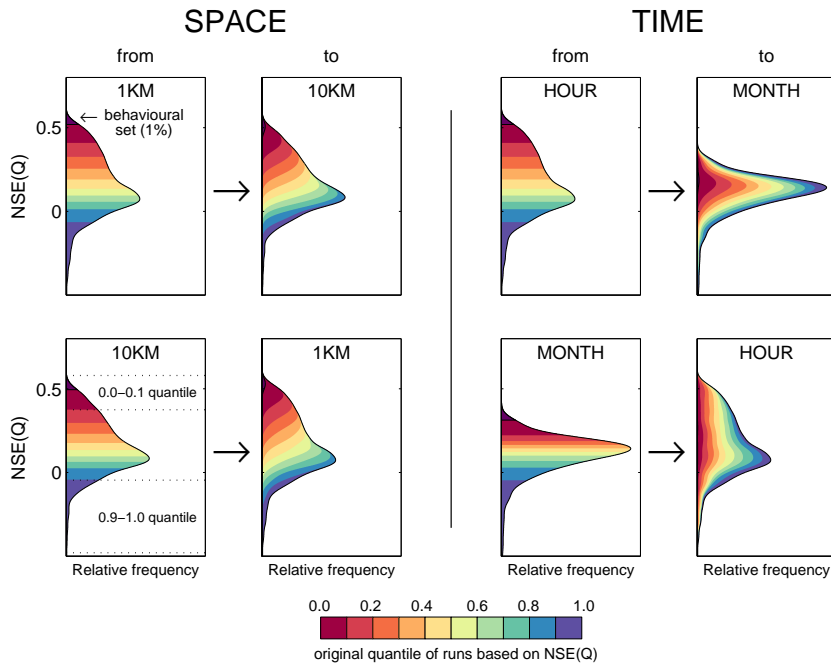
Figure 4.9: Impact of parameter transfer on model performance. The panels show the distribution of the NSE(Q) fitted with a kernel density for 3,150 runs. On the left hand side of the arrow the red area represents the best 10% of the runs, each colour interval increasing with 10% to the full data set (100%, purple). The selected behavioural runs are indicated separately with a black line (best 1%). The panel on the right hand side of the arrow shows the distribution of the model performance for the coloured selections when evaluated at another spatial (left) or temporal (right) resolution. When the direction of the colours changes from the left panel to the right panel, this implies a low parameter transferability. The data for the first two columns are based on hourly discharges, the data for the second two columns are based on the 1×1 km model.

formly applied forcing, and we tested the effect of spatially distributed soil- and land use parameters by aggregating them for lower resolutions (Figure 4.8). Despite distributed forcing and the decrease in variation in soil- and land use parameters, the model parameters showed low sensitivity to the spatial resolution. A possible explanation could be the sub-grid parameterizations of the VIC model for land use and elevation, which decrease the effect of up-scaling these parameters to other resolutions, as shown by *Haddeland et al.* (2002). However, we think that Table 4.3 shows that spatial variability is underestimated by calibrating and applying the most sensitive parameters uniformly over the basin.

The models in this study are configured in a similar way as many current day large-domain hydrological models, using common data like the Harmonized World Soil Database and uniform application of the most sensitive parameters. As such, this study is likely representative for many large-domain studies. The limited sensitivity for spatial resolution is arguable because our implementation of VIC substantially underestimates the spatial variability in nature, and, importantly, that similar issues in representing spatial variability is a common problem in large-domain hydrological modelling (see, e.g., the model configuration in *Mizukami et al.*, 2015). Many studies have considered spatial variability in forcing (*Adams et al.*, 2012; *Lobligeois et al.*, 2014) and soil parameters (*Mohanty and Skaggs*, 2001; *Western et al.*, 2004). *Kim et al.* (1997) accounted for heterogeneity in soil hydraulic properties using stochastic methods, based on the scaling theory of *Miller and Miller* (1956). In fact, the effect of stochastic soil parameterizations on parameter transferability would be a valuable research topic (*Maxwell and Kollet*, 2008). We argue here that the high spatial transferability may occur because the current generation of land-surface models have an inadequate representation of spatial variability and hydrologic connectivity, providing a strong motivation to substantially improve the representation of spatial and temporal variability in models. This not only implies increasing the spatial (and temporal) resolution of the model, but also including more relevant hydrological processes. Promising techniques have been developed to allow spatial distribution of calibrated parameters, for example with Multiscale Parameter Regionalization (MRP, *Samaniego et al.*, 2010; *Kumar et al.*, 2013), which could and should be applied for large-domain hydrological models.

Limitations of this case study
The results in our study are based on a limited number of model configurations for a single basin, so the results presented here are only intended to provide an example of the behaviour in the current generation of land-surface models. Our results show a low sensitivity for the spatial resolution, whether applied with distributed forcing or not. The observed impact of spatial resolution can therefore almost completely be attributed to the effect of spatially distributed soil and land use parameters (including the calibrated ones), which could be substantially underestimated. The impact of temporal resolution on parameter transferability is large. We employed the temporal resolutions for which most hydrological observations are available, thus our results are relevant for practical applications. Based on the work of *Chaney et al.* (2015) we expect that parameter transferability will be lower for arid climates than the numbers we obtained, and based on

the work of *Kavetski et al.* (2011) we expect that parameter transferability will be lower for more parsimonious models. The general message from our study is the surprisingly high spatial transferability, highlighting the need for a focused research effort to improve the representation of spatial variability in large-domain distributed models (GHMs). A possible path forward is to develop computationally frugal process representations, as for example presented by *Hazenberg et al.* (2015) for hillslope processes.

4.5   Summary and conclusions

A VIC model for the Thur basin was run with four different spatial resolutions (1×1 km, 5×5 km, 10×10 km, lumped) and evaluated at three different temporal resolutions (hourly, daily, monthly). The forcing was applied both uniformly and distributed over the catchment, and the drainage network for the routing routine was defined independent of the model resolution. Three objective functions were used to evaluate model performance: KGE(Q), NSE(Q) and the NSE(logQ). The model was run 3,150 times using a Hierarchical Latin Hypercube Sample and the best 1% of the runs was selected as behavioural and used for further analysis. Parameter transferability was quantified by evaluating the overlap in behavioural sets for different temporal and spatial resolutions. From the results we can draw the following conclusions:

· Both the spatial resolution and the temporal resolution of the VIC model had a significant impact on the model performance, either expressed in terms of KGE(Q), NSE(Q), or NSE(logQ). The model performance evaluated at a monthly time step consistently increased with decreasing spatial resolution, while for the daily and hourly time step no clear relation with spatial resolution could be found. Generally, the models applied with spatially distributed forcing performed better than the models applied with uniform forcing.

· The spatial resolution of the model had little impact on the parameter distribution of the behavioural sets. On the other hand, the temporal resolution significantly impacted the distribution of at least four out of seven parameters, both when applied with uniformly and distributed forcing.

· Parameters could to a large extent be transferred across the spatial resolutions, while parameter transferability over the temporal resolutions was less trivial. Parameter transferability between the hourly and daily time step was found to be feasible, but the monthly time step lead to substantially different parameter values. This is crucial information, because many studies tend to calibrate the VIC model on the monthly time step (see Chapter 7). The results of this study suggest that the output from models calibrated on a monthly time step cannot be interpreted or analysed on a daily or hourly time step. This might seem obvious, but it should be recognized that the increasing spatial resolution of large-domain land-surface models might increase the expectations concerning temporal resolution as well, as described in Chapter 7.

· We also investigated if parameters could be transferred across both the spatial and the temporal resolution simultaneously. Parameter transferability decreases when the ratio between the original and the intended spatial and/or temporal resolution increases. The ratio of temporal resolutions has a larger negative effect on

parameter transferability than the ratio of spatial resolutions. It was also shown that parameter transferability depends on the objective function. When the NSE (logQ), which tends to put more emphasize on low flows, is used as evaluation criterion, the parameter values at a monthly time step overlap much more with the daily and hourly time steps than when KGE(Q) or NSE(Q) are used as objective functions. This means that parameter transferability across temporal resolution also depends on the time scale of the process to which a particular parameter refers.

The most important result of our study is that it showed high parameter transferability across spatial resolution, even when forcing was applied in a distributed fashion. This is indicative of a substantial underestimation of the actual spatial variability represented by the VIC simulations. We did, however, construct our model according to current day standards for large-domain land-surface models, raising the point that the high spatial transferability may occur because the current generation of models have an inadequate representation of spatial variability and hydrologic connectivity. The results therefore provide strong motivation to further investigate and substantially improve the representation of spatial and temporal variability in large-domain hydrological models. Large-domain hydrological models have many applications, from water footprints (*Gleeson et al.*, 2012) and water scarcity (*Hoekstra*, 2014), to global water use (*Wada and Bierkens*, 2014) and electricity supply (*Van Vliet et al.*, 2012), but the spatial variability in the models is very likely underestimated, which increases the uncertainty in the model results.

Chapter 5

Subjective modelling decisions impact hydrological predictions

It is generally acknowledged in the environmental sciences that the choice of a computational model impacts the research results. In this study we show, with an example of hydrological modelling of floods and drought, that modelling decisions during the model configuration, beyond the model choice, also impact the model results. In our carefully designed experiment we investigated four modelling decisions in ten nested basins: the spatial resolution of the model, the spatial representation of the forcing data, the calibration period, and the performance metric. The simulation of both hydrological extremes was affected by the four modelling decisions, with differing significance and magnitude. The flood characteristics were mainly affected by the performance metric, whereas the drought characteristics were mainly affected by the calibration period. Modelling decisions during model configuration introduce subjectivity from the modeller. Multiple working hypotheses during model configuration can provide insights on the impact of such subjective modelling decisions.

5.1   Introduction

In jury sports, such as gymnastics, the jury is supposed to objectively determine the outcome of the game. In a study on home advantage for the Summer Olympic Games it was, however, shown that jury sports experience a significant home advantage, in contrast to sports which are based on objective measurements (*Balmer et al.*, 2003). It seems that the jury is actually making subjective decisions, despite their expert knowledge and all the rubrics and directives that have been drafted in order to objectify their decision.

It is generally acknowledged that models in Earth and environmental sciences are affected by several sources of uncertainty (*Oreskes et al.*, 1994). Uncertainty can, for example, stem from the randomness of natural processes (so-called aleatoric uncertainty), or from an insufficient representation of the involved processes (epistemic uncertainty). There is agreement that the model choice, basically the choice for a particular representation of the processes, affects the output and thus the results of the study, as shown by numerous model inter-comparison studies (see e.g. *Joussaume et al.* (1999) on climate modelling, *Clark et al.* (2015) on hydrological modelling, *Freni et al.* (2009) on urban storm water modelling and *Bennett et al.* (2013) on benchmarking environmental models). The modeller or expert acts as jury to determine the most appropriate model for the question at hand (*Crout et al.*, 2009), while model inter-comparison studies provide the modeller with rubrics and directives to judge the model performance in a fair way. As such, the model choice can be justified based on expert knowledge and the rubrics and directives from model inter-comparison studies. It should be noted, however, that expert knowledge is actually a mixture of opinion and knowledge (*Krueger et al.*, 2012). The opinion-part of expert knowledge introduces subjectivity in the model choice, in the same way that the gymnastics jury at the Olympic Games showed to cause home advantage: different experts could make different choices based on the same information. Furthermore, model choice is only the first decision in a sequence of decisions a modeller has to make during model configuration. The impact of those modelling decisions is currently overlooked in most, if not all, model inter-comparison studies, and an assessment of their relative importance is lacking.

Several studies in different research disciplines have shown that individual modelling decisions during model configuration can have a large impact on model results. *Cosgrove et al.* (2003), for example, showed how the length of the spin-up period affects NLDAS simulations, illustrating the large effects that chosen spin-up periods can have on land surface modelling. This study explicitly validates a spin-up modelling decision in the NLDAS project. Different model configurations are, however, not always identified as 'modelling decisions', and the subjectivity of these decisions is hardly ever acknowledged. For instance, *Ettema et al.* (2009) showed that 24% more annual precipitation over the Greenland ice sheet was obtained from a high-resolution regional climate model (RCM) compared to coarser resolution RCM output. Though it may depend on the available data or the available computational resources, the spatial resolution of the model is often a choice of the modeller. *Neal et al.* (2010) compared three parallelization methods to model 2D flood inundations, where each method - i.e. modelling decision - had particular

drawbacks. Neither *Ettema et al.* (2009) nor *Neal et al.* (2010) explicitly discuss their results as modelling decisions.

Some sources of uncertainty, for example concerning model choice (model inter-comparison studies) or "optimal" model parameters (calibration studies), have been scrutinized in detail, whereas other sources of uncertainty, such as modelling decisions during model configuration, received considerably less attention or are not recognized as such, although they might have an equally relevant impact on the model results. A possible explanation can be that the uncertainty caused by modelling decisions is introduced before the first model calculations start, making it difficult to quantify this source of uncertainty. We note, however, a slowly growing interest in describing and comprehending modelling decisions and their impact on model output, for example in the fields of water resources (*Maier and Dandy*, 2000) and hydrology (*Clark and Kavetski*, 2010; *Kavetski and Clark*, 2010; *Clark et al.*, 2011a; *Mendoza et al.*, 2015a,b; *Fenicia et al.*, 2016). The attention for this topic in the hydrologic sciences is a logical extension of the ambition to improve realism in hydrological models (e.g. *McDonnell et al.*, 2007; *Clark et al.*, 2016a), which are generally known for their conceptual nature, especially compared to other environmental sciences such as meteorology and oceanography.

In this Chapter, we argue that the choice for a particular model is only one of several modelling decisions, and we illustrate the importance of modelling decisions during model configuration through an example from hydrology. In particular, we investigate the impacts of four modelling decisions on the simulation of hydrological extremes (floods and drought). The aim is to demonstrate the impact of multiple modelling decisions on model results, and to raise awareness to recognize the uncertainty introduced by modelling decisions. Novel in this Chapter is that we systematically investigate and quantify the statistical significance of multiple modelling decisions.

5.2 Modelling decisions in hydrology

The sources of uncertainty in hydrological modelling have been a grateful inspiration for an abundance of scientific literature (e.g. *Wagener and Gupta*, 2005; *Liu and Gupta*, 2007), and have led to methods to estimate and quantify uncertainty (among others *Beven and Binley*, 1992; *Vrugt and Sadegh*, 2013). *Vrugt and Sadegh* (2013) developed a Bayesian evaluation framework that explicitly recognizes six different sources of uncertainty (parameters, forcing, initial state, model structure, output, and new states), as shown in the rectangle in Figure 5.1. Hydrological modelling, however, is surrounded by modelling decisions, as shown in the ellipse in Figure 5.1 and discussed in *Clark et al.* (2011a, 2015). These modelling decisions do not only introduce uncertainty not incorporated in the Bayesian evaluation framework in Figure 5.1, they also influence the uncertainty estimated with the developed framework. For example, they determine the prior in a Bayesian framework, or parameter uncertainty as affected by the parameter boundaries.

Many modelling decisions are relevant during the process (sometimes called '*the art*') of modelling. In this Chapter we focus on four modelling decisions for which the scientific literature provides ambiguous advice to the hydrological modeller: the spatial res-
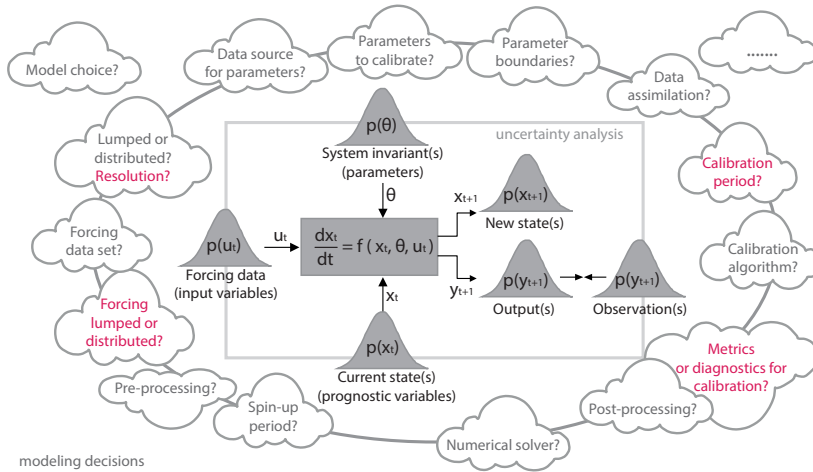
Figure 5.1: Bayesian model evaluation framework for a hydrological model, based on Figure 1 of Vrugt and Sadegh (2013), which explicitly recognizes uncertainty in parameters, forcing data, initial state, model structure, output and model state. The Bayesian framework is surrounded by decisions that a modeller has to make during model configuration. Note that the modelling decisions in this figure are non-exhaustive. The modelling decisions discussed in this Chapter are highlighted in red.

olution, the spatial representation of the forcing, the calibration period, and the performance metric. We recognize that, since we focus on only four modelling decisions, the results of our study will be impacted by many other modelling decisions that we made during model configuration. This is further discussed in Section 5.5.

Spatial resolution

An important decision that modellers make when setting up a distributed hydrological model is the choice of the spatial resolution. This choice is often bounded by the available data or the calculation time. Nowadays, both the availability of spatially distributed data at high resolution and the computational power are increasing. This has lead to the call for large-scale hyper-resolution hydrological modelling (*Wood et al.*, 2011), which aims to improve the physical realism of the models. Figure 7.3 illustrates that the spatial resolution at which the Variable Infiltration Capacity (VIC) model (*Liang et al.*, 1994) is applied in the scientific literature has increased over the years. Several studies have investigated the effect of spatial resolution (e.g. *Haddeland et al.*, 2002; *Liang et al.*, 2004; *Troy et al.*, 2008), but the reported results are ambiguous. *Troy et al.* (2008) for example, found a high sensitivity of the optimal parameter values to the spatial resolution, whereas in Chapter 4 exactly the opposite was found for the same model. In this study, we compare three different spatial resolutions, ranging from the so-called hyper-resolution as advocated by *Wood et al.* (2011) (1×1 km) to 'regional scale' hydrology (10×10 km) representing the finest test resolution of *Troy et al.* (2008), and an intermediate spatial resolution

(5×5 km).

Spatial representation of forcing

Another important choice for distributed hydrological modelling is the spatial representation of the forcing data. In this Chapter, we explore the question: do we apply the forcing in a lumped fashion over the basin, or in a distributed fashion? The choice of forcing data is in many applications a matter of choice between existing datasets, whose spatial resolution is already determined. One could select global data sets like WATCH or ERA-Interim, which are available at 0.5° or 0.25° resolution worldwide. Otherwise one needs to invest time and resources in high-resolution forcing data, e.g. obtained directly from meteorological stations or weather radars. Several studies already compared predictive accuracy and summary metrics for hydrological models fed with spatially distributed and uniform forcing data, starting with *Wilson et al.* (1979), followed by e.g. *Beven and Hornberger* (1982); *Obled et al.* (1994); *Nicótina et al.* (2008); *Zhao et al.* (2013); *Lobligeois et al.* (2014). None of the studies based on a large range of catchments (*Zhao et al.*, 2013; *Lobligeois et al.*, 2014) reported consistent results. The benefit of distributed data depends on the spatial variability of rainfall in the region and at the time scale of interest, as pointed out by *Lobligeois et al.* (2014). In the Thur, the basin where this study is conducted (see Section 2.1), topography causes a high spatial variability in rainfall. Therefore, spatially distributed forcing could potentially be of added value, although this could differ for the flood (short time scale) and drought (long time scale) event. In this Chapter, we compare uniformly applied (representing global datasets like WATCH and ERA-Interim) versus spatially distributed (representing gauge networks or radars) forcing datasets.

Calibration period

The choice of the calibration period is critical for studies where models are used to extrapolate observations in time, for example to investigate the effects of climate or land use change. Future high or low flow events may be beyond the range of historically observed events (*Wagener et al.*, 2010), suggesting that parameter values obtained from calibration on current day observations may not be the most suitable for a future climate. To mimic this effect, several studies applied a differential split sample test (*Klemeš*, 1986b), in which the calibration and validation period are significantly different in terms of precipitation and flow regime (see e.g. *Coron et al.*, 2012; *Li et al.*, 2012; *Merz et al.*, 2011). *Coron et al.* (2012) showed that the effect of the chosen calibration period on average runoff volume differed per catchment considered, and *Li et al.* (2012) concluded that some parameters are more sensitive to that choice than others. Further, *Merz et al.* (2011) found that many parameters which are assumed to be time-invariant are actually not. These considerations make it extremely difficult for a modeller to decide on an appropriate calibration period. In this Chapter we compare a high flow calibration period to a low flow calibration period (shown in Figure 5.3), thus applying the differential split sample test . Note, however, that the length of the calibration period can also impact the modelling results (see amongst others *Vaze et al.* (2010) and Chapter 3). This point is further discussed in Section 5.5.

Performance metric

The Nash-Sutcliffe Efficiency (NSE, *Nash and Sutcliffe*, 1970) is the most widely used performance metric in hydrology, even though several caveats have been identified (*Schaefli and Gupta*, 2007). Alternatives for the NSE have been proposed, for example the Kling-Gupta Efficiency (KGE, *Gupta et al.*, 2009), which consists of a correlation term, a bias term, and a measure of relative variability (see Section 2.3). Since the call for a more process-based evaluation of hydrological models (*McDonnell et al.*, 2007; *Gupta et al.*, 2008; *Clark et al.*, 2016a, and Chapter 7), hydrological signatures have become more popular as performance metric. Hydrological signatures - e.g. the slope of the flow duration curve - help in providing insights on how adequate process representations are (*Sawicz et al.*, 2011). In this study we compare the NSE(Q) and NSE(logQ), for floods and drought respectively, with the KGE(Q).

5.3   Methods

This study has been conducted on the Thur basin (1703 km^2) and its nine (nested) sub-basins of various sizes. A description of these basins and the available data is provided in Section 2.1. Because the sub-basins are nested, the ten basins are not completely independent. Five basins have upstream nested catchments: Frauenfeld, St.Gallen, Jonschwil, Halden, and the Thur (see Figure 2.4). An overview of the set-up of this study is shown in Figure 5.2. The simulation period is 1 August 2002 – 31 August 2003, characterized by three flood events in the Thur basin (August, September 2002) as well as the severe 2003 drought (June, July, August 2003) as shown in Figure 5.3 and described in Section 4.2. The rapid succession of these two contrasting hydrological events makes this period very suitable for our analysis, because only a limited simulation period is required to cover both extremes.
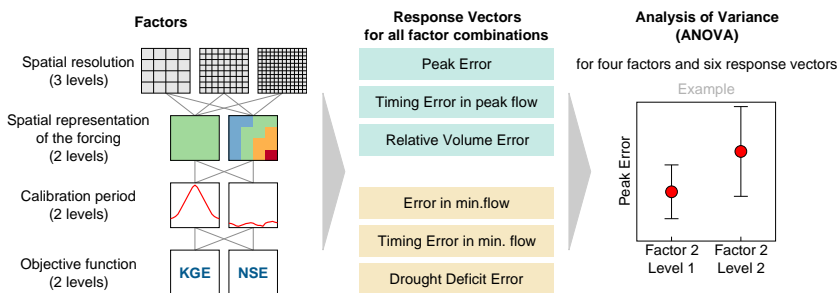
Figure 5.2: Flow chart of the methodology. Hydrological models were constructed and calibrated to represent the different factors. Subsequently, response vectors were determined by obtaining the error between modelled and observed (events indicated in Figure 5.3) for several flood and drought characteristics, for the different factor-combinations. Finally, ANOVA was conducted to test the significance of the factors on the response vector. The threshold for significance in the ANOVA was set at 0.05.
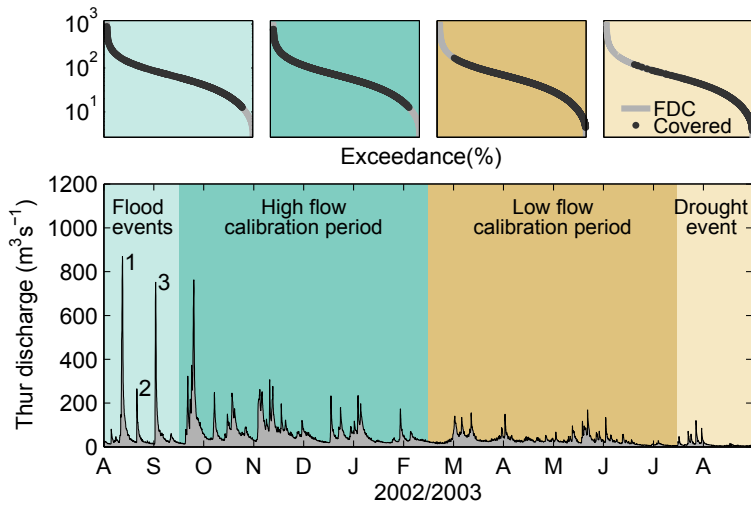
Figure 5.3: Hourly discharge in the Thur basin, with the calibration period and validation period (flood and drought events) indicated. The numbers indicate the three flood events that have been studied in more detail. Upper panels: The flow duration curve based on 39 years of hourly discharge data (light grey). The dark grey dots indicate which part of the flow duration curve is covered in the validation/calibration period.

Model and calibration

The model choice is probably one of the most important decisions a modeller makes. The goal of this study, however, was to show the impact of modelling decisions during model configuration, beyond the model choice. To illustrate this, the impact of modelling decisions has been tested for one widely used hydrological model: the Variable Infiltration Capacity model (see Section 2.2).

Three VIC models (version 4.1.2.i) were configured with different spatial resolutions (1×1 km, 5×5 km, 10×10 km). The model was run at an hourly time step (solving both the water balance and the energy balance) for the period 1 May 2002 − 31 August 2003, where the first three months were used as spin-up period. Figure 4.2 in Chapter 4 shows that three months are enough to remove the effect of initial conditions (Figure 4.2). Total runoff was routed through the channel network using the MizuRoute routine (*Mizukami et al.*, 2016) which is described in Section 2.2. Because drought events usually have a process time scale in the order of weeks or months, they do not require to be evaluated at an hourly resolution. Therefore, the model output has first been aggregated from an hourly to a daily time step to analyse the drought event. Finally, six models were configured; three different spatial resolutions, with two different spatial representations of forcing.

The models were run with a complete parameter sample (see Section 2.3). The seven sampled parameters (Table 2.4) have been applied uniformly over the catchment, where-

as the other soil- and land use parameters have been applied in a distributed fashion (separate value for each grid cell) based on data provided by the Swiss Federal Institute for Forest, Snow and Landscape Research (WSL, swisstopo license JA100118) and the Harmonized World Soil Database (*FAO et al.*, 2012), as described in Section 2.1). The use of a complete parameter sample instead of a calibration algorithm allowed us to make a fair comparison between the different model configurations, avoiding pitfalls like local minima, sensitivity to starting values, or sensitivity to the calibration algorithm.

'Calibration' (selection of behavioural parameter sets) is performed by identifying the best performing 1% (32 runs) of the 3,150 runs, based on the NSE(Q) for the flood events, and the NSE(logQ) for the drought event. Because the calibration period is one of the modelling decisions investigated, the best performing 32 runs have been determined based on both the high flow and low flow calibration period (Figure 5.3). To test the effect of performance metric selection, the best performing runs based on KGE(Q) have been determined next to the NSE(Q) and NSE(logQ) for the flood and drought event respectively. An implication of our approach – i.e. to select the best performing 1% of the runs rather than an objective-function-based threshold level – is that the selected runs can have a relatively low model performance. We do not expect that this influences our results to a significant extent, either positively or negatively, because we investigate the sensitivity of several characteristics for modelling decisions, rather than evaluating the model performance.

In total, six different model configurations were tested: three spatial resolutions and two spatial representations of forcing (uniform, distributed), which have been calibrated on two periods (a high flow calibration period from mid-September to mid-February, and a low flow calibration period from mid-February to mid-July), with two objective functions (NSE(Q) versus KGE(Q) for the flood events, and NSE(logQ) versus KGE(Q) for the drought event).

Flood and drought characteristics

To investigate the effect of subjective modelling decisions on extreme hydrological events, the error in flood and drought characteristics between observations and simulations were investigated for different model configurations (Figure 5.2). The flood characteristics have been validated for three flood events, and the drought characteristics have been validated for the drought event (Figure 5.3).

The three main characteristics of a flood event are the peak height, the timing, and the volume (*Lobligeois et al.*, 2014). For each behavioural model run, the peak error, timing error and relative volume error compared to observations were computed. The peak error (ΔQ_p) describes the difference between the maximum observed (Q^p_{obs}) and simulated (Q^p_{sim}) discharges:

$$\Delta Q_p = Q^p_{\text{sim}} - Q^p_{\text{obs}}, \tag{5.1}$$

The timing error is defined as the difference, in hours, between the observed and the modelled peak:

$$\Delta t_p = t(Q^p_{\text{sim}}) - t(Q^p_{\text{obs}}), \tag{5.2}$$

where $t(Q^p_{\text{sim}})$ is the timing of the modelled peak and $t(Q^p_{\text{obs}})$ is the timing of the observed peak. Both the peak error and the timing error are sensitive to small discharge fluctuations. The Relative Volume Error (RVE) is the relative difference in total flow volume between observed and modelled discharge:

$$\text{RVE} = \frac{\sum(Q_{\text{sim}} - Q_{\text{obs}})}{\sum Q_{\text{obs}}}, \tag{5.3}$$

where $\sum(Q_{\text{sim}} - Q_{\text{obs}})$ is the summation of the difference in the simulated (Q_{sim}) and observed (Q_{obs}) discharge over all the time steps in the flood event. To determine the beginning and the end of the flood event, an adapted version of the method of *Lobligeois et al.* (2014) is used, which is based on a threshold level Q_0. The lowest (modelled) discharge Q_{min} in four days before and four days after the observed discharge peak is determined. Then the threshold level, based on the defined Q_{min} is calculated:

$$Q_0 = \max_{t-4,t+4}(Q^p_{\text{obs}} \cdot 0.25, Q_{min} + 0.05 \cdot (Q^p_{\text{obs}} - Q_{min})). \tag{5.4}$$

The flood event starts as soon as the discharge exceeds threshold level Q_0, and ends when the discharge drops below Q_0. With this definition, the flood event cannot start earlier than four days before the observed peak discharge, and the end of the flood event cannot be later than four days after the observed peak discharge (eight days in total). The response times in our system are short (in terms of several hours up to one day for the largest basin, the Thur) and therefore four days should be sufficient to capture the flood event.

The error between simulations and observations for three specific drought characteristics has been investigated (Figure 5.2). Drought duration and deficit are the two most common characteristics for a drought event (*Van Loon et al.*, 2014). However, drought duration was difficult to determine because the drought event was occasionally interrupted by short discharge peaks. For ecology and navigation, the minimum flow is a relevant indicator, and therefore the error in minimum flow and the error in timing of the minimum flow have been determined, in addition to the error in drought deficit. All errors in drought characteristics have been computed using a daily time step. The error in minimum flow ΔQ_{min} is defined as

$$\Delta Q_{min} = Q_{min,\text{sim}} - Q_{min,\text{obs}}, \tag{5.5}$$

which is simply the difference between the lowest simulated discharge ($Q_{min,\text{sim}}$), and the lowest observed discharge ($Q_{min,\text{obs}}$) during the drought event. The error in the timing of the minimum flow Δt_{min} is defined in the same way as the timing error for the peak flow events;

$$\Delta t_{min} = t(Q^{min}_{\text{sim}}) - t(Q^{min}_{\text{obs}}). \tag{5.6}$$

Here, $t(Q^{min}_{\text{sim}})$ is the timing of the lowest simulated discharge, and $t(Q^{min}_{\text{obs}})$ is the timing of the lowest observed discharge. In order to define drought deficit, first a variable threshold level τ (*Hisdal et al.*, 2004) for drought was defined. In this study, a drought starts as the discharge drops below the lowest 10% (Q_{90}) of the observations. The threshold level was determined based on 39 years of daily observations, identifying the lowest

10% of the discharge with a moving window of 31 days (15 days before and 15 days after the date for which the threshold level is determined). Drought deficit is then defined as the integral of the deviations (d) between the threshold level and the actual discharge (*Van Loon et al.*, 2014). The deviation is defined as:

$$d(t) = \begin{cases} \tau(t) - Q(t) \text{ if } Q(t) < \tau(t) \\ 0 \qquad\quad \text{ if } Q(t) \geq \tau(t). \end{cases} \tag{5.7}$$

The total deficit D for a drought is then defined as:

$$D = \sum_{t=1}^{T} d(t) \cdot \Delta t. \tag{5.8}$$

The duration T of a drought is assumed to be the complete drought event. The error in the drought deficit is the difference between the observed deficit D_{obs} and the simulated deficit D_{sim}:

$$\Delta D = D_{\text{sim}} - D_{\text{obs}}. \tag{5.9}$$

Analysis of variance

After computing the error in flood and drought characteristics for all the behavioural runs and for the different model configurations as shown in Figure 5.2, Analysis of Variance (ANOVA) was conducted (*Ott and Longnecker*, 2010). ANOVA allows to test the hypothesis that the means of several groups (in this case, for example, the peak error obtained with three different spatial resolutions) are drawn from the same normal distribution. The ANOVA test provides the probability (from zero, zero probability, to one, certainty) of this hypothesis. Analysis of Variance was conducted for four factors (the modelling decisions), and has been applied to six response vectors (the errors in flood and drought characteristics), as shown in Figure 5.2. If the probability $p < 0.05$, the factor was assumed to have significant impact on the response vector.

The aim of the study in this Chapter is to demonstrate that modelling decisions significantly impact the simulation of hydrological extremes. This can directly be demonstrated by evaluating if any of the investigated decisions significantly (p-value lower than 0.05) impacts the error in any of the flood or drought characteristics. To investigate how persistent the impact of the modelling decision is on the flood and drought characteristics, the results of the ten investigated basins are compared. To get insight in the underlying mechanisms causing the impact of subjective modelling decisions, it was also investigated how the decisions impact the parameter distribution, using ANOVA.

5.4   Results

Flood characteristics

In this section we focus on three flood events (Figure 5.3). Figure 5.4 shows how the different model configurations impact the peak error (panel a), timing error (panel d), and relative volume error (panel g) for the three flood events in the Thur basin. Although the

magnitude of the error differs per event, the relative difference between the configurations is more or less stable over the events, except for the timing error. Figure 5.4 also shows to what extent the impact of modelling decisions on the error in characteristics of the three flood events is significant, using ANOVA (panels c, f and i).

Figure 5.4c shows that the peak error for all basins and for all three flood events is significantly affected by the spatial representation of the forcing, the calibration period and the performance metric. Resolution plays a significant role in some basins for some events. The impact of the four investigated modelling decisions on the timing error (Figure 5.4f) is less clear. The spatial representation of the forcing affects many basins for the first and second event, but for the third event the calibration period impacts more basins significantly. The performance metric significantly affects the timing error in at least six basins. The relative volume error (Figure 5.4i) is mainly impacted by the performance metric, followed by the spatial representation of the forcing and the calibration period. Spatial resolution has considerable effects on the relative volume error only in the smaller basins. The simulated flood events are mainly affected by the performance metric, followed by the calibration period and the spatial representation of the forcing, respectively. The spatial resolution plays a minor role. The flood peak is the characteristic most affected by subjective modelling decisions. A summary of the results is given in Table 5.1.

Drought characteristics

Figure 5.4 shows how the different model configurations affect the error in minimum flow (panel b), timing error (panel e) and deficit error (panel h) in the Thur basin. The results show that the calibration period has a large impact on the error in drought characteristics. Figure 5.4c shows that in all basins the calibration period significantly impacts the error in the minimum flow. The spatial representation of the forcing is important for the error in the minimum flow in four basins, and the spatial resolution only in one basin. Using the performance metric KGE(Q) as opposed to NSE(logQ) significantly affects the error in minimum flow in seven out of ten basins. For the timing error in the minimum flow we recognize the same pattern as for the timing error in the peak flow (Figure 5.4f); the impact of the modelling decisions on the timing error does not show a consistent pattern over the ten basins, although the calibration period has a significant impact in all basins. The spatial resolution and the performance metric show to be important in at least six basins. For the deficit error (Figure 5.4i), the choice of the calibration period seems to be the most important decision, with a significant impact in eight out of ten basins. The spatial representation of the forcing and the performance metric significantly affects the deficit error in five basins. Spatial resolution significantly affects the deficit error in only four basins. These results show that the drought characteristics are mainly affected by the calibration period, followed by the performance metric, the spatial resolution and the spatial representation of the forcing. The summary in Table 5.1 reveals that the timing error in the minimum flow experiences most impact from the investigated subjective modelling decisions.
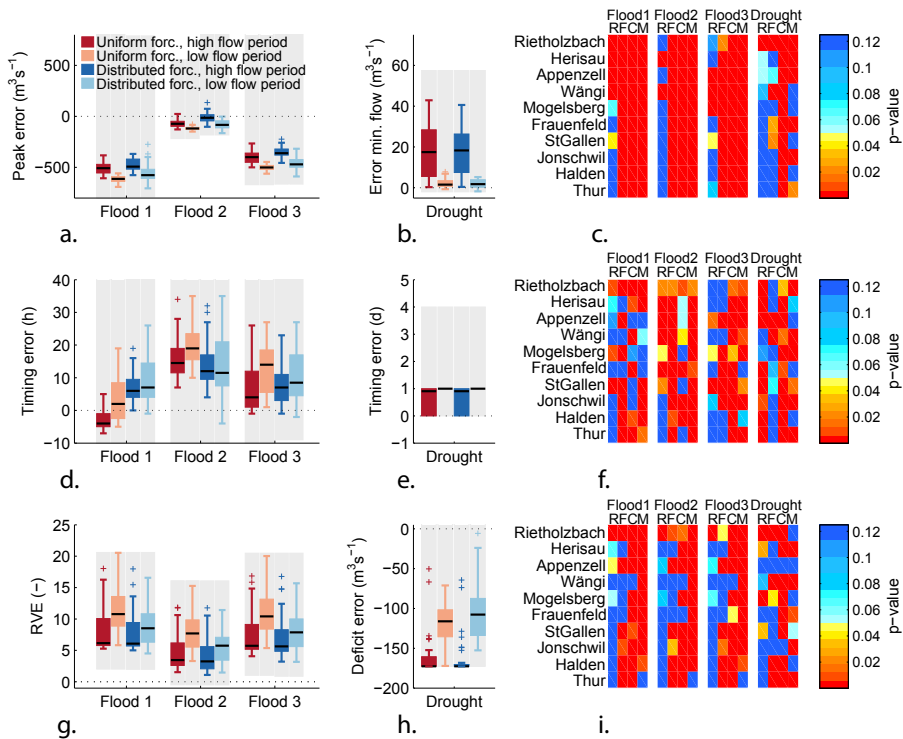
Figure 5.4: The impact of the investigated modelling decisions on the error in the three flood characteristics (peak error, timing error, Relative Volume Error) and the three drought characteristics (error in minimum flow, timing error in minimum flow, and deficit). Panel a, d, g: The distribution of the behavioural sets (best 1% expressed in NSE(Q)) for the error in flood characteristics for three flood events in the Thur (1×1 km resolution). Panels b, e, h: The distribution of the behavioural sets (best 1% expressed in NSE(logQ)) for the error in drought characteristics for the drought event in the Thur (1×1 km resolution). The dashed line indicates the optimum (no difference between modelled and observed). The grey boxes show the distribution of the complete parameter sample. For clarity, the impact of spatial resolution and performance metric are not shown. Panel c, f, i: ANOVA p-value of the impact of Resolution (R), Forcing (F), Calibration period (C), and performance Metric (M) on the error in flood and drought characteristics, for the 10 basins. The basins are ordered from small to large.

Table 5.1: *Number of catchments (out of 10 in total) for which the modelling decisions (spatial resolution R, spatial representation of forcing F, calibration period C, and performance metric M) significantly impact the error in flood and drought characteristics (p<0.05). For the error in flood characteristics, the average for the three flood events is given.*

Characteristic	R	F	C	M	Average
Peak error	3	10	10	10	8.3
Timing error	4.3	7.3	7	8	6.7
Relative volume error	1.3	6	8.7	9	6.3
Average for flood characteristics	2.9	7.8	8.6	9	
Error in min. flow	1	4	10	6	5.3
Timing error in min. flow	8	2	10	6	6.5
Deficit error	4	5	8	5	5.5
Average for drought characteristics	4.4	3.7	9.3	5.7	

Impact on parameter distribution

Table 5.2 provides an overview of the percentage of basins for which the distribution of the sampled parameters (Table 2.4) was significantly affected by any of the four modelling decisions, using ANOVA. For the flood events, spatial resolution had the lowest impact on the parameter distribution. Most basins and most parameters were affected by the calibration period, followed by the performance metric. The most affected parameters are the Depth_1, the depth of soil layer 1, and C, the velocity parameter of the routing scheme.

For the drought event, calibration period is by far the most important modelling decision that determines the parameter distribution. At a distance, this is followed by the performance metric. Especially the infiltration shape parameter, b_i, and the parameter describing the base flow relation, d_s are affected by the modelling decisions.

It is interesting to note that the affected parameters differ for flood and drought events. For the flood events, mainly the parameters impacting the response time are influenced by the modelling decisions, whereas for the drought event the infiltration and base flow parameter are mostly affected by the decisions. For drought, the calibration period is by far the most important decision, whereas spatial resolution only plays a very minor role. For floods the calibration period is most important, followed closely by the performance metric, and the spatial representation of the forcing.

5.5 Discussion

The main point of this Chapter was to demonstrate that subjective modelling decisions, beyond the model choice, affect the simulation of hydrological extremes. Some of the results can be hydrologically explained. An example is the limited impact of the spatial representation of the forcing on the error in drought characteristics. A hydrological drought

Table 5.2: Number of catchments (out of 10 in total) for which the parameters were significantly (p<0.05) affected by spatial resolution R, spatial representation of forcing F, calibration period C, and performance metric M. For the flood characteristics, NSE(Q) on an hourly time step is used as objective function. To investigate the effect of the performance metric M, NSE(Q) is compared with KGE(Q) on an hourly time step. For the drought characteristics, NSE(logQ) on a daily time step is used as objective function. The effect of the performance metric M is investigated by comparing NSE(logQ) with KGE(Q) on a daily time step.

Parameter	NSE(Q) (h)				NSE(logQ) (d)			
	R	F	C	M	R	F	C	M
b_i	2	7	9	4	0	5	8	8
d_s	4	8	5	5	3	5	8	8
d_m	2	5	7	5	0	3	10	5
$expt_1$	1	2	9	6	2	3	8	4
$Depth_1$	2	7	7	9	1	4	8	6
C	3	5	8	9	0	5	6	5
D	0	2	3	8	0	3	7	4

is mostly caused by lack of precipitation. When little or no precipitation is falling, the spatial resolution of the precipitation data is not relevant since it will remain (nearly) zero throughout the catchment. Although it is not the purpose of this paper to hydrologically explain the results, we would like to stress that some modelling decisions are more logical than others.

We also want to note that the results of this study depend on model decisions that we, as modellers, made for the experimental set-up. We only investigated the effect of four modelling decisions, although many more decisions were made while setting up our experiments. Clear examples of these decisions are the uniform application of the sampled parameters, the length of the calibration period, and the selection of the best 1% of the model runs as 'behavioural'. Furthermore, we also made important decisions on the parameters included in the sensitivity analysis, their boundaries and the sampling strategy adopted (Section 2.3).

The uniform application of the sampled parameters can decrease the effect of spatial resolution. Most likely, the spatial resolution will become a more important modelling decision when the sampled parameters are applied in a distributed fashion. A randomly distributed sample would, however, be a heavy computational burden. One potential approach is the use of spatial regularization methods, where transfer functions are formulated to relate the model parameters to physical characteristics (e.g. *Samaniego et al.*, 2010). With this method, spatially distributed parameters can be sampled by perturbing the coefficient of the transfer function. However, no pedo-transfer functions have been identified yet for the VIC model.
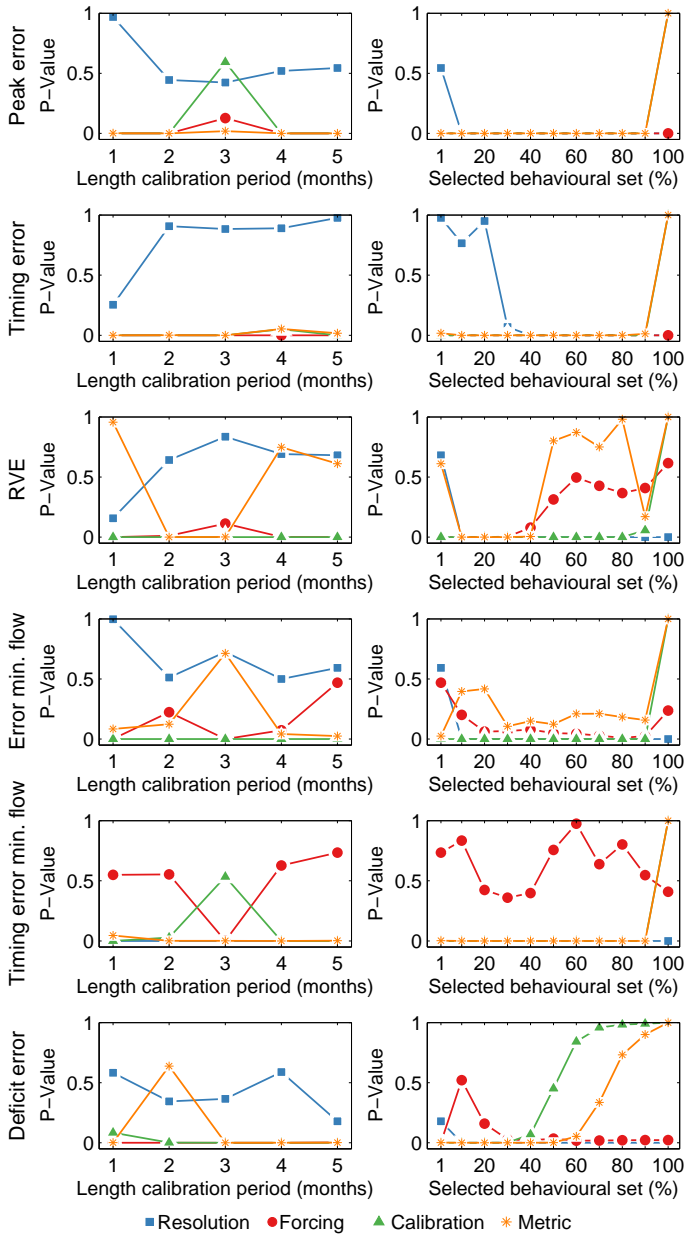
Figure 5.5: ANOVA p-value versus length of the calibration period (left panels) and size of the sample selected as 'behavioural' (right panels) for the four investigated modelling decisions. For clarity, only the results for the Thur basin are shown, and for the error in flood characteristics only the results for Flood event 1 (see Figure 5.3).

The length of the calibration period was fixed to five months. For the Rietholzbach catchment this was shown to be a sufficient period to obtain stable parameters (Chapter 3). Obviously, a shorter calibration period would lead to different model results. Therefore, the analysis as shown in Figure 5.2 has been repeated with five different calibration periods; the initial five months, each time shortened with one month up to a calibration period of one month only. The different calibration periods have been obtained by decreasing the period each time with 15 days at the beginning of the period and 15 days at the end of the period. Figure 5.5 shows that the investigated modelling decisions still have significant impact on the error in characteristics of the hydrological extremes for a shorter calibration period. Most modelling decisions that have shown to significantly impact the error in the characteristics based on a five-month calibration period, remain significant for shorter calibration periods and vice versa. *Vaze et al.* (2010) showed that model parameters are more resilient for climate change when they have been calibrated using a period of 20 years or longer and the mean annual rainfall did not change by more than 15% decrease or 20% increase. This implies that the impact of the choice of calibration period decreases when the length of the calibration period is increased. However, our experimental set-up, with a large parameter sample, did not allow a very long calibration period for computational reasons.

The selection of the best 1% of the sample as 'behavioural' is not so much a modelling decision as it is a decision in the research set-up. To investigate the effect of this choice, the analysis as shown in Figure 5.5 was repeated with 10 different sample sizes; 1% (the initial size), and 10% up to 100% of the sample, each time increasing with 10%. Figure 5.5 shows that choosing a larger sample affects the results, but in most cases it increases the significance level of the modelling decisions concerning the error in characteristics of the extremes. The figure also shows that - as expected - the choices of calibration period and performance metric approach p=1 (a very high probability that the two samples are drawn from the same distribution, i.e. no significant difference between the two samples) when the complete sample (100%) is used as 'behavioural'. In other words, when the complete parameter sample is used, it becomes unimportant which period or metric is used for calibration because essentially no calibration is performed. A remarkable result is that the spatial resolution and the spatial representation of the forcing remain important for the complete sample. They apparently impact the model output in such a manner that the complete parameter sample changes significantly.

Given the caveats discussed above, the order of importance of the investigated modelling decisions on hydrological extremes could change if other modelling decisions or experimental configurations would be adopted. Nevertheless, the conclusion that subjective modelling decisions significantly impact the simulation of hydrological extremes remains valid.

5.6   Summary and conclusion

Computational models in Earth and environmental sciences have to deal with uncertainty, which is partially augmented by subjective modelling decisions (e.g., model choice, performance metric selection). The impact of model choice on model results is generally

acknowledged, whereas the uncertainty introduced by modelling decisions during the configuration of the model is often neglected. In this Chapter we showed, with an example for hydrological extremes, that modelling decisions beyond the model choice, affect model results significantly.

We investigated four modelling decisions (spatial resolution of the model, spatial representation of the forcing data, calibration period, and performance metric) and examined the impact of these decisions on the error in three flood characteristics and three drought characteristics. Errors in flood characteristics were mainly affected by the performance metric, followed by the calibration period, the spatial representation of the forcing, and the spatial resolution of the model. The most important modelling decision for the error in drought characteristics was the calibration period, followed by the performance metric, the spatial resolution, and the spatial representation of the forcing, respectively. However, the impact of the investigated modelling decisions on hydrological extremes differed from basin to basin, even though the compared basins had much in common in terms of climate and land use. The results undeniably show that modelling decisions impact the simulation of hydrological extremes. This is undesirable, because it implies that the predicted severity of a hydrological extreme would depend on the (subjective) decisions made by the modeller

For every model in Earth and environmental sciences, several decisions have to be made. We should be aware of the large impact that these decisions have on the final model results. Preferably, uncertainty introduced by subjective decisions should be excluded, which is difficult as long as humans are involved in the modelling process. We speculate that the impact of subjective modelling decisions decreases with increasing realism in Earth and environmental models. A first step to advance process understanding is to get a better comprehension of the uncertainty introduced by adopting different modelling decisions. This paper provides a compelling example on how to conduct such an assessment for a specific hydrological application. Synthetic experiments might also have a valuable contribution to our understanding of the sensitivity of simulations to modelling decisions, although synthetic experiments in particular are dictated by modelling decisions. Further, it is critical to constrain the spectrum of options (or hypotheses) for a particular modelling decision based on the information that can be extracted from different data sources (e.g. *Gupta et al.*, 2008), which provide additional knowledge on the behaviour of environmental systems. Uncertainty associated to a particular modelling decision can be characterized through multiple working hypotheses (*Clark et al.*, 2011a), carefully selected to avoid over-confident portrayals of environmental processes.

Chapter 6

Uncertainty in the hydrological impact of climate change

While hydrological projections are of vital importance, particularly for water infrastructure design and food production, they are also prone to different sources of uncertainty. Using a multi-model set-up we investigated the uncertainty in hydrological projections for the period 2070-2100 associated with the parameters of hydrological models, hydrological model structure, and General Circulation Models (GCMs) needed to force the hydrological model, for 605 basins throughout the contiguous United States. We investigated the sign of the projected change in mean annual runoff. The hydrological model parameters influenced the sign of change in 5 to 34% of the basins, depending on the hydrological model and GCM forcing. The hydrological model structure led to uncertainty in the sign of the change in 13 to 26% of the basins, depending on GCM forcing. This uncertainty could largely be attributed to the conceptualization of snow processes in the hydrological models. In 14% of the basins, none of the hydrological models was behavioural, which could be related to catchments with high aridity and intermittent flow behaviour. In 41 to 69% of the basins, the sign of the change was uncertain due to GCM forcing, which could be attributed to disagreement among the climate models regarding the projected change in precipitation. The results demonstrate that even the sign of change in mean annual runoff is highly uncertain in the majority of the investigated basins. If we want to use hydrological projections for water management purposes, including the design of water infrastructure, we clearly need to increase our understanding of climate and hydrological processes and their feedbacks.

6.1 Introduction

A thorough understanding of the terrestrial part of the hydrological cycle is vitally important to ensure water resource management meets the many demands for water: From domestic, agricultural and industrial consumers (including hydro-power) and inland navigation, to design of infrastructure to reduce the risk of flooding and drought. How continental water resources will be affected by climate change remains, however, uncertain (*Milly et al.*, 2005; *Clark et al.*, 2016b). Hydrological projections are themselves prone to uncertainty, as they are the result of a long chain of models, with each step along the chain introducing uncertainty into the projection (*Clark et al.*, 2016b).

Due to a lack of a closed system (*Oreskes et al.*, 1994), models in the Earth sciences are all subject to conceptualization; processes are neglected, simplified, or scaled up or down in the model structure compared to the physical reality. This holds also true for hydrological models. A result of this conceptuality is that the parameters of hydrological models need to be constrained, which is critical for long-term projections exhibiting non-stationarity (*Milly et al.*, 2008; *Wagener et al.*, 2010; *Merz et al.*, 2011). The practice of using a single parameter set in hydrological models (e.g. *Teutschbein et al.*, 2015) can profoundly affect the conclusions drawn from the modelling exercise (*Andréassian et al.*, 2009). The conceptuality in hydrological models also leads to sensitivity of the results to the choice of the hydrological model; different conceptualizations lead to different results. Lastly, hydrological models require meteorological input variables, so-called forcing. Long-term hydrological projections are therefore prone to the uncertainty in General Circulation Model (GCM) projections.

A clear demonstration of uncertainty introduced by hydrological model conceptualization is the study of *Sheffield et al.* (2012) on global trends in drought. This study was conducted in response to the results of *Dai et al.* (2004), who demonstrated that the surface area experiencing severe agricultural drought has globally more than doubled since the 1970s. This conclusion was based on a water balance model where the evapotranspiration was obtained using the empirical Thornthwaite equation, which only has temperature as a variable input (*Seneviratne*, 2012). This equation neglects the fact that the causal link between drought and temperature can reverse; if no soil moisture is available for evaporation, the sensible heat flux increases (*Sheffield et al.*, 2012; *Seneviratne*, 2012). Furthermore, other evaporation drivers such as wind speed and relative humidity are neglected. *Sheffield et al.* (2012) compared the estimate of areas in agricultural drought obtained using the Thornthwaite equation with the results obtained with the more physically founded Penman-Monteith equation, which does account for wind speed and relative humidity. It was demonstrated that, when using the Penman-Monteith formulation, no strong increase in very dry areas is found. The representation of underlying (physical) principles of hydrological processes in the hydrological model can thus have a profound effect on the results and conclusion of a study.

Using a comprehensive multi-parameter multi-model multi-basin approach, we show the role that parameters, hydrological model structure, and GCM choice plays in the un-

certainty of hydrological projections. Hereby, we focus on the sign of change in long term mean annual runoff. We deliberately set the bar low because agreeing on the sign in the mean is the first step in robust projections; all other relevant variables, such as peak flows or drought (*Roudier et al.*, 2016), are even harder to project, as they are related to runoff variability rather than mean runoff. The large sample of basins employed in this study, 605, spread over the contiguous United States (CONUS, *Newman et al.*, 2015) provides the opportunity to study this uncertainty across a range of climate and catchment conditions, and to attribute the uncertainty to particular hydrological processes (*Gupta et al.*, 2014).

6.2 Methods

We apply three frequently used hydrological models, which were applied in a lumped fashion to 605 basins throughout the contiguous United States (*Newman et al.*, 2015): VIC 4.1.2h (*Liang et al.*, 1994), SAC-SMA combined with SNOW-17 (*Newman et al.*, 2015), and the TUWmodel following the structure of HBV (*Parajka et al.*, 2007). A short model description is provided in Section 2.2.

For each model, a representative set of parameters that capture the essential hydro-climatological features was identified. By using a selection of representative parameter sets, essentially a GLUE-based approach (*Beven and Binley*, 1992), we indirectly account for uncertainty in the observed forcing and validation data. For each model, a large range of parameters was sampled using a Sobol'-based Latin Hypercube sample: 17 parameters for VIC (*Demaria et al.*, 2007; *Chaney et al.*, 2015; *Mendoza et al.*, 2015c), 18 parameters for SAC (*Newman et al.*, 2015; *Lhomme*, 1997), and 15 for HBV (*Parajka et al.*, 2007; *Uhlenbrook et al.*, 1999; *Abebe et al.*, 2010). Physically realistic parameter boundaries were determined, based on the literature: see Tables 6.1, 6.2, and 6.3. Each parameter was sampled 100 times, plus 100 base runs with average parameter values. This implies that for each of the 605 basins, SAC was run 1900 times, VIC 1800 times, and HBV 1600 times. The hydrological models were forced with daily Daymet observed meteorological variables and the model output from the different parameter sets was compared with daily USGS observed discharges over a 23 year period (1985–2008). The period 1980–1985 was used as spin-up for the model. The parameter sets were considered behavioural as soon as they fulfilled a criterion that minimizes the Euclidean distance between observations and simulations for three components: the correlation, the relative variability, and the relative bias (KGE, *Gupta et al.*, 2009, see Eq.2.11). The parameter sets needed to result in a KGE of at least 0.5 with a daily time step over 23 years in order to be considered behavioural, see Figure 6.1a. If none of the parameter sets fulfilled the performance criterion, the hydrological model structure was considered non-behavioural.

The constrained hydrological models were forced with statistically downscaled and bias corrected GCM output for a historical (1980–2008) and future (2065–2100) period and run with a daily time step (Figure 6.1b). The first five years were used as spin-up period and ignored in the analysis. Five different climate models from the Coupled Model Intercomparison Project Phase 5 (CMIP5, RCP8.5) were employed: CNRM-CM5, IPSL-CM5A-MR, CCSM4, MPI-ESM-MR, and INM-CM4. The five selected GCMs represent the different families in the climate model genealogy (*Knutti et al.*, 2013). From each family the
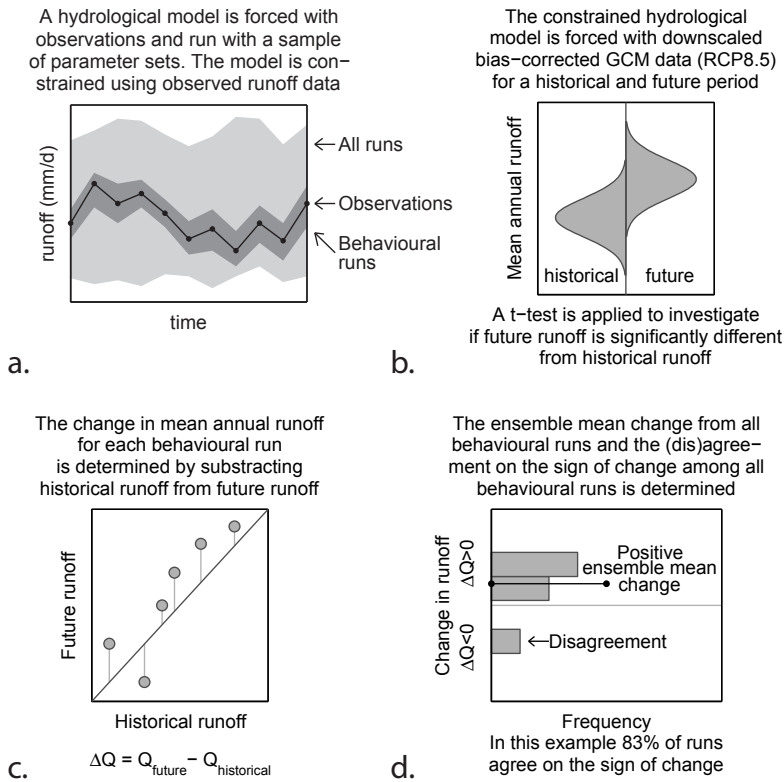
Figure 6.1: Overview of the conducted procedure. This procedure was repeated for fifteen combinations of three hydrological models and five General Circulation Models (GCMs). a) The hydrological model was run with a sample of parameter sets for the period 1985-2008. The model was forced with observed Daymet data. The size of the parameter sample differed per model, from 1600 to 1900 (see Section 6.2). Behavioural runs were identified based on the KGE(Q) on 23 years of daily discharge data. b) The constrained model, i.e. the behavioural parameter sets, were forced with GCM data for the historical period 1985-2008 and future period 2070-2100. For each run, the mean annual runoff was determined for both periods. c) The change is defined as the difference in mean annual runoff between the historical and future period, and determined per run. d) The sign of the ensemble mean change is determined, as well as the agreement among the different parameter sets on the sign of the ensemble mean change. The agreement is defined as the percentage of runs that project the same sign of change (positive change = increasing mean annual runoff, negative change = decreasing mean annual runoff) as the ensemble mean change. Finally, the sign of the ensemble mean change is compared for different combinations of hydrological models and GCMs.

member with the smallest bias in temperature and precipitation for the contiguous United States (*Sheffield et al.*, 2013) was selected.

In order to test if the projected change in mean annual runoff was significant, a t-test was applied, which compared the distribution in mean annual runoff over the behavioural parameter sets for the historical period with the distribution in mean annual runoff over the behavioural parameter sets for the future period (Figure 6.1b). The threshold in order to be qualified as significant was $p<0.05$. In order to apply the t-test, we set a pragmatic lower boundary of at least 10 parameter sets that needed to be behavioural. For VIC, SAC, and HBV, 0.5, 0.8, and 0.7% of the basins, respectively, had less than 10 but at least 1 behavioural parameter set. In these basins, the significance of the projected change could not be tested. The intention was to exclude the basins with a non-significant change from the analysis, following the approach of *Knutti and Sedláček* (2013) where agreement and significance are combined in a robustness metric. However, it turned out that none of the basins experienced a consistent non-significant change over different hydrological models and GCMs. Therefore none of the basins was excluded.

To determine the change in annual runoff for a basin, the simulated mean annual runoff over the period 1985–2008 was compared with the simulated mean annual runoff over the period 2070–2100 for each behavioural parameter set. The difference in mean annual runoff between both periods is the projected *change* (Figure 6.1c). The ensemble mean change was then determined as the mean change in annual modelled runoff for all the behavioural parameter sets (Figure 6.1d). In this study we particularly focus on the sign of the ensemble mean change (i.e., an increase or decrease in mean annual runoff).

We relate the sign of change in mean annual runoff to three sources of uncertainty: hydrological model parameters, hydrological model structure, and climate forcing. To identify the uncertainty induced by the representative parameter sets, the agreement among the representative set of parameters on the sign of the ensemble mean change was determined per basin (Figure 6.1d). To investigate the effect of hydrological model structure, the sign of the ensemble mean change projected for the three different hydrological models was compared per basin. To investigate the impact of the type of climate forcing on hydrological projections, the sign of the ensemble mean change obtained by forcing the same hydrological model with five different GCM outputs was compared.

Subsequently, the basins have been divided into three categories: basins in which the three hydrological models agree on the sign of the change, basins where the three hydrological models disagree, and basins where the three hydrological models are non-beha- vioural. For all basins, eight catchment and climate characteristics were identified. With a t-test, the climate and catchment characteristics for each category were compared to the complete sample of basins, to identify which characteristics were significantly ($p<0.05$) different for the three categories. To investigate the effect of the GCM, the sign of the ensemble mean change from the same hydrological model forced with five different GCMs, was compared. In this case, the basins were divided in two categories: basins where the model forced with different GCMs consistently agrees on the

sign of the change and basins where the model output disagrees using different GCMs. Subsequently, these categories have been related to four climate change characteristics and tested on significance ($p < 0.05$) using a t-test.

The VIC model was chosen as reference hydrological model because this model is widely applied for climate impact assessments, and the CNRM-CM5 GCM was chosen as reference GCM because it was shown that this model has the lowest bias in CONUS (*Sheffield et al.*, 2013). We do, however, also discuss the results from the other hydrological model - GCM combinations.

6.3 Results

Even when only considering behavioural hydrological model runs, there was disagreement in 11% of the basins on the sign of change when the VIC model was forced with CNRM-CM5 output. In these basins, some parameter sets lead to an increase in mean annual runoff, while other parameter sets lead to a decrease in mean annual runoff in the future. Figure 6.2a shows the spatial distribution of these basins. In the majority of the basins (59%) there is unanimous agreement on the sign of the change. In 29% of the basins no representative set of parameters could be identified, i.e., the hydrological model could not capture the hydrological behaviour of the basin and is therefore qualified as non-behavioural. In 1% of the basins the change was not significant, or less than 10 parameters were identified as representative (see Section 6.2). The percentage of basins in which there is unanimous agreement on the sign of change depends on the employed hydrological model and the GCM (see Appendix Figure A.1). On average, VIC leads to agreement in most basins (55% averaged over five GCMs), followed by SAC (46%), and HBV (43%). HBV has the largest fraction of basins where the model was non-behavioural (44%), compared to VIC (29%) and SAC (22%).

The sign of the change in mean annual runoff is also affected by the hydrological model structure. The model structure leads to disagreement in 26% of the basins when the hydrological models are forced with CNRM-CM5 output. Figure 6.2b also shows the basins in which there is agreement among the three employed hydrological models on the sign of the change in mean annual runoff. In 46% of the basins, three or two (if one model was non-behavioural) hydrological models agreed on the sign of the change. In 14% of the basins two hydrological models were non-behavioural, such that a mutual model comparison was impossible. In the remaining 14% of the basins, none of the employed hydrological models was behavioural. The agreement among the hydrological models varied for different GCMs (see Appendix Figure A.2). With CCSM4, the highest agreement among hydrological models was established (in 59% of the basins), with INM-CM4 the lowest agreement (45%).

The structure in the spatial distribution of the agreement among the hydrological models, as demonstrated in Figure 6.2b, suggests a link with catchment and climate characteristics. We compared several characteristics for three different agreement categories (agreement, dark and light blue dots in Figure 6.2b; no agreement, orange and red dots in Figure 6.2b; three models non-behavioural, black dots in Figure 6.2b). Figure 6.3 shows
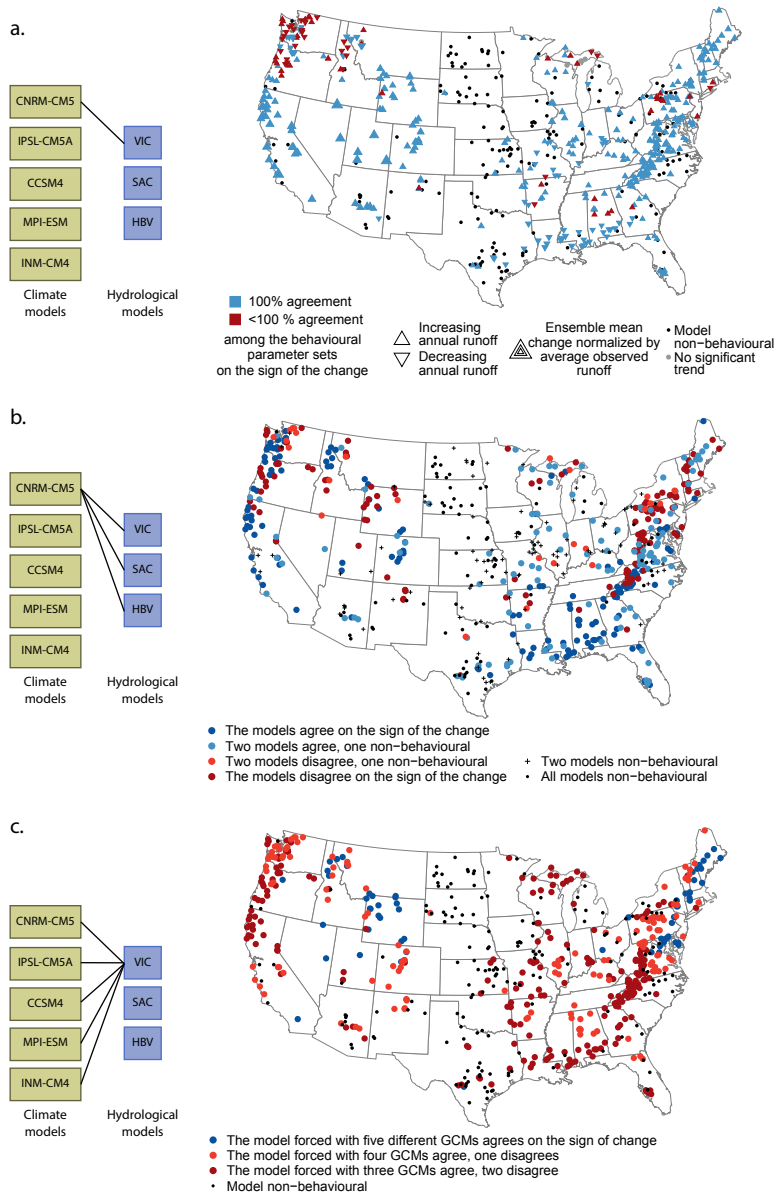
Figure 6.2: *Distribution of uncertainty in the sign of change over the contiguous United States. a) The agreement among the different model runs (representing different behavioural parameter sets) of the VIC model on the sign of the ensemble mean change in mean annual runoff. The direction of the triangle-marker shows the sign of the ensemble mean change, the size of the marker indicates the relative projected change. b) Agreement among the three different hydrological models (all forced with CNRM-CM5) on the sign of the ensemble mean change. c) Agreement on the sign of the ensemble mean change when the same hydrological model (VIC) is forced with data from five different GCMs.*

the spatial distribution and summarizes the characteristics of the basins within each category for the models forced with CNRM-CM5. In the next section, we only discuss the characteristics which were consistently significantly different when the three hydrological models were forced with different GCMs.

The basins in which the three hydrological models agree on the sign of the change (Figure 6.3a) experience a significantly lower projected temperature change compared to all other basins. Furthermore, these basins have on average a lower aridity, less or no dry periods, and a lower snow-day ratio. A logical explanation, namely that the basins in which the models agree experience a significantly larger change in precipitation as demonstrated in Figure 6.3b, was found not to be consistent among the five GCMs. The basins in which the hydrological models disagree on the sign of the change (Figure 6.3c) are characterized by a consistently higher slope and elevation, and related to that, a significantly higher snow-day ratio (Figure 6.3d). Disagreement among the hydrological models can thus be attributed to the conceptualization of snow accumulation and melt processes. The basins in which none of the hydrological models was able to capture the hydrological behaviour (Figure 6.3e) have a significantly higher aridity, and intermittent stream flow behaviour (no-flow periods). This is related to basins with a larger area and a lower slope (Figure 6.3f). These results imply that all models have difficulty in mimicking dry conditions, where the interplay between soil moisture and evapotranspiration becomes important (*Seneviratne et al.*, 2010). For the basins in Figure 6.3e, a significantly lower change in precipitation is projected, consistent among the five GCMs.

Uncertainty in the sign of change introduced by the climate forcing shows to have impact in the majority of the basins: In 60% of the basins different GCM forcing leads to a different sign of change when the VIC model is employed. Figure 6.2c shows that only in 11% of the basins, the sign of the change is consistent when the same hydrological model (VIC) is forced with different climate model outputs. In the remaining 29% of the basins, the hydrological model was non-behavioural. When HBV is applied with five GCMs, 16% of the basins show a consistent sign of change, for SAC only 8% (see Appendix Figure A.3).

Climate change characteristics can explain the (dis)agreement on the sign of the change when the hydrological model is forced with five different GCMs. The spatial distribution and the related climate change characteristics of two different categories (agreement, dark blue dots in Figure 6.2c; disagreement, orange and red dots in Figure 6.2c) are shown in Figure 6.4 for the VIC model. In the next section, we only discuss the characteristics which were consistently significantly different among the three different hydrological models when forced with five GCMs.

The basins with agreement on the sign of the change (Figure 6.4a) are characterized by a significantly lower standard deviation in the projected change in precipitation (Figure 6.4b), i.e., the GCMs agree more on the projected change in precipitation. The basins in which the models disagree on the sign of the change (Figure 6.4c) have a significantly larger change in precipitation, a smaller change in temperature, a smaller standard deviation in temperature among the five GCMs, but a larger standard deviation in precipi-
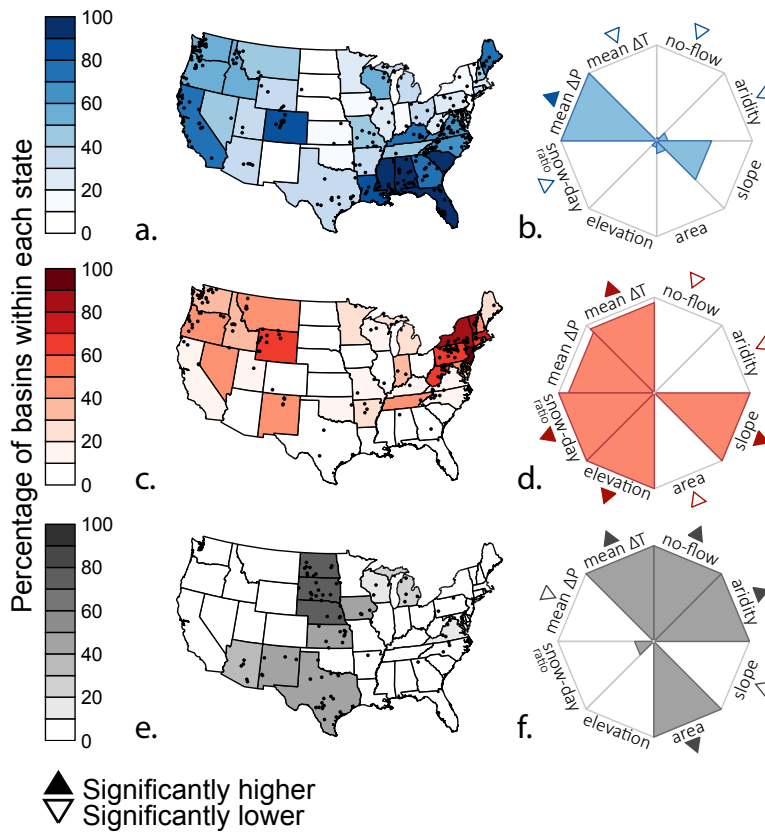
Figure 6.3: *Distribution of the (dis)agreement in the sign of the ensemble mean change in mean annual runoff among the three hydrological models, linked to catchment and climate characteristics. The characteristics have been standardized from 0 to 1, where 0 represents the lowest value of the characteristic for the three displayed groups, and 1 the highest value of the characteristics for the three displayed groups. The hydrological models were forced with CNRM-CM5 data. a) Spatial distribution of the basins in which the three hydrological models agree on the sign of the ensemble mean change. b) Standardized catchment and climate characteristics for the basins in panel a. c) Spatial distribution of the basins in which the three hydrological models disagree on the sign of the ensemble mean change. d) Standardized catchment and climate characteristics for the basins in panel c. e) Spatial distribution of the basins in which the hydrological models were non-behavioural. f) Standardized catchment and climate characteristics for the basins in panel e.*
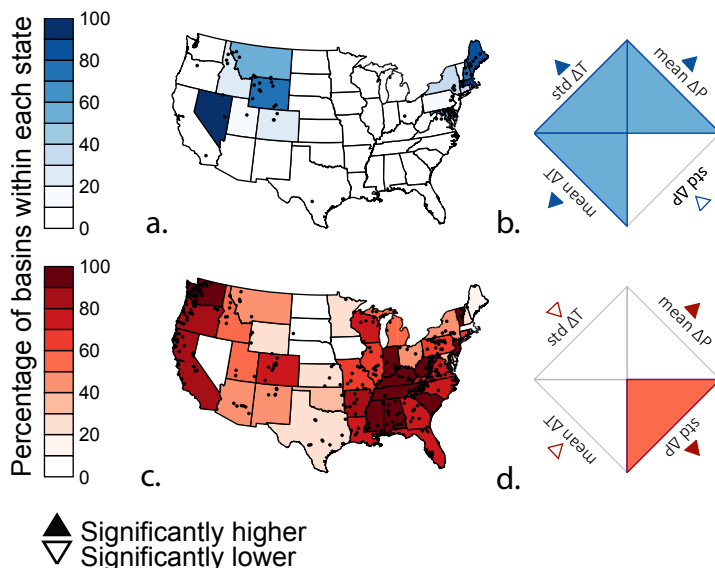
Figure 6.4: Distribution of the (dis)agreement in the sign of the ensemble mean change in mean annual runoff when the same hydrological model is forced with data from five different GCMs, linked to projected climate change characteristics. The climate change characteristics have been standardized from 0 to 1, where 0 represents the lowest value of the characteristic among the two displayed groups, and 1 the highest value of the characteristic among the two displayed groups. The employed hydrological model is VIC. a) Spatial distribution of the basins in which the five outputs from the same hydrological model forced with five different GCMs agrees on the sign of the ensemble mean change. b) Standardized climate characteristics for the basins in panel a. c) Spatial distribution of the basins in which the hydrological model forced with five different GCMs disagrees on the sign of the change. d) Standardized climate characteristics for the basins in panel c.

tation among the five GCMs (Figure 6.4d). It may seem illogical that both the basins with agreement (Figure 6.4a) and the basins with disagreement (Figure 6.4c) have a significantly higher mean change in projected precipitation (Figure 6.4b,d), but this is because the basins in each category have been compared to all other basins, including the basins in which the hydrological model was non-behavioural. The results in Figure 6.4 show that precipitation is the driving force for hydrological models, and that the disagreement among GCMs regarding the change in precipitation introduces uncertainty in the hydrological projection, more than temperature does.

6.4 Discussion

Uncertainty in the parameters of the hydrological models affected the sign of change in 5 to 34% of the basins, dependent on the employed hydrological model and climate forcing. A more stringent criterion to identify representative parameter sets could decrease

this uncertainty, but at the same time increases the number of basins in which the models are non-behavioural. Another approach would be to constrain the hydrological models on internal consistency, by using observations from different hydrological states and fluxes (*Koster et al.*, 2010; *Rakovec et al.*, 2016).

Furthermore, this study revealed that hydrological model structure creates uncertainty in the sign of the change in 13 to 26% of the basins, dependent on the GCM forcing. The representation of snow processes in hydrological models is currently inadequate. The exact mechanisms behind the impact on runoff of a precipitation shift from snow towards rain are not yet well understood (*Berghuijs et al.*, 2014). In 14% of the basins all three hydrological models were non-behavioural, which was related to arid conditions and intermittent stream flow behaviour. Our lack of understanding in these processes is of concern particularly as aridity is expected to increase in the future (*Berg et al.*, 2016). Improved conceptualization of these processes is needed, which will probably entail scrutinizing soil moisture - evapotranspiration feedbacks in detail (*Roderick et al.*, 2015).

The largest source of uncertainty were the GCMs, influencing the sign of change in 41 to 69% of the basins dependent on the employed hydrological model. The large impact of GCMs on the uncertainty is confirmed in many other studies, such as *Addor et al.* (2014). The uncertainty could be attributed to disagreement among the GCMs regarding the projected precipitation. In order to improve the agreement on the sign of the change in mean annual runoff, the disagreement in the projected precipitation among the GCMs needs to be reduced. This is, however, not a trivial task (*Knutti and Sedláček*, 2013).

One could expect that uncertainty in the parameters is closely related to uncertainty in the hydrological model structure. This is, however, not the case. Figure 6.5 provides an overview of the investigated sources of uncertainty and their combined spatial distribution. Depending on which combination of hydrological model and GCM is employed, 0 to only 5% of the basins experience both uncertainty from the parameters and from the hydrological model structure, and not from the climate model. The combination hydrological model - GCM is more frequent (3 - 14%) and the combination parameter sets and GCM is most frequent (3 - 19%). In 1 to 16% of the basins, all three factors lead to uncertainty. It should be noted that several sources of uncertainty have not been considered in this study, such as the statistical downscaling technique of the GCM output, and the emission scenario. It has been shown that these factors also introduce uncertainty in hydrological projections (*Gutmann et al.*, 2014; *Addor et al.*, 2014), and therefore the uncertainty in this study is likely even underestimated.

6.5 Conclusion

We demonstrated that robustness in hydrological projections is controlled by the parameters of hydrological models, the hydrological model structure, and the GCM forcing. This being so, hydrological models should always be used with caution in climate change impact studies (e.g. *Van Vliet et al.*, 2012; *Roudier et al.*, 2016). In most basins, even the sign of the change in mean annual runoff is highly uncertain. We found that this uncertainty is associated with snow and soil moisture evaporation processes dictated by the hydro-
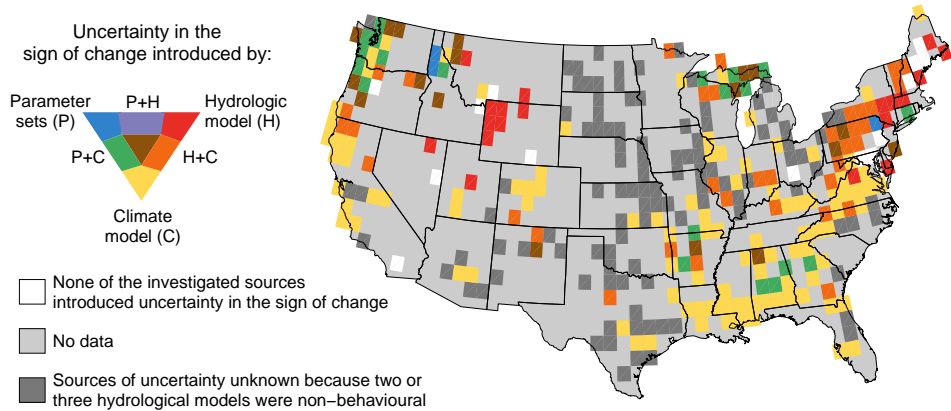
Figure 6.5: Distribution of the combined investigated sources of uncertainty. Spatial coverage is obtained by determining a grid-based maximum likelihood: For each basin, the relevant sources of uncertainty have been determined where CNRM-CM5 was used as reference model when the three hydrological models were compared, and VIC was used as reference when the five climate models were compared. Subsequently, for each grid cell with a size of $1 \times 1°$, the most frequent sources-of-uncertainty combination from the basins within that grid cell have been determined.

logical model, processes that have been recognized before as being essential for the energy balance (*Lahoz and De Lannoy*, 2014), and with (dis)agreement among GCMs regarding the projected change in precipitation. Our study included many of the headwaters of reservoirs that are currently used to store water for domestic supplies and irrigation, for example the Delaware River that feeds the Pepacton reservoir, the main drinking water reservoir for New York City. The compelling social relevance of these basins provides a strong motivation to further enhance our understanding of climate and hydrological processes and their interplay.

Table 6.1: *Selected parameters and their boundaries for the VIC model (see Section 2.2). Parameter 1 to 7 were shown to be the most sensitive parameters based on Demaria et al. (2007); Chaney et al. (2015) and Chapter 2. Parameter 8 to 14 were selected for their impact on snow and/or evapotranspiration processes. Parameter 15 - 17 are usually hard-coded in the VIC model, but were shown to be highly sensitive in Mendoza et al. (2015c) and are therefore included in the sampling. LB = lower boundary, UB = upper boundary.*

	Name	Unit	LB	UB	Description
1	b_i	-	10^{-5}	0.4	Variable infiltration curve parameter
2	d_s	-	10^{-4}	1.0	Fraction of d_m where non-linear base flow starts
3	d_m	mm d^{-1}	0.1	50	Maximum base flow
4	w_s	-	0.2	1.0	Fraction of maximum soil moisture where non-linear base flow starts
5	expt$_1$	-	4.0	30	Exponent of the Brooks-Corey relation layer 1
6	Depth$_1$	m	0.1	3.0	Depth of soil layer 1
7	Depth$_2$	m	0.1	3.0	Depth of soil layer 2
8	Ts$_{max}$	°C	0.0	3.0	Max temperature where snowfall can occur
9	Ts$_{min}$	°C	Ts$_{max}$-0.01	Ts$_{max}$-3.0	Min temperature where rainfall can occur
10	SR	-	5^{-5}	0.5	Surface roughness of the snow pack
11	RZT$_1$	-	0.5	2	Multiplication factor for rootzone thickness layer 1
12	RZT$_2$	-	0.5	2	Multiplication factor for rootzone thickness layer 2
13	RZT$_3$	-	0.5	2	Multiplication factor for rootzone thickness layer 3
14	R$_{min}$	-	0.1	10	Multiplication factor for minimum stomatal resistance of the vegetation
15	newalb	-	0.7	0.99	New snow albedo
16	albaa	-	0.88	0.99	Base in snow albedo function for accumulation
17	albtha	-	0.66	0.98	Base in snow albedo function for melt

Table 6.2: Selected parameters and their boundaries for the SAC model (see Section 2.2). The parameter boundaries are based on Newman et al. (2015), the Priestley-Taylor parameter (number 18) has been adapted based on Lhomme (1997). LB = lower boundary, UB = upper boundary.

	Name	Unit	LB	UB	Description
1	MFAX	mm $°C^{-1}$ $6h^{-1}$	0.8	3.0	Max melt factor
2	MFMIN	mm $°C^{-1}$ $6h^{-1}$	0.01	0.79	Min melt factor
3	UADJ	km $6h^{-1}$	0.01	0.40	Wind adjustment factor for rain on snow
4	SI	mm	1.0	3500	snow water equivalent for 100% snow area
5	SCF	-	0.1	5.0	Undercatch correction factor
6	PXTEMP	$°C$	-3.0	3.0	Temperature for rain/snow transition
7	UZTWM	mm	1.0	800	Upper zone max storage of tension water
8	UZFWM	mm	1.0	800	Upper zone max storage of free water
9	LZTWM	mm	1.0	800	Lower zone max storage of tension water
10	LZFPM	mm	1.0	800	Lower zone max storage of free water
11	LZFSM	mm	1.0	1000	Lower zone max storage of secondary free water
12	UZK	d^{-1}	0.1	0.7	Upper zone free water lateral depletion rate
13	LZPK	d^{-1}	1^{-5}	0.025	Lower zone primary free water depletion rate
14	LZSK	d^{-1}	1^{-3}	0.25	Lower zone secondary free water depletion rate
15	ZPERC	-	1.0	250	Max percolation rate
16	REXP	-	0.0	6.0	Exponent of the percolation equation
17	PFREE	-	0.0	1.0	Fraction percolating from the upper to the lower zone
18	P-T	-	1.0	1.74	Priestley-Taylor coefficient

Table 6.3: *Selected parameters and their boundaries for the HBV model (see Section 2.2). The selected parameters are based on Parajka et al. (2007), the parameter boundaries have been widened based on Uhlenbrook et al. (1999) and Abebe et al. (2010). The Priestley-Taylor parameter (number 15) is based on Lhomme (1997). LB = lower boundary, UB = upper boundary.*

	Name	Unit	LB	UB	Description
1	SCF	-	0.1	5.0	Snow correction factor
2	DDF	mm $°C^{-1}$ day^{-1}	0.04	12	Degree day factor
3	Tr	$°C$	0.0	3.0	Temperature above which precipitation is rain
4	Ts	$°C$	Tr-0.01	Tr-3	Temperature below which precipitation is snow
5	Tm	$°C$	-3.0	3.0	Temperature where melt starts
6	LP	-	0.0	1.0	Evaporation reduction threshold
7	FC	mm	0.0	2000	Max soil moisture storage
8	BETA	-	0.0	20	Non-linear shape coefficient
9	K0	d	0.0	2.0	Storage coefficient of very fast response
10	K1	d	2.0	30	Storage coefficient of fast response
11	K2	d	30	250	Storage coefficient of slow response
12	L	mm	0.0	100	Reservoir threshold
13	PERC	mm d^{-1}	0.0	100	Percolation rate
14	BMAX	d	0.0	30	Max base flow of low flows
15	P-T	-	1.0	1.74	Priestley-Taylor coefficient

Chapter 7

Model evaluation: spatio-temporal scales

A meta-analysis on 192 peer-reviewed articles reporting on applications of the Variable Infiltration Capacity (VIC) model in a distributed way reveals that the spatial resolution at which the model is applied has increased over the years, while the calibration and validation time interval has remained unchanged. We argue that the calibration and validation time interval should keep pace with the increase in spatial resolution in order to resolve the processes that are relevant at the applied spatial resolution. We identified six time concepts in hydrological models, which all impact the model results and conclusions. Process-based model evaluation is particularly relevant when models are applied at hyper-resolution, where stakeholders expect credible results both at a high spatial and temporal resolution.

7.1 Introduction

One of the famous paradoxes of the Greek philosopher Zeno of Elea (∼450 BC) concerns a shot arrow (*Fearn*, 2001): *If one shoots an arrow, and cuts its motion into such small time steps that at every step the arrow is standing still, the arrow is motionless, because a concatenation of non-moving pieces cannot create motion*. Only ages later, this reasoning could be refuted by the invention of integral and differential calculus by Newton and Leibniz (*Stillwell*, 1989), accepting infinitely small rates of change. Motion is a change of location over time, thus motion links time and space.

In hydrology, it is essential to understand and predict the motion of water within the Earth system, which implies that both space and time have to be considered. In hydrological models space can be accounted for by using distributed (spatially explicit) models, where space is *'cut in small pieces'*, to paraphrase Zeno. Different types of distributed hydrological models exist; *Todini* (1988) distinguished roughly two different classes. The first class consists of distributed differential models. These models explicitly simulate lateral fluxes by means of differential equations. The second class are the distributed integral models which consist of one-dimensional columns and ignore lateral fluxes between the columns (lateral fluxes can be accounted for with an extra routing scheme, as for example applied in Chapter 4, although this does not allow for lateral re-distribution). These models have a wide application in land-surface modelling (*Clark et al.*, 2015). In this discussion we focus on the latter.

The constant development in computational power, the increased understanding of physical processes, and the increased availability of high spatial resolution hydrological information stimulated the development of increasingly complex and distributed hydrological models (*Boyle et al.*, 2001; *Liu and Gupta*, 2007). Increasing the spatial resolution of Global Hydrological Models (GHMs) has been labelled as one of the current 'Grand Challenges' in hydrology by *Wood et al.* (2011) and *Bierkens et al.* (2015), who call for global modelling at the so-called spatial hyper-resolution (∼1 km and smaller). Arguably, there is a growing societal need for hydrological information at the (sub-)km scale. Whereas model products at the $1°$ or $0.5°$ resolution may provide relevant information for policy makers at the (inter)national level, hyper-resolution results will become relevant for local water managers or even individual farmers (see e.g. *Bastiaanssen et al.*, 2007). The challenge is not to simply provide information based on a model with default parameters, but to provide credible information which matches the actual situation in the field at a temporal resolution which is consistent with the spatial resolution of the model. The temporal and spatial scales are linked through the characteristic speed (including both velocity and celerity as discussed in *McDonnell and Beven*, 2014) of the involved hydrological processes (*Blöschl and Sivapalan*, 1995), the so-called *process scale*. Figure 7.1 shows that there is a general tendency for the temporal process scale to decrease with the spatial process scale, although there is quite a broad bandwidth and local changes might occur stepwise. Policy makers might be able to deal with model products at a monthly resolution, whereas resource managers and farmers expect, at the spatial hyper-resolution, credible model products with a daily or hourly resolution.
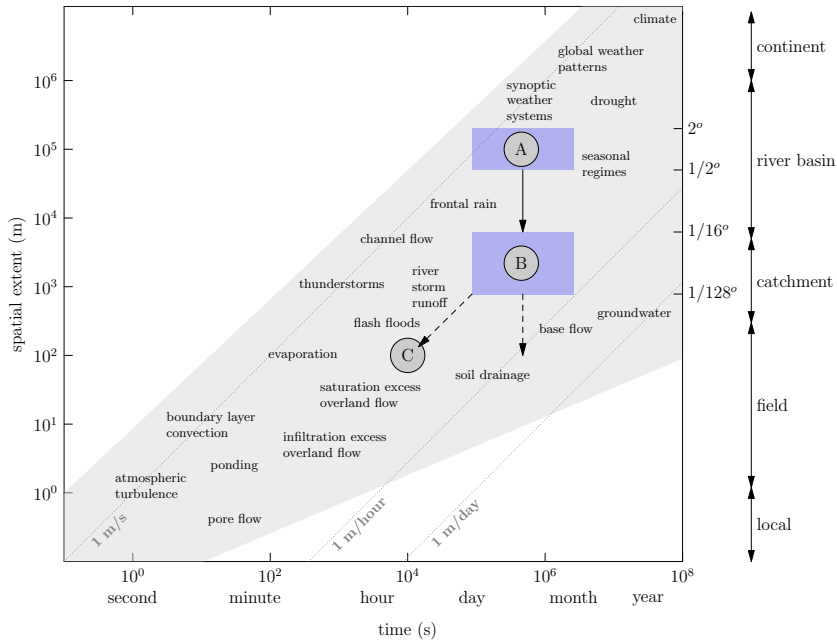
Figure 7.1: The time and space scales of several hydro-meteorological processes. Adapted from Brutsaert (2005) and Blöschl and Sivapalan (1995), who based it on Orlanski (1975); Dunne (1978); Fortak (1982); Anderson and Burt (1990). The blue areas indicate the temporal and spatial resolution at which the VIC model has been applied, when it was initially developed (A) and nowadays (B). The dashed arrow pointing downwards shows the ambitions of spatial hyper-resolution modelling, whereas the dashed arrow pointing towards (C) shows the temporal and spatial resolution of hyper-resolution modelling if it would follow the direction of characteristic velocity of hydro-meteorological processes.

Although increasing the resolution of hydrological models is claimed to provide the opportunity to improve physical process representation in hydrological models (*Bierkens et al.*, 2015; *Bierkens*, 2015), as discussed in Chapter 4, almost every hydrological model requires calibration of the model parameters (*Beven*, 2012). Models can contain conceptual parameters, which have no directly measurable physical meaning and thus need calibration. In addition, the measurement scale of parameters which do have a physical meaning often differs from the model scale, making calibration necessary to determine the effective parameter values to account for sub-grid variability (*Kim and Stricker*, 1996). *Beven and Cloke* (2012) responded to the hyper-resolution challenge by emphasizing that the focus of hydrological modelling should be on determining and accounting for epistemic uncertainty and appropriate parameterizations at different spatial resolutions, rather than on maximizing the spatial resolution. Increasing the spatial resolution of the model (towards hyper-resolution) is not a solution to sub-grid variability, since many of the relevant processes take place on even smaller scales (*Wood et al.*, 1992; *Kim and Stricker*,

1996; *Arora et al.*, 2001; *Montaldo and Albertson*, 2003; *Beven and Cloke*, 2012; *Clark et al.*, 2015). Hence, despite their increasing spatial resolution, also GHMs require calibration in order to obtain effective parameters, and validation to determine model credibility. Even if a correct physical representation of hydrological processes is impossible, the goal of the model should be to mimic realism and hydrological processes as closely as possible (*Wagener and Gupta*, 2005; *Kirchner*, 2006; *McDonnell et al.*, 2007). This implies that the models should be subject to a process-based calibration and validation procedure (*Gupta et al.*, 1998, 2008; *Clark et al.*, 2011b). Since different hydrological processes dominate at different scales (Figure 7.1), the temporal and spatial scale are linked. Because the spatial resolution of GHMs is currently being increased to meet societal needs (*Wood et al.*, 2011), the temporal resolution should decrease accordingly to meet these needs. This should be reflected in the calibration and validation time interval of the model, in order to guarantee model credibility at the required temporal and spatial resolution.

7.2   Time scales

A short review of scientific literature about scaling issues provides the impression that the focus has mostly been on the spatial scale and/or resolution rather than on its temporal counterpart (*Klemeš*, 1983; *Dooge*, 1986; *Gupta et al.*, 1986; *Dooge*, 1988; *Feddes*, 1995; *Kalma and Sivapalan*, 1995; *Sposito*, 1998; *Beven*, 1995; *Bierkens et al.*, 2000; *Gentine et al.*, 2012). Many concepts have been developed to describe representative areas and volumes (*Gray et al.*, 1993). In soil physics, the Representative Elementary Volume (REV) is an often used concept which describes the volume for which a measurement can be considered representative (*Whitaker*, 1999). *Wood et al.* (1988) explored a similar concept with applications in hydrology, namely the Representative Elementary Area (REA), the critical area at which the pattern of small-scale heterogeneity becomes unimportant. *Reggiani et al.* (1998) proposed the Representative Elementary Watershed (REW), allowing closure of the balance equations averaged over time and space. Similar concepts which statistically integrate temporal variations have not been reported in the literature. The lack of attention for the temporal scale is remarkable, because hydrological states and fluxes are mostly studied as a function of time. As an illustration of the lack of attention for the aspects of temporal scale, it should be noted that in the recent papers by *Wood et al.* (2011) and *Bierkens et al.* (2015) on spatial hyper-resolution modelling, the temporal resolution of these models is referred to only once. One of the reasons why the development of a Representative Elementary Timestep (RET) is more complex, is that several different time concepts play a role in hydrological modelling.

As a guideline and first step for the discussion on time dimensions in hydrological models, we identify six time concepts which in practice are often mixed up and misinterpreted. A distinction is made between 'scale', which is defined as a continuous variable, 'resolution', defined as discrete variable being a model property, and 'time interval', which is a discrete variable independent of the used model. The six concepts are:

1. The process time scale

2. The input resolution
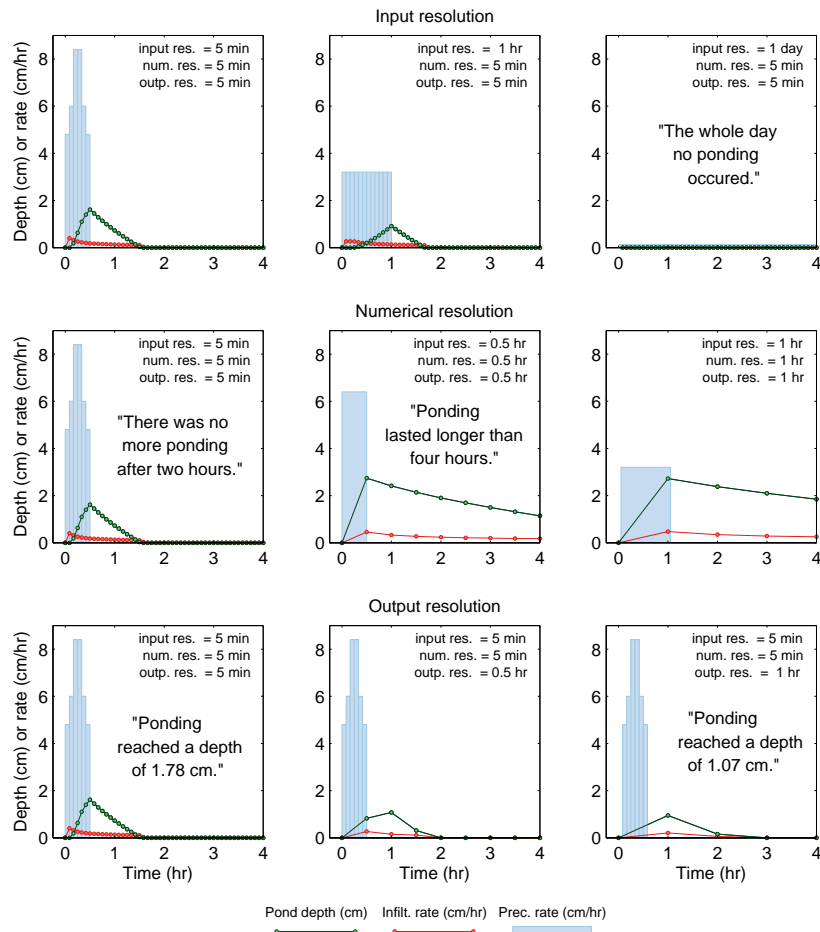
Figure 7.2: Application of the Green-Ampt infiltration scheme for different input resolutions (upper row), different numerical resolutions (middle row), and different output resolutions (lower row). For each set-up, the model was fed with the same extreme precipitation event of 32 mm of rain in 30 minutes (4 mm in the first 5 minutes, 5 mm in minutes 5-10, 7 mm in minutes 10-20, 5 mm in minutes 20-25 and 4 mm in minutes 25-30). The model parameters have been kept constant; saturated hydrologic conductivity K_s=0.044 cm/hr, initial volumetric soil moisture content θ_i=0.1, saturated volumetric soil moisture content θ_s=0.5, matric pressure at wetting front Ψ=22.4 cm. Each of the three time concepts impacts the conclusions that are drawn from the model results, which shows that calibration and validation at the appropriate time interval are essential to resolve the processes taking place.

3. The numerical resolution ("time step")
4. The output resolution ("temporal resolution")
5. The calibration/validation time interval
6. The interpretation time interval

Firstly, the *process time scale* is defined, as the characteristic time scale of the hydrological process considered. This is the typical time period over which the process takes place. Infiltration excess overland flow, for instance, has a relatively short time scale, whereas regional groundwater flow has a longer time scale. The end-user determines which process is most relevant in the modelling procedure.

Secondly, the temporal resolution of the input data or *input resolution* is relevant for the modelled process. The input resolution of the forcing data can differ from the output resolution of the model, and this can impact the results of the model. An example is given in the upper panels of Figure 7.2, showing an application of the Green-Ampt (*Green and Ampt*, 1911) infiltration model.

The *numerical resolution* (or the "time step") of the model is the time interval over which the model calculates the states and the fluxes internally. A model can only deterministically resolve a process if the numerical resolution is higher than the characteristic time scale of the process. The panels in the second row of Figure 7.2 show how the numerical resolution impacts model output for the process of ponding, which leads to different conclusions about ponding, based on the model output.

The *output resolution* (often referred to as simply "temporal resolution") is the time interval at which the model output yields the states and fluxes. This time interval can be equal to the numerical resolution of the model, or aggregated from the numerical resolution. The modelled process can only be identified if the output time interval is shorter than the characteristic time scale of the process, which is shown in the lower panels of Figure 7.2.

The *calibration and validation time interval* of the model is defined here as the time interval at which the model output is being confronted with observations. Calibration and validation of the model output can be conducted at another time interval than the output resolution, by aggregating the model output. Calibration and validation should be performed at a time interval smaller than or equal to the time scale of the process that is relevant for the end-user.

Finally, the *interpretation time interval* is defined as the time interval at which the model output is eventually analysed or interpreted. This can be equal to the calibration time interval, or the model output can be further aggregated resulting in a larger interpretation time interval (e.g. from daily to monthly). Since the model has not been validated or calibrated on time intervals smaller than the calibration time interval, the credibility of the results will be unknown for time interval smaller than the calibration time interval.

It is critical to note that some of these time concepts are necessarily equal or larger than related time concepts, sometimes for logical reasons (the output resolution cannot be higher than the numerical resolution), sometimes for model credibility reasons (the interpretation time interval should not be smaller than the calibration time interval). It is also important to note that the first time concept, the process scale, explicitly links the temporal and the spatial scale (*Stommel*, 1963; *Blöschl and Sivapalan*, 1995; *Brutsaert*, 2005). Conversely, the spatial resolution of a model will set a minimum temporal resolution determining which processes need to be resolved.

7.3 Example for VIC model studies

To illustrate the development of calibration/validation time interval and spatial resolution in large-domain hydrological modelling, we carried out a meta-analysis on the use of GHMs. The Variable Infiltration Capacity (VIC) model (*Liang et al.*, 1994) was chosen for this analysis, because it is widely used and therefore enough studies were available for a meta-analysis. The VIC model is mentioned explicitly in *Bierkens et al.* (2015) as a type of model being run at the spatial hyper-resolution. Sub-grid variability is parameterized as a distribution of responses without explicit treatment of the pattern. We believe this model is representative for the much larger class of global hydrological models.

The VIC model was initially constructed to couple climate model output to hydrological processes: it is capable of solving both the energy and the water balance, see the model description in Section 2.2. *Lohmann et al.* (1996) developed a horizontal routing model to couple the individual grid cells of the VIC model. This facilitated the distributed application of VIC for rainfall-runoff processes at large domains. No explicit definition of a spatial derivative or scale appears in the equations of the VIC model, the spatial resolution of the model only appears in the routing scheme through the horizontal flow velocity (see *Kampf and Burges* (2007) for a description of space-time representation in other distributed hydrological models).

In our analysis we assembled 242 peer-reviewed studies that used the VIC model. Of these, 192 studies used the model in a distributed way and performed a calibration or validation on the model output. Figure 7.3 presents a space-time perspective on the application of the VIC model during the past two decades. As expected, the spatial resolution at which the model is applied has increased steadily over the years (Figure 7.3a). While the model was initially constructed for spatial resolutions in the order of $0.5°$ to $2°$, it is now mostly applied at $1/8°$ and smaller. The main driver for the increase in spatial resolution is the availability of high-resolution spatial data-sets, like presented by *Maurer et al.* (2002).

This increase, however, does not apply to the employed calibration and validation time interval. Figure 7.3b shows that the time interval at which the model has been calibrated and validated has remained steady over the years. So, while the spatial resolution of the model has increased, the model output is still calibrated and validated at the original coarse time interval. Processes with a short time scale, which become more important when the spatial resolution increases, will likely be overlooked during the calibration and validation of the model if the time interval is too coarse. Chapter 4 already showed
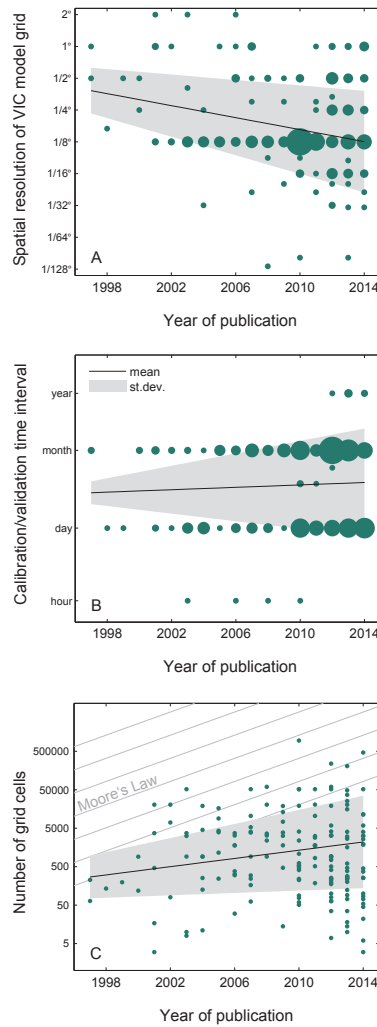
Figure 7.3: The year of publication versus the highest spatial resolution of the VIC model that was used in the study (a), the smallest time interval on which the calibration and/or validation of the VIC model was performed (b), and the total number of grid cells in the study (c) based on 192 peer-reviewed studies. The grey lines in (c) show the slope of computational power increase according to Moore's law (Moore, 1965). The point size is proportional to the number of studies that were published in a certain year with a certain spatial resolution or calibration/validation time interval. If the spatial resolution was given in kilometres, it was assumed that $1°=100$ km. For the total number of grid cells, catchment size was divided by cell size, assuming that $1°=100$ km, unless the number of grid cells was explicitly given. To obtain the mean and the standard deviation, both were calculated per year on the logarithmic scale and with linear regression a line was fitted through these points. An overview of the publications used to produce this figure can be found in Appendix Table B.1.

that calibration on a coarser time interval does not guarantee credible results for shorter time intervals (see also *Kavetski et al.*, 2011; *Littlewood and Croke*, 2013). Figure 7.1 indicates the initial development scale of the VIC model ('A'), the scale where it is heading to right now ('B'), and the direction where it should go in order to resolve relevant hydrometeorological processes ('C'). Therefore, the VIC model with a high spatial resolution should be calibrated and/or validated at a time interval short enough to catch the processes relevant at those particular spatial scales.

Two causes for the discrepancy in the joint development of spatial resolution and calibration time interval come to mind: lack of computational power, or a lack of (using) observations with a high temporal frequency. Figure 7.3c shows that the total number of grid cells that was used in the studies has on average increased over time. This is as expected: computational power has increased significantly over the years. According to Moore's law (*Moore*, 1965), computational power roughly doubles every two years. The grey lines in Figure 7.3c indicate the corresponding slope in computational power on a log-log scale. The largest numbers of grid cells per year likely indicate the limit of technical capability. Overall, the trend in the studies, even in the higher quantiles, is much lower than the computational limit, suggesting that computational power is not a constraint for most studies. This implies that, nowadays, the main constraint for calibration and validation of distributed hydrological models at a certain time interval (Figure 7.3b) is not the computational power, but the lack of (using) observations with a high temporal frequency. A possible explanation for this may be that many (global) studies rely on data from the Global Runoff Data Centre (GRDC), which are often available only at the monthly time interval. Also important is that for large basins, the typical application scale of VIC and other GHMs, flow is often regulated by dams for hydro-power and flood control. Naturalized flows for these basins are often estimated at the monthly time interval. Our results reinforce the conclusion of *Kirchner* (2006) that field observations should account for the spatial and temporal heterogeneity of hydro-meteorological processes, and the statement from *Kavetski et al.* (2011) that in most cases, temporal resolution is fixed by the data collection procedure.

7.4 Problem statement and outlook

The meta-analysis on VIC studies showed that the spatial resolution at which the model is applied has increased over the years, while the calibration time interval has remained steady (Figure 7.3). The examples are shown for the VIC model only, but we have the impression that the obtained trends apply for all GHMs. There is a general tendency to move towards higher spatial resolution in large-domain hydrological models (induced by e.g. *Wood et al.*, 2011; *Bierkens et al.*, 2015), whereas the available data for calibration and validation are model independent.

Although coarse temporal resolution data can be used to constrain model uncertainty, the ambition to move towards spatial hyper-resolution hydrological models with predictive capabilities should keep pace with the data that are required to run, calibrate and validate the models. Increasing the spatial resolution of the model implies modelling different relevant hydro-meteorological processes, which in turn requires calibration and

validation to be performed on a smaller time interval. It requires a community effort to increase the availability of high temporal resolution data for calibration and validation of large-domain hydrological models. Especially for large-domain studies, where data collection from all the separate basins at different institutes and countries is very time consuming (explaining the success of the GRDC), the data need to be gathered at and accessible from one point. It should also be recognized that discharge data only, especially at a monthly time scale, do not provide sufficient information for a process-based model evaluation at the spatial hyper-resolution scale. Possible paths forward are the use of tracer data to identify different flow paths (*Tetzlaff et al.*, 2015), the use of multiple objectives (*Gupta et al.*, 1998), and the use of satellite and remote sensing data (*Pan et al.*, 2008), all at a representative spatial ánd temporal resolution.

We acknowledge that calibration and validation at the appropriate time interval is only one of the many challenges of spatial hyper-resolution hydrological modelling. Even with enough observations available for calibration and validation, disinformative data (*Beven and Westerberg*, 2011), correct sub-grid parameterizations (*Beven et al.*, 2015) and model structural uncertainty (*Clark et al.*, 2015) remain outstanding challenges. However, we believe that all these challenges can only be tackled if the models are subject to critical and process-based evaluation and validation (*Gupta et al.*, 2008; *Clark et al.*, 2011b). In the end, the goal is to model hydrological processes in an appropriate way (*Beven*, 2006b; *McDonnell et al.*, 2007).

Along with an increased spatial resolution of the model products, there will be a shift in users' expectations of those products. Whereas coarse-scale (0.5° to 1°) products may provide relevant information for policy makers at the national or state level, products at the spatial hyper-resolution (0.1 to 1 km) are potentially of interest to a much wider range of users, including for instance farmers that want to schedule their irrigation. At the sub-kilometre scale, new processes such as infiltration excess overland flow and ponding can (and should) be resolved, but at the same time these processes cannot be explicitly resolved at a daily or monthly time interval. Thus, the recent call for increasing the spatial resolution of distributed hydrological models (*Wood et al.*, 2011; *Bierkens et al.*, 2015) should not focus solely on the spatial resolution, but should aim to increase the evaluation time interval simultaneously, at a balanced rate consistent with the characteristic time and space scales of the relevant hydrological processes (Figure 7.1). We believe that such a balanced approach will serve scientific and societal needs best.

Chapter 8

Reflection on hydrological modelling practice

This thesis provided an evaluation of several aspects of hydrological modelling practice. The introduction discussed six separate steps in the modelling process (*Beven*, 2012) from different perspectives (engineering, verificationism, falsificationism, Bayesianism). Several steps have been dealt with in more detail in the subsequent chapters of this thesis.

8.1 Overview of findings and lessons learned

In Chapter 3, three parameter identification methods and their data requirements were compared for a parsimonious two-parameter hydrological model. The three methods were: automatic calibration (DREAM), recession analysis, and Boussinesq theory. Automatic calibration and recession analysis had the same data requirements for the catchment in which they were tested, namely 5 months of data. Both methods did, however, come up with significantly different parameter values. With Boussinesq theory, a physics-based method which allowed parameter estimation a-priori based on catchment characteristics, the uncertainty in some physical characteristics caused the estimated range of feasible parameters to be too wide to be of added value.

Chapter 4 dealt with the transferability (and thus sensitivity) of parameters across spatial and temporal resolutions. It was shown that parameters were hardly sensitive to the spatial resolution (a high transferability), while the parameters were very sensitive to the temporal resolution (especially from hourly/daily to a monthly time step). The low sensitivity to temporal resolution could be caused by the uniform application of the most sensitive parameters: it was demonstrated, using several sub-catchments, that the spatial variability was strongly underestimated. There are, however, currently no strategies to apply the most sensitive parameters in a distributed fashion, because this is an ill-posed problem. Possible paths forward are for example Multiscale Parameter Regionalization (MPR, *Kumar et al.*, 2013), but this first requires that a relation between physical catchment characteristics and model parameters is established, which is not a trivial task. Hence, increasing the spatial resolution of the model does not automatically imply that the spatial variability is better represented in the model.

Chapter 5 showed that decisions during model configuration, basically subjective decisions of the modeller, significantly impact the simulation of hydrological extremes. It is argued that, although many modellers are aware of the impact of modelling decisions on the model output, many studies, for example model inter-comparison studies, do not yet explicitly incorporate this knowledge.

Chapter 6 demonstrated, for a large sample of catchments, that the uncertainty associated with the parameters in hydrological models, the hydrological model choice, and the GCM choice, is so large that even the sign (i.e., increase or decrease) of the projected change in mean annual discharge is unknown in the majority of the investigated catchments. This uncertainty could be attributed to the conceptualization of snow processes in the hydrological models and disagreement among climate models regarding the change in precipitation. Furthermore, it was demonstrated that processes related to aridity and intermittent flow behaviour are not yet well captured in the investigated hydrological models.

In Chapter 7 it is argued, based on a literature study of 192 peer-reviewed publications, that the calibration and validation time interval should keep pace with the spatial resolution of the model. Different hydrological processes are relevant at different spatial scales; the calibration and validation time interval should reflect the temporal scale of these processes in order to estimate the credibility of the model.

Based on the conclusions of each chapter, three main findings can be identified:

- Not only the hydrological model choice (and thus the underlying theory), but also the configuration of the model determines the outcome of the model as shown in Chapter 3 for different calibration methods, in Chapter 5 for different performance metrics and several other modelling decisions, and in Chapter 7 for different modelling time steps. This finding puts all the results of this thesis into perspective. Actually, Chapter 2 discusses a concatenation of subjective modelling decisions, such as modelling choice, sampling methodology, used input data, etc., and the results in this thesis depend on those modelling decisions.

- Sufficient data are needed to constrain and evaluate a hydrological model. Although it is already quite a cliché to conclude this, Chapter 4 clearly demonstrated this need. The unique advantage of a distributed model over a lumped model is the spatial distribution of forcing and parameters, but we need ways to identify the most sensitive parameters in a distributed fashion without ending up with an ill-posed problem. The literature survey in Chapter 7 also provides a motivation to improve data availability to stimulate more thorough model evaluation. Figure 8.1 illustrates that model complexity and data availability are related if we want to increase our understanding of the hydrological system. As such, it is possible that the approach adopted in Chapter 3, where we employ a very simple model with a large amount of data, could add more information to our system understanding than the more complex model employed in Chapter 7, where we hypothesise that insufficient data is available to evaluate the model.
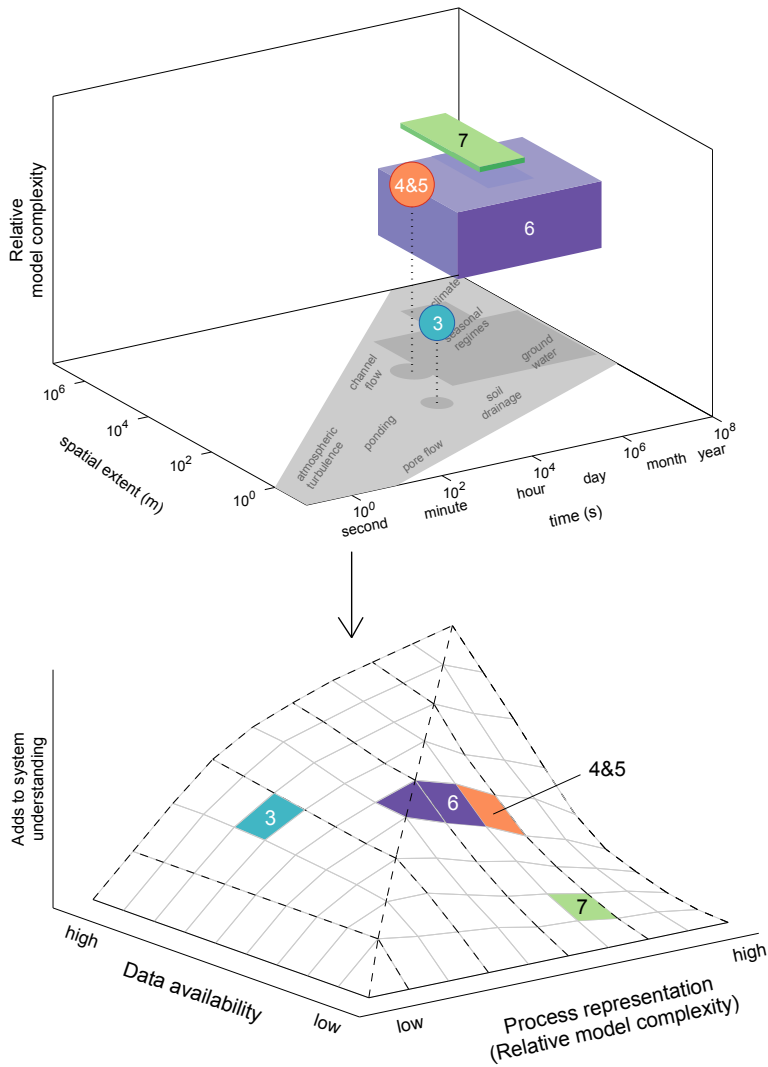
Figure 8.1: The relative complexity of the hydrological models employed in this thesis (upper panel, see also Figure 1.1) related to the required data availability in order to increase our understanding of the system (lower panel). The numbers refer to the Chapters in this thesis. The lower panel is adapted from Grayson and Blöschl (2000).

· As a results of the above two findings, I conclude that the uncertainty in hydrological modelling studies is large. Chapter 6 demonstrated that the uncertainty is even so great, that in a climate impact study for the contiguous United States, it could not even be determined whether the mean annual discharge would increase or decrease as a result of climate change for a large part of the catchments. For me this is a clear motivation to state that we should focus on a scientific approach - increasing our understanding of the system - rather than an engineering approach - applying our understanding -, because our current understanding on some topics seems insufficient. The opinion that we should focus on understanding rather than on predicting is shared by *Gupta and Nearing* (2014).

From the work presented in this thesis, several challenges become clear, which will be discussed in the following sections.

8.2 Hydrological modelling: Growing our knowledge, testing theories

In the introduction, it was argued that hydrologic science is too often mixed up with engineering approaches. Furthermore, it was argued that the difference between science and engineering in hydrological modelling can be larger or smaller, dependent on the science philosophy one adopts. In principle, every researcher should have the freedom to approach scientific questions in his or her own way. I do think, however, that falsificationism and Bayesianism stimulate the growth of scientific knowledge more than verificationism does. Verificationism, according to which confirming a model with observations is sufficient to perceive the model as an establishment of the 'truth', does not seem to stimulate critical reflection on one's own findings, and increases the risk to mix up science and engineering approaches. In a falsification framework, we would constantly try to falsify the underlying theory of our model, and subsequently adapt or reject this theory. In a Bayesian framework, we would continuously update the probability of a theory to be true. I would therefore argue to have a more critical attitude towards our hydrological models and embrace the falsification or Bayesian approach. A requirement for these scientific approaches is to re-establish the connection between a theory (a hypothesis) and a hydrological model.

It is challenging to disentangle the different theories in a hydrological model and to test these theories, individually and combined, either in the laboratory or under natural conditions. But it is necessary, to understand which processes are relevant where and when, and which assumptions are justified where and when. *Weiler and Beven* (2015) wrote, in their discussion if we need a Community Hydrological Model, that '*there are too many models in hydrology*'. Generally, I think there cannot be enough models, because every model implies a different theory which we could test. What, however, is lacking is a general overview or insight regarding the representation of the underlying processes to investigate how the different hydrological models (and thus theories or combinations of theories) are related to each other, such as for example demonstrated with the model genealogy of *Knutti et al.* (2013) for climate models. This model genealogy was useful in Chapter 6 because it enabled us to estimate the uncertainty associated with the choice of climate models by using only a sub-sample of all available models. Whereas *Knutti et al.* (2013)

based their genealogy on model output, ideally we would like to obtain a model genealogy based on the underlying theories. Therefore, more attention should be given to the theories underlying hydrological models.

My plea to use falsificationism or Bayesianism in hydrological modelling to grow scientific knowledge, and the required re-establishment of the link between models and theories, does not exclude nor favour the physics-based or the conceptual modelling approaches. Both tracks of modelling, sometimes referred to as the 'Newtonian' and the 'Darwinian' approaches, respectively (*Sivapalan et al.*, 2011; *Harman and Troch*, 2014), have clear advantages and drawbacks. The advantage of the physics-based Boussinesq theory as applied in Chapter 3 is that the parameters could directly be linked to physical catchment characteristics, but one could wonder if representing a catchment as a single hillslope is a correct conceptualization of the system. On the other hand, Chapter 4 showed that in a distributed conceptual model, we have difficulties applying the most sensitive parameters in a distributed fashion because there is no established link between catchment characteristics and model parameters. The way ahead should therefore be found in pluriformity (as also advocated by *Harte*, 2002, and *Sivapalan et al.* 2011): lumped and distributed, conceptual and physical, large scale and small scale, case studies and multi-catchment studies. The only condition we should set for scientific hydrological modelling studies is that they test a theory (hypothesis-based), which is falsifiable.

Model complexity should relate to the available data, as shown in Figure 8.1. Promising observation technologies, e.g. groundwater observations as explored with the Gravity Recovery And Climate Experiment (GRACE) satellite (*Famiglietti et al.*, 2011), rainfall observations using widespread microwave links (*Rios Gaona et al.*, 2015), fibre-optic cables for distributed temperature sensing (*Bense et al.*, 2016), the use of cosmic-ray neutron sensors to measure soil moisture (*Iwema et al.*, 2015), or neutron tomography to determine connectivity in porous media (*Schaap et al.*, 2008) lead to opportunities to define and test new theories. Furthermore, they provide new opportunities to falsify or test established theories and models. If we really want to go down the falsificationism or Bayesianism road, we need thorough testing of our hypotheses, so-called 'crash-testing' (*Andréassian et al.*, 2009). It is not enough to confirm our model with discharge data, we need to test the internal consistency (e.g. *Rakovec et al.*, 2016), and quantify the model performance with more than just a single measure providing a global estimate of the average performance, for example by using signatures of several states and fluxes (*Gupta et al.*, 2008; *Sawicz et al.*, 2011). New observation technologies provide an outstanding opportunity to do such crash-testing. We should, however, not only follow an inductive approach (*Gupta and Nearing*, 2014), where new observations lead to the formulation of new insights, but also appreciate a parallel path, where new insights can lead to new measurements or new observation techniques. The growth of scientific knowledge is a continuous interplay between new theories and new observations.

A recent initiative to structure hydrological theories for different parts of the hydrological cycle is the SUMMA framework (*Clark et al.*, 2015), which allows to combine different theories for different processes. This is an interesting start for structured hypothesis testing

that should be combined with sophisticated observations in order to constrain the model sufficiently. Another approach is the ABC framework (*Vrugt and Sadegh*, 2013) where, through calibration, the weaknesses in a model can be identified and subsequently improved. However, as demonstrated in several chapters of this thesis (most pronounced in Chapter 5), the model output not only depends on the underlying theory, but also on many modelling decisions throughout the modelling process. Therefore, if we want to test the theory in a modelling framework, every modelling decision should be considered as part of the tested theory. Working with several hypotheses (*Chamberlin*, 1890; *Clark et al.*, 2011a) should therefore not only apply to the underlying theories, but also to the applied model configuration.

Summarizing, I think the growth of hydrological scientific knowledge using hydrological models is best stimulated within a Bayesian or falsification framework. In order to do this, the link between a theory and a hydrological model should be re-established: theories should be perceived as hypotheses, and must be falsifiable. I argue that we should conduct thorough hydrological model evaluation, including testing for internal consistency if the model combines several theories. The required observations depend on the theory being tested. Therefore, advanced hydrological observations beyond discharge and precipitation are needed to test and eventually confirm or falsify the theory. Furthermore, the decisions during model configuration should be explicitly recognized as elements of the tested theory. The proposed approach reflects the quote at the beginning of this thesis: '*Theories crumble, but good observations never fade*'.

8.3   On repeatability and reproducibility

Another condition for scientific hydrological modelling, recently advocated by *Hutton et al.* (2016), is that modelling studies should be reproducible. Unfortunately, *Hutton et al.* (2016) refrain from providing a clear definition on what they see as 'reproducibility' in the context of computational hydrology. In our comment to this opinion paper (*Melsen et al.*, 2017) we discuss two definitions:

· The complete model study, including the exact numbers, should be reproducible (bit-reproducibility);

· The conclusions of the model study should be reproducible (conclusion-reproducibility).

Bit-reproducibility is reproducing the results of the same computer code combined with the same data. This is not always trivial because specific platforms or software packages might be needed. However, if one succeeds in running the same code with the same data, this is not much more than a repetition exercise, because the same computer code combined with the same data should always lead to the same results (as no epistemic uncertainty is involved), and different persons pushing the "enter" button to run the same computer code with the same data should also lead to the same results (as no subjectivity is involved). This view is shared by *Easterbrook* (2014), who states that repeatability and reproducibility are often mixed up in scientific computing. More interesting for growing scientific knowledge is whether the conclusions of a model study are reproducible. In

computational hydrology, this can be established by testing the same model with different data, or testing a different model set-up with the same data. By testing the reproducibility of model conclusions, one actively aims to verify (logical positivist) or falsify (in the sense of Popper) the underlying theory.

Reproducibility of model conclusions provides founded reasons to improve the transparency of modelling studies and to increase the availability of model code and data, especially in the light of a falsification or Bayesian approach. First of all, it increases the transparency of the tested theory or set of theories (which can be quite complex in some hydrological models). This closely relates to the point in the previous section, that we should give more attention to the theories underlying a model. Secondly, it provides insights in the subjective modelling decisions that were made during the model configuration and which can have a profound impact on the modelling results, as demonstrated in Chapter 5. Furthermore, it enhances the opportunity to build on already existing theories. Finally, most importantly, it allows to test the theories for different situations or catchments, to investigate the limits of applicability of a theory.

8.4   Hydrology as a connecting science

Hydrology is *'a distinct geoscience that cuts across the traditional atmospheric, ocean and solid earth sciences'* (*Eagleson*, 1991). It is especially this integrated part, and the engineering part of hydrology, that cause a lack of attention to the science of hydrology according to *Yevjevich* (1968) and *Klemeš* (1986a), as discussed in Chapter 1. Although I clearly share the opinion that the engineering approach obstructs progress in hydrologic sciences – which by no means implies that I do not see the value of engineering, especially as I have lived a large part of my life below sea level –, I do perceive it as valuable that hydrology is considered as an integrated science.

Our enhanced process understanding, for example concerning soil-moisture precipitation feedbacks (*Guillod et al.*, 2015), the role of vegetation and vegetation dynamics in the hydrological cycle (*Schymanski et al.*, 2009; *Williams et al.*, 2012) and on the the interaction between available soil moisture and transpiration (*Baldocchi et al.*, 2004), taught us that some hydrological processes cannot be considered in isolation from other sciences. The development of land-surface models clearly demonstrates the insight that hydrological behaviour depends on amongst others meteorology, soil science, and ecology, and vice versa. Enhanced understanding of hydrological processes therefore implies that we have to extend the field of interest and the field of study (*Paola et al.*, 2006).

Another process only recently picked up in hydrological modelling, is the role of human beings on the hydrological system (see for example *Wada et al.*, 2014; *Van Loon et al.*, 2016). Now that the current era is called '*The Anthropocene*', the geologic time period where the impact of humans on the Earth is larger than the impact of natural systems, it would be obsolete not to recognize and study the impact of human beings on the hydrological cycle. A frequently used concept like 'naturalized stream flows' not only completely ignores the processes that are taking place, but also does not allow for thorough model evaluation, because no observations are available for artificial data series. We should aim

to increase our understanding of the impact of human beings on the hydrological cycle, whether it be indirect (e.g. through climate change or land use change) or direct (e.g. through groundwater abstraction).

The goal of science is to increase our understanding of the system under study. The strongest way of understanding might be expressed by identifying 'universal laws' (*Carnap*, 1966) - regularities observed at all times and all places. Although hydrology and hydrological modelling have some universal laws they try to respect, such as the conservation of mass when establishing a water balance, these universal laws are based on physics rather than on hydrological understanding. So far, no universal laws for hydrological processes have been identified - Darcy's law, for example, might work perfectly in a lab-setting but is difficult in a field-setting with large spatial variation in hydrological conductivity. This can be explained on the basis of the science classification as defined by Auguste Comte (1798-1857), who is sometimes perceived as the first modern philosopher of Science (*Bourdeau*, 2014). Comte classified sciences according to their generality and the complexity of the investigated system, which resulted in the following order from more general applicable laws to less general applicable laws and from a less complex system to a more complex system under study: mathematics, astronomy, physics, chemistry, biology, sociology. According to his theory, within mathematical sciences general laws can be established which are always true, independent of knowledge of any other science. Astronomy is one step lower in the generality-hierarchy, because astronomy needs mathematical laws to define astronomical laws, and one step higher in the complexity-hierarchy because more factors than mathematics alone influence astronomy. Earth and Life sciences can be classified at the level of biology; within these sciences laws are less general and the system under study is more complex. Hydrologic sciences has to combine the laws of all the sciences that are higher in the generality-hierarchy, which includes chemistry, physics (see the example of the conservation of mass) and mathematics. At the same time, the system under study is highly complex as it is influenced by many factors, such as vegetation and humans which have been discussed earlier in this section. The aim to identify 'universal laws' in hydrology is therefore probably not feasible. This does imply, however, that every hydrological theory can be falsified somewhere, because no law and thus no theory either holds generally true. Therefore, if we test a theory on 100 basins and the result is that we can falsify this theory in 10 of them, we should not reject this theory immediately but try to investigate which factor or process influences the confirmation or falsification of this specific theory; we have to identify the limits of applicability for each theory. I believe that such an approach would stimulate the growth of hydrological knowledge.

Appendices
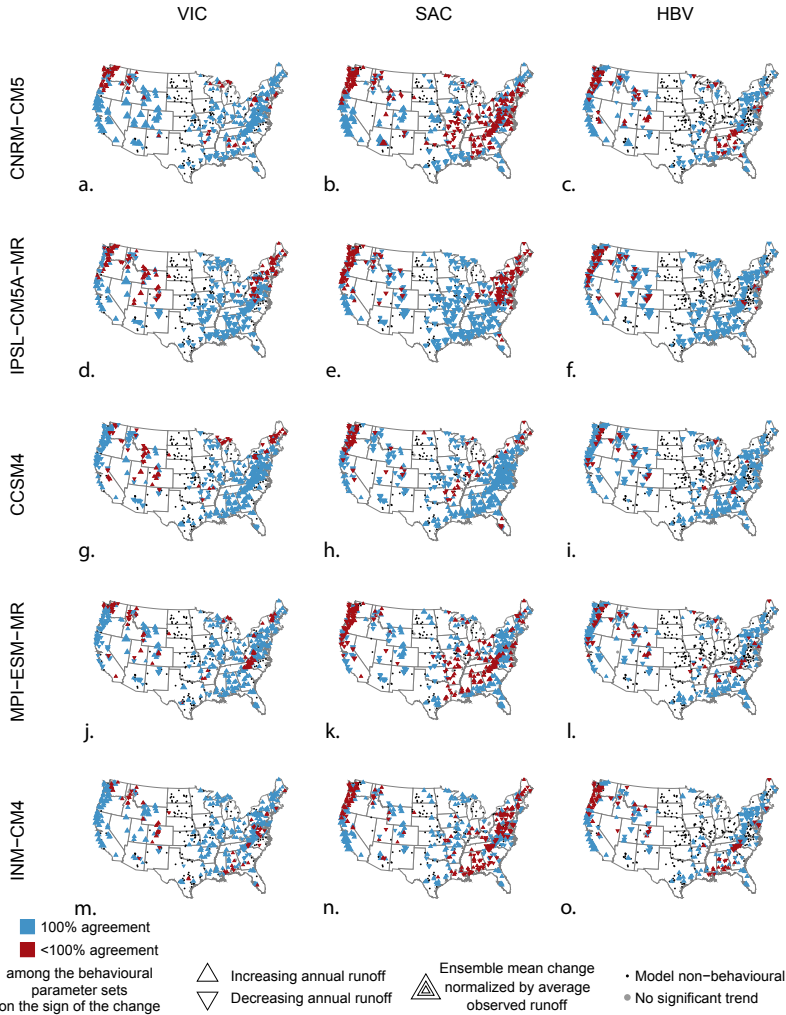
Figure A.1: The agreement among the different model runs (representing different behavioural parameter sets) on the sign of the ensemble mean change in mean annual runoff for three hydrological models (columns) forced with five different GCMs (rows). The direction of the triangle-marker shows the sign of the ensemble mean change, the size of the marker indicates the relative projected change. a) VIC forced with CNRM-CM5. b) SAC forced with CNRM-CM5. c) HBV forced with CNRM-CM5. d) VIC forced with IPSL-CM5A-MR. e) SAC forced with IPSL-CM5A-MR. f) HBV forced with IPSL-CM5A-MR. g) VIC forced with CCSM4. h) SAC forced with CCSM4. i) HBV forced with CCSM4. j) VIC forced with MPI-ESM-MR. k) SAC forced with MPI-ESM-MR. l) HBV forced with MPI-ESM-MR. m) VIC forced with INM-CM4. n) SAC forced with INM-CM4. o) HBV forced with INM-CM4.
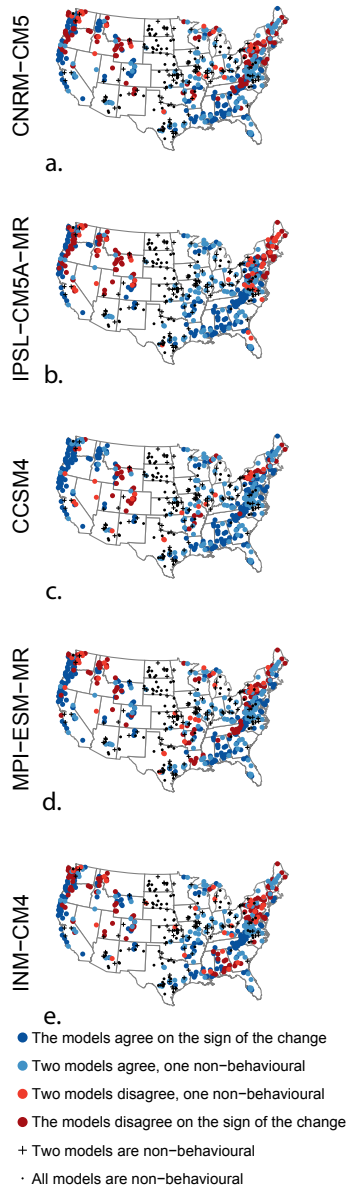
Figure A.2: The agreement among the three hydrological models on the sign of the ensemble mean change in mean annual runoff, forced with five different GCMs. a) The three hydrological models are forced with CNRM-CM5 data. b) The three hydrological models are forced with IPSL-CM5A-MR data. c) The three hydrological models are forced with CCSM4 data. d) The three hydrological models are forced with MPI-ESM-MR data. e) The three hydrological models are forced with INM-CM4 data.
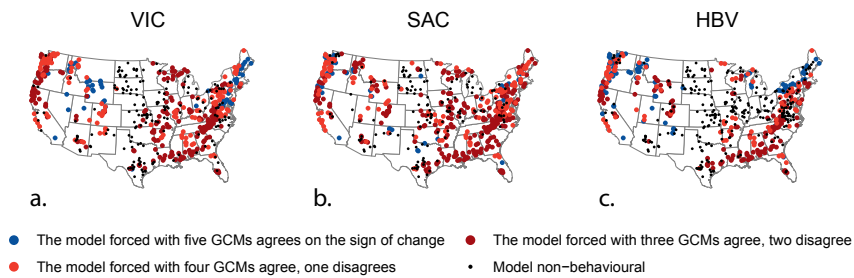
Figure A.3: The agreement on the sign of the ensemble mean change in mean annual runoff in the output from the same hydrological model forced with five different GCMs. a) Agreement when VIC is forced with data from five different GCMs. b) Agreement when SAC is forced with data from five different GCMs. c) Agreement when HBV is forced with data from five different GCMs.

Table B.1: All articles used to create Figure 7.3, with their highest spatial resolution (in degrees) and the time interval used for calibration and validation.

| Authors	Journal	Year	Title	Spat.	Temp.
Abdullah, F.A., and D.P. Lettenmaier	J. Hydrol.	1997	Application of regional parameter ...	1.000	monthly
Acharya, A., et al.	J. Hydrol.	2011	Modeled streamflow response ...	0.125	monthly
Adam, J.C. , et al.	J. Geophys. Res.	2007	Simulation of reservoir influences ...	1.000	monthly
Agboma, C.O., et al.	J. Hydrol.	2009	Intercomparison of the total storage ...	0.300	monthly
Ahmad, S., et al.	Adv. Water Resour.	2010	Estimating soil moisture ...	0.125	daily
Andreadis, K.M., and D.P. Lettenmaier	Adv. Water Resour.	2006	Assimilating remotely sensed ...	0.125	daily
Arora, V.K., and G.J. Boer	J. Climate	2006	The temporal variability of ...	2.000	monthly
Ashfaq, M., et al.	J. Geophys. Res.	2010	Influence of climate model ...	0.125	daily
Bao, Z., et al.	J. Hydrol.	2012	Comparison of regionalization ...	0.250	monthly
Bao, Z., et al.	J. Hydrol.	2012	Attribution for decreasing ...	0.250	monthly
Bao, Z., et al.	Hydr. Process.	2012	Sensitivity of hydrological ...	0.250	monthly
Bohn, T.J., et al.	Env. Res. Letters	2007	Methane emissions from ...	1.000	daily
Bohn, T.J., et al.	J. Hydrometeorol.	2010	Seasonal Hydrologic Forecasting ...	0.125	monthly
Bowling, L.C., and D.P. Lettenmaier	J. Hydrometeorol.	2010	Modeling the Effects of ...	0.125	hourly
Chang, J., et al.	Quater. Int.	2014	Impact of climate change ...	0.500	daily
Cherkauer, K.A., and D.P. Lettenmaier	J. Geophys. Res.	1999	Hydrologic effects of frozen soils ...	0.500	daily
Christensen, N.S., et al.	Climatic Change	2004	The effect of climate change on ...	0.125	daily
Christensen, N.S., and D.P. Lettenmaier	Hydrol. Earth Syst. Sc.	2007	A multimodel ensemble approach ...	0.125	daily
Costa-Cabral, M., et al.	Climatic Change	2013	Snowpack and runoff response ...	0.125	monthly
Crow, W.T., et al.	J. Geophys. Res.	2003	Multiobjective calibration of ...	0.125	hourly
Cuo, L., et al.	J. Hydrol.	2013	The impacts of climate change ...	0.250	daily
Demaria, E.M.C. , et al.	J. Hydrol.	2013	Climate change impacts on ...	0.250	daily
Demaria, E.M.C., et al.	Int. J. River Bas. Manag.	2014	Satellite precipitation in ...	0.125	monthly
Díaza, A., et al.	Int. J. River Bas. Manag.	2013	Multi-annual variability of ...	0.125	daily
Drusch, M., et al.	Gephys. Res. Lett.	2005	Observation operators for the ...	0.125	daily
Eum, H., et al.	Hydr. Process.	2014	Uncertainty in modelling the ...	0.063	daily
Feng, X., et al.	J. Hydrometeorol.	2008	The Impact of Snow Model ...	0.125	daily
Ferguson, C.R., et al.	Int. J. Remote Sens.	2010	Quantifying uncertainty in ...	0.125	monthly
Ferguson, C.R., et al.	J. Hydrometeorol.	2012	A Global Intercomparison of ...	0.250	daily
Gao, H., et al.	J. Hydrometeorol.	2004	Using a Microwave Emission ...	0.125	daily
Gao, H., et al.	J. Hydrometeorol.	2006	Using TRMM/TMI to Retrieve ...	0.125	daily
Gao, H., et al.	J. Hydrometeorol.	2007	Copula-Derived Observation ...	0.125	daily
Gao, H., et al.	Int. J. Remote Sens.	2010	Estimating the water budget ...	0.500	monthly
Gao, Y., et al.	J. Geophys. Res.	2011	Evaluating climate change ...	0.125	monthly
Garg, V., et al.	J. Hydr. Eng.	2013	Hypothetical scenario-based ...	0.250	yearly
Gebregiorgis, A., and F. Hossain	J. Hydrometeorol.	2011	How Much Can A Priori Hydrologic ...	0.125	daily
Gebregiorgis, A.S., et al.	Water Resour. Res.	2012	Tracing hydrologic model ...	0.125	daily
Gu, H., et al.	Stoch. Environ. Res. Risk Ass.	2014	Impact of climate change ...	0.125	daily
Guerrero, M., et al.	Int. J. River Bas. Manag.	2013	Parana River morphodynamics ...	0.125	monthly
Guo, J., et al.	J. Hydrol.	2004	Impacts of different precipitation ...	0.125	daily
Guo, J., et al.	Proc. Env. Sci.	2011	Daily runoff simulation in ...	0.042	daily
Haddeland, I., et al.	Gephys. Res. Lett.	2006	Anthropogenic impacts on ...	0.500	monthly
Haddeland, I., et al.	J. Hydrometeorol.	2006	Reconciling Simulated Moisture ...	0.125	hourly
Haddeland, I., et al.	J. Hydrol.	2006	Effects of irrigation on the ...	0.500	daily
Hamlet, A.F., et al.	J. Climate	2005	Effects of Temperature and ...	0.125	monthly
Hamlet, A.F., and D.P. Lettenmaier	Water Resour. Res.	2007	Effects of 20th century warming ...	0.125	monthly
Hidalgo, H.G., et al.	J. Hydrol.	2013	Hydrological climate change ...	0.500	monthly
Hillarda, Y., et al.	Remote Sens. Environ.	2003	Assessing snowmelt dynamics ...	0.125	daily
Huang, M., et al.	J. Geophys. Res.	2003	A transferability study of model ...	0.130	daily
Hurkmans, R.T.W.L., et al.	Water Resour. Res.	2008	Water balance versus land ...	0.088	daily
Hurkmans, R.T.W.L., et al.	Water Resour. Res.	2009	Effects of land use changes ...	0.050	daily
Hurkmans, R., et al.	J. Climate	2010	Changes in Streamflow Dynamics ...	0.088	daily
Jayawardena, A.W., et al.	J. Hydrolog. Eng.	2002	Meso-Scale Hydrological ...	1.000	daily
Kam, J., et al.	J. Climate	2013	The Influence of Atlantic ...	0.125	daily
Lakshmi, V., et al.	Gephys. Res. Lett.	2004	Soil moisture as an ...	0.125	monthly
Li, J., et al.	J. Hydrometeorol.	2007	Modeling and Analysis ...	0.042	daily
Li, H., et al.	J. Hydrometeorol.	2013	A Physically Based Runoff ...	0.063	monthly
Liang, X. and Z. Xie	Adv. Water Resour.	2001	A new surface runoff ...	0.125	daily
Liang ,X., and Z. Xie	Global Planet. Change	2003	Important factors in land ...	0.125	daily
Liang, X., et al.	J. Geophys. Res.	2003	A new parameterization ...	0.125	daily
Liang, X., et al.	J. Hydrol.	2004	Assessment of the effects ...	0.031	daily
Liu, Z., et al.	Hydr. Process.	2010	Impacts of climate change on ...	0.500	daily
Liu, L., et al.	J. of Flood Risk Manag.	2013	Hydrological analysis for water ...	0.010	daily
Liu, H., et al.	Hydrol. Earth Syst. Sc.	2013	Soil moisture controls on ...	0.500	monthly
Liu, X., et al.	Hydrol. Earth Syst. Sc.	2014	Effects of surface wind speed ...	0.250	monthly
Livneh, B., et al.	J. Climate	2013	A Long-Term Hydrologically ...	0.063	monthly
Lohmann, D., et al.	Hydrolog. Sci. J.	1998	Regional scale hydrology: ...	0.167	daily
Lu, X., and Q. Zhuang	J. Geophys. Res.	2012	Modeling methane emissions ...	0.333	daily
Lucas-Picher, P., et al.	Atmosphere-Ocean	2003	Implementation of a ...	0.405	monthly
Luo, Y., et al.	J. Hydrometeorol.	2005	The Operational Eta Model ...	0.125	monthly
Luo, L., and E.F. Wood	Gephys. Res. Lett.	2007	Monitoring and predicting ...	0.125	monthly
Luo, L., and E.F. Wood	J. Hydrometeorol.	2008	Use of Bayesian Merging ...	0.125	monthly
Lutz, E.R., et al.	Water Resour. Res.	2012	Paleoreconstruction of cool ...	0.063	monthly

Author	Journal	Year	Title	Value	Frequency
Mao, D., and K.A. Cherkauer	J. Hydrol.	2009	Impacts of land-use change ...	0.125	monthly
Mao, D., et al.	Water Resour. Res.	2010	Development of a coupled ...	0.125	daily
Marshall, M., et al.	Climate Dynamics	2012	Examining evapotranspiration ...	1.000	monthly
Matheussen, B., et al.	Hydr. Process.	2000	Effects of land cover change ...	0.250	monthly
Maurer, E.P., et al.	J. Geophys. Res.	2001	Evaluation of the land ...	0.125	monthly
Maurer, E.P., et al.	J. Climate	2002	A Long-Term Hydrologically ...	0.125	monthly
McGuire, M., et al.	J. Water Resour. Plan. Manage.	2006	Use of Satellite Data for ...	0.125	monthly
Meng, L., and S.M. Quiring	Int. J. Climatol.	2010	Observational relationship of ...	0.500	monthly
Miguez-Macho, G., et al.	Bull. Am. Meteor. Soc.	2008	Simulated water table ...	0.008	monthly
Miller, W.P., et al.	J. Water Res. Pl. Manag.	2012	Water Management Decisions ...	0.125	monthly
Minihane, M.R.	Phys. Chem. Earth	2012	Evaluation of streamflow ...	0.250	monthly
Mishra, V., et al.	J. Hydrometeorol.	2010	Parameterization of Lakes ...	0.125	daily
Mishra, V., and K.A. Cherkaue	Agric. For. Meteorol.	2010	Retrospective droughts in ...	0.125	monthly
Mishra, V., et al.	J. Hydrometeorol.	2010	Assessment of Drought due ...	0.125	monthly
Mishra, V., et al.	Int. J. Clim.	2010	A regional scale assessment ...	0.125	monthly
Mishra, V., et al.	Global Planet. Change	2011	Lake Ice phenology of ...	0.125	daily
Mishra, V., et al.	Global Planet. Change	2011	Changing thermal dynamics ...	0.125	daily
Mishra, V., and K.A. Cherkauer	J. Geophys. Res.	2011	Influence of cold season ...	0.125	daily
Mo, K.C.	J. Hydrometeorol.	2008	Model-Based Drought Indices ...	0.500	monthly
Mo, K.C., et al.	J. Hydrometeorol.	2012	Uncertainties in North American ...	0.500	monthly
Munoz-Arriola, F., et al.	Water Resour. Res.	2009	Sensitivity of the water resources ...	0.125	monthly
Nijssen, B., et al.	Water Resour. Res.	1997	Streamflow simulation for ...	0.500	monthly
Nijssen, B., et al.	J. Climate	2001	Global Retrospective Estimation ...	2.000	monthly
Nijssen, B., et al.	Climatic Change	2001	Hydrologic sensitivity of global ...	1.000	monthly
Niu, J., et al.	J. Hydrol.	2013	Impacts of increased CO2 ...	1.000	monthly
Niu, J., and J. Chen	Hydrological Sciences Journal	2014	Terrestrial hydrological responses ...	1.000	daily
Niu, J., and B. Sivakumar	Stoch Environ Res Risk Assess	2014	Study of runoff response to ...	1.000	monthly
Niu, J., et al.	Hydrol. Earth Syst. Sc.	2014	Teleconnection analysis of ...	1.000	daily
Null, S.E., and J.H. Viers	Water Resour. Res.	2013	In bad waters: Water year ...	0.125	monthly
O'Donnell, G.M., et al.	J. Geophys. Res.	2000	Macroscale hydrological modeling ...	0.500	monthly
Oubeidillah, A.A., et al.	Hydrol. Earth Syst. Sc.	2014	A large-scale, high-resolution ...	0.042	monthly
Ozdogan, M.	Hydrol. Earth Syst. Sc.	2011	Climate change impacts on ...	0.125	monthly
Pan, M., and E.F. Wood	J. Hydrometeorol.	2006	Data Assimilation for ...	0.500	daily
Parada, L.M., and X. Liang	J. Geophys. Res.	2004	Optimal multiscale Kalman ...	0.125	daily
Parada, L.M., and X. Liang	J. Geophys. Res.	2008	Impacts of spatial resolutions ...	0.125	daily
Park, D., and M. Markus	J. Hydrol.	2014	Analysis of a changing ...	0.125	daily
Parr, D., and G. Wang	Global and Planetary Change	2014	Hydrological changes in the ...	0.030	daily
Qiao, L., et al.	Water Resources Management	2014	Climate Change and ...	0.125	daily
Qin, S., et al.	Int. J. Remote Sens.	2013	Development of a hierarchical ...	0.030	daily
Raje, D., and R. Krishnan	Water Resour. Res.	2012	Bayesian parameter uncertainty ...	1.000	monthly
Raje, D., et al.	Hydr. Process.	2014	Macroscale hydrological modelling ...	1.000	monthly
Ray, R.L., et al.	Remote sensing of environment	2010	Landslide susceptibility mapping ...	0.010	daily
Rhoads, J., et al.	J. Geophys. Res.	2001	Validation of land surfacemodels ...	1.000	daily
Rosenberg, E.A., et al.	Water Resour. Res.	2011	Statistical applications of...	0.063	monthly
Rosenberg, E.A. , et al.	Hydrol. Earth Syst. Sc.	2013	On the contribution of ...	0.125	daily
Saurral, R.I.	J. Hydrometeorol.	2010	The Hydrologic Cycle of the ...	0.125	monthly
Schaller, M.F., and Y. Fan	J. Geophys. Res.	2009	River basins as groundwater ...	0.125	monthly
Schumann, G.J.-P., et al.	Water Resour. Res.	2013	A first large-scale flood ...	0.250	daily
Sheffield, J., et al.	J. Geophys. Res.	2003	Snow process modeling ...	0.125	daily
Sheffield, J., et al.	J. Hydrometeorol.	2012	Representation of Terrestrial ...	0.500	monthly
Shi, X., et al.	Env. Res. Letters	2011	The role of surface energy ...	1.000	weekly
Shi, X., et al.	J. Climate	2013	Relationships between Recent ...	1.000	monthly
Shrestha, R.R., et al.	Hydr. Process.	2012	Modelling spatial and ...	0.063	monthly
Shrestha, K.Y., et al.	J. Hydrometeorol.	2014	An Atmospheric-Hydrologic ...	0.250	daily
Shrestha, R.R., et al.	J. Hydrometeorol.	2014	Evaluating Hydroclimatic ...	0.063	daily
Shrestha, R.R., et al.	Hydr. Process.	2014	Evaluating the ability of a ...	0.063	monthly
Shukla, S., et al.	Hydrol. Earth Syst. Sc.	2012	Value of medium range ...	0.500	2-weeks
Shukla, S., et al.	Hydrol. Earth Syst. Sc.	2012	On the sources of global land ...	0.500	monthly
Sinha, T., et al.	J. Hydrometeorol.	2010	Impacts of Historic Climate ...	0.125	weekly
Sinha, T. and K.A. Cherkauer	J. Geophys. Res.	2010	Impacts of future climate ...	0.125	weekly
Sinha, T. and A. Sankarasubramanian	Hydrol. Earth Syst. Sc.	2013	Role of climate forecasts and ...	0.125	monthly
Slater, A.G., et al.	J. Geophys. Res.	2007	A multimodel simulation of ...	1.000	monthly
Sridhar, V., et al.	Climate Dynamics	2013	Explaining the hydroclimatic ...	0.125	monthly
Stephen, H., et al.	Hydrol. Earth Syst. Sc.	2010	Relating surface backscatter ...	0.125	daily
Su, F., et al.	J. Geophys. Res.	2005	Streamflow simulations of ...	1.000	monthly
Su, F., et al.	J. Geophys. Res.	2006	Evaluation of surface water ...	1.000	monthly
Su, F., et al.	J. Hydrometeorol.	2008	Evaluation of TRMM Multisatellite ...	0.125	daily
Su, F., and D.P. Lettenmaier	J. Hydrometeorol.	2009	Estimation of the Surface ...	0.125	monthly
Tang, C., et al.	J. Hydrol.	2009	Spatial and temporal soil ...	0.125	monthly
Tang, Q and D.P. Lettenmaier	Int. J. Remote Sens.	2010	Use of satellite snow-cover ...	0.063	monthly
Tang, C., et al.	J. Hydrol.	2011	Relationships between ...	0.125	monthly
Tang, Q., et al.	J. Hydrometeorol.	2012	Predictability of Evapotranspiration ...	0.063	daily
Tang, C., et al.	Global Planet. Change	2012	Assessing streamflow sensitivity ...	0.063	monthly
Tang, C., and R.L. Dennis	Global Planet. Change	2014	How reliable is the offline ...	0.125	monthly
Fan, Y., et al.	J. Hydrometeorol.	2011	Verification and Intercomparison ...	0.125	monthly
Vano, J.A., et al.	J. Hydrometeorol.	2012	Hydrologic Sensitivities of ...	0.125	monthly
VanShaar, J.R., et al.	Hydr. Process.	2012	Effects of land-cover changes ...	0.125	monthly
Vicuna, S., et al.	J. Am. Water Resour. As.	2007	The sensitivity of California ...	0.125	monthly

Voisin, N., et al.	J. Hydrometeorol.	2008	Evaluation of Precipitation ...	0.500	monthly
Voisin, N.,et al.	Weather Forecast.	2011	Application of a Medium-Range ...	0.250	daily
Wang, A., et al.	J. Geophys. Res.	2008	Integration of the variable ...	0.125	monthly
Wang, J., et al.	Int. J. Clim.	2010	Quantitative assessment of climate ...	0.125	monthly
Wang, G.Q., et al.	Hydrol. Earth Syst. Sc.	2012	Assessing water resources in ...	0.500	daily
Werner, A.T., et al.	Atmosphere-Ocean	2013	Spatial and Temporal Change ...	0.063	daily
Wen, Z., et al.	Water Resour. Res.	2012	A new multiscale routing ...	0.031	daily
Wenger, S.J., et al.	Water Resour. Res.	2010	Macroscale hydrologic ...	0.063	daily
Wojcik, R., et al.	J. Hydrometeorol.	2008	Multimodel Estimation of ...	0.125	hourly
Wood, A.W., et al.	J. Geophys. Res.	2002	Long-range experimental ...	0.125	monthly
Wood, A.W., et al.	J. Geophys. Res.	2005	A retrospective assessment ...	0.125	monthly
Wu, Z., et al.	Atmosphere-Ocean	2007	Thirty-Five Year (1971-2005) ...	0.300	daily
Wu, Z.Y., et al.	Hydrol. Earth Syst. Sc.	2011	Reconstructing and analyzing ...	0.300	daily
Wu, H., et al.	Water Resour. Res.	2014	Real-time global flood ...	0.125	daily
Xia, Y., et al.	J. Geophys. Res.	2012	Continental-scale water ...	0.125	daily
Xia, Y., et al.	Hydr. Process.	2012	Comparative analysis of ...	0.125	monthly
Xia, Y., et al.	Hydr. Process.	2014	Evaluation of NLDAS-2 ...	0.125	daily
Xie, Z., et al.	J. Hydrometeorol.	2007	Regional Parameter Estimation ...	0.500	monthly
Yang, G., et al.	J. Hydrometeorol.	2010	Hydroclimatic Response of ...	0.125	daily
Yang, G., et al.	Landscape Urban Plan.	2011	The impact of urban development ...	0.125	daily
Yang, G., and L.C. Bowling	Water Resour. Res.	2014	Detection of changes in ...	0.125	daily
Yearsley, J.	Water Resour. Res.	2012	A grid-based approach for ...	0.063	daily
Yong, B., et al.	Water Resour. Res.	2010	Hydrologic evaluation of ...	0.063	daily
Yong, B., et al.	J. Hydrometeorol.	2013	Spatial-Temporal Changes of ...	0.063	daily
Yuan, F., et al.	Can. J. Rem. Sens.	2004	An application of the VIC-3L ...	0.250	daily
Yuan, X., et al.	Hydr. Sci. J.	2009	Sensitivity of regionalized ...	0.500	monthly
Yuan, X., et al.	J. Hydrometeorol.	2013	Probabilistic Seasonal ...	0.250	monthly
Zeng, X., et al.	J. Hydrometeorol.	2010	Comparison of Land-Precipitation ...	0.125	monthly
Zhang, X., et al.	Phys. Chem. Earth	2012	Modeling and assessing ...	0.031	monthly
Zhang, B., et al.	Agr. Water Manage.	2012	Drought variation trends in ...	0.500	yearly
Zhang, B., et al.	Theor. Appl. Climatol.	2013	A drought hazard assessment ...	0.500	yearly
Zhang, B., et al.	Hydr. Process.	2014	Assessing the spatial and ...	0.500	yearly
Zhang, X., et al.	J. Hydrometeorol.	2014	A Long-Term Land Surface ...	0.250	monthly
Zhang, B., et al.	Hydr. Process.	2014	Spatiotemporal analysis of climate ...	0.500	yearly
Zhao, F., et al.	J. Hydrometeorol.	2012	Application of a Macroscale ...	0.050	daily
Zhao, X.,and P. Wu	Natural Hazards	2013	Meteorological drought over ...	0.500	yearly
Zhao, Q., et al.	Env. Earth Sci.	2013	Coupling a glacier melt model ...	0.083	daily
Zhao, F., et al.	J. Hydrol.	2013	The effect of spatial rainfall ...	0.050	daily
Zhu, C., and D.P. Lettenmaier	J. Climate	2007	Long-Term Climate and ...	0.125	monthly
Ziegler, A.D., et al.	J. Climate	2003	Detection of Intensification in ...	2.000	monthly
Ziegler, A.D., et al.	Climatic Change	2005	Detection of time for plausible ...	0.125	monthly

Bibliography

Abbaspour, K., J. Yang, I. Maximov, R. Siber, K. Bogner, J. Mieleitner, J. Zobrist, and R. Srinivasan (2007), Modelling hydrology and water quality in the pre-alpine/alpine Thur watershed using SWAT, *J. Hydrol.*, *333*, 413–430, doi: 10.1016/j.jhydrol.2006.09.014.

Abebe, N., F. Ogden, and N. Pradhan (2010), Sensitivity and uncertainty analysis of the conceptual HBV rainfall-runoff model: Implications for parameter estimation, *J. Hydrol.*, *389*, 301–310, doi: 10.1016/j.jhydrol.2010.06.007.

Adams, R., A. Western, and A. Seed (2012), An analysis of the impact of spatial variability in rainfall on runoff and sediment predictions from a distributed model, *Hydrol. Process.*, *26*, 3263–3280, doi: 10.1002/hyp.8435.

Addor, N., O. Rössler, N. Köplin, M. Huss, R. Weingartner, and J. Seibert (2014), Robust changes and sources of uncertainty in the projected hydrological regimes of Swiss catchments, *Water Resour. Res.*, *50*, 7541–7562, doi: 10.1002/2014WR015549.

Andersen, O., S. Seneviratne, J. Hinderer, and P. Viterbo (2005), GRACE-derived terrestrial water storage depletion associated with the 2003 European heat wave, *Geophys. Res. Lett.*, *32*, L18405, doi: 10.1029/2005GL023574.

Anderson, M., and T. Burt (1990), *Process Studies in Hillslope Hydrology*, chap. 11. Subsurface runoff, pp. 365–400, Wiley & Sons, Chichester, UK.

Andréassian, V., C. Perrin, L. Berthet, N. Le Moine, J. Lerat, C. Loumagne, L. Oudin, T. Mathevet, M. Ramos, and A. Valéry (2009), HESS opinions: Crash tests for a standardized evaluation of hydrological models, *Hydrol. Earth Syst. Sc.*, *13*, 1757–1764, doi: www.hydrol-earth-syst-sci.net/13/1757/2009/.

Arora, V., F. Chiew, and R. Grayson (2001), Effect of sub-grid-scale variability of soil moisture and precipitation intensity on surface runoff and streamflow, *J. Geophys. Res.*, *106*(D15), 17,073–17,091, doi: 10.1029/2001JD900037.

Baker, M., and R. Horton (1936), Historical development of ideas regarding the origin of springs and groundwater, *Transactions, American Geophysical Union*, *17*(2), 395–400, doi: 10.1029/TR017i002p00395.

Baldocchi, D., L. Xua, and N. Kiang (2004), How plant functional-type, weather, seasonal drought, and soil physical properties alter water and energy fluxes of an oak-grass savanna and an annual grassland, *Agr. Forest Meteorol.*, *123*, 13–39, doi: 10.1016/j.agrformet.2003.11.006.

Balmer, N., A. Nevill, and A. Williams (2003), Modelling home advantage in the Summer Olympic Games, *J. Sport Sci.*, *21*, 469–478, doi: 10.1080/0264041031000101890.

Bastiaanssen, W., R. Allen, P. Droogers, G. D'Urso, and P. Steduto (2007), Twenty-five years modeling irrigated and drained soils: State of the art, *Agr. Water Manage.*, *92*, 111–125, doi: 10.1016/j.agwat.2007.05.013.

Becker, A., and U. Grünewald (2003), Flood risk in Central Europe, *Science*, *300*(5622), 1099, doi: 10.1126/science.1083624.

Bennett, J., D. Robertson, P. Ward, H. Hapuarachchi, and Q. Wang (2016), Calibrating hourly rainfall-runoff models with daily forcings for streamflow forecasting applications in meso-scale catchments, *Environ. Modell. Softw.*, *76*, 20–36, doi: 10.1016/j.envsoft.2015.11.006.

Bennett, N., B. Croke, G. Guariso, J. Guillaume, S. Hamilton, A. Jakeman, S. Marsili-Libelli, L. Newham, J. Norton, C. Perrin, S. Pierce, B. Robson, R. Seppelt, A. Voinov, B. Fath, and V. Andréassian (2013), Characterising performance of environmental models, *Environ. Modell. Softw.*, *40*, 1–20, doi: 10.1016/j.envsoft.2012.09.011.

Bense, V., T. Read, and A. Verhoef (2016), Using distributed temperature sensing to monitor field scale dynamics of ground surface temperature and related substrate heat flux, *Agr. Forest Meteorol.*, *220*, 207–215, doi: 10.1016/j.agrformet.2016.01.138.

Berg, A., K. Findell, B. Lintner, A. Giannini, S. I. Seneviratne, B. van den Hurk, R. Lorenz, A. Pitman, S. Hagemann, A. Meier, F. Cheruy, A. Ducharne, S. Malyshev, and P. Milly (2016), Land-atmosphere feedbacks amplify aridity increase over land under global warming, *Nat. Clim. Change*, *6*, 869–874, doi: 10.1038/nclimate3029.

Berghuijs, W., R. Woods, and M. Hrachowitz (2014), A precipitation shift from snow towards rain leads to a decrease in streamflow, *Nat. Clim. Change*, *4*, 583–586, doi: 10.1038/nclimate2246.

Bergström, S. (1976), Development and application of a conceptual runoff model for Scandinavian catchments, *Tech. rep.*, SMHI Report RHO 7, Norrköping.

Bergström, S. (1992), The HBV model - its structure and applications, *Tech. Rep. 4*, SMHI reports hydrology.

Beven, K. (1995), Linking parameters across scales: subgrid parameterizations and scale dependent hydrological models, *Hydrol. Process.*, *9*, 507–525, doi: 10.1002/hyp.3360090504.

Beven, K. (2006a), On undermining the science?, *Hydrol. Process.*, *20*, 3141–3146, doi: 10.1002/hyp.6396.

Beven, K. (2006b), Searching for the Holy Grail of scientific hydrology: Qt=H(S,R,Δt)A as closure, *Hydrol. Earth Syst. Sc.*, *10*, 609–618, doi: 10.5194/hess-10-609-2006.

Beven, K., and A. Binley (1992), The future of distributed models: Model calibration and uncertainty prediction, *Hydrol. Process.*, *6*, 279–298, doi: 10.1002/hyp.3360060305.

Beven, K., and A. Binley (2014), GLUE: 20 years on, *Hydrol. Process.*, *28*, 5897–5918, doi: 10.1002/hyp.10082.

Beven, K., and H. Cloke (2012), Comment on "Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water" by Eric F. Wood et al., *Water Resour. Res.*, *48*, W01801, doi: 10.1029/2011WR010982.

Beven, K., and G. Hornberger (1982), Assessing the effect of spatial pattern of precipitation in modeling stream flow hydrographs, *Water Resour. Bull.*, *18*(5), 823–829, doi: 10.1111/j.1752-1688.1982.tb00078.x.

Beven, K., and I. Westerberg (2011), On red herrings and real herrings: disinformation and information in hydrological inference, *Hydrol. Process.*, *25*, 1676–1680, doi: 10.1002/hyp.7963.

Beven, K., H. Cloke, F. Pappenberger, R. Lamb, and N. Hunter (2015), Hyperresolution information and hyperresolution ignorance in modelling the hydrology of the land surface, *Sc. China: Earth Sc.*, *58*(1), 25–35, doi: 10.1007/s11430-014-5003-4.

Beven, K. J. (2006c), A manifesto for the equifinality thesis, *J. Hydrol.*, *320*, 18–36, doi: 10.1016/j.jhydrol.2005.07.007.

Beven, K. J. (2012), *Rainfall-Runoff modelling, The Primer - 2nd Edition*, vol. Ch.1. Down to Basics: Runoff Processes and the Modelling Process, John Wiley & Sons.

Beven, K. J., and M. J. Kirkby (1979), A physically based, variable contributing area model of basin hydrology, *Hydrol. Sci. Bull.*, *24*, 43–69, doi: 10.1080/02626667909491834.

Bierkens, M., P. Finke, and P. De Willigen (2000), *Upscaling and Downscaling Methods for Environmental Research*, Springer, London, UK.

Bierkens, M., V. Bell, P. Burek, N. Chaney, L. Condon, C. David, A. De Roo, P. Döll, N. Drost, J. Famiglietti, M. Flörke, D. Gochis, P. Houser, R. Hut, J. Keune, S. Kollet, R. Maxwell, J. Reager, L. Samaniego, E. Sudicky, E. Sutanudjaja, N. Van de Giesen, H. Winsemius, and E. Wood (2015), Hyper-resolution global hydrological modelling: What's next?, *Hydrol. Process.*, *29*(2), 310–320, doi: 10.1002/hyp.10391.

Bierkens, M. F. P. (2015), Global hydrology 2015: State, trends, and directions, *Water Resour. Res.*, *51*(7), 4923–4947, doi: 10.1002/2015WR017173.

Biswas, A. (1970), *History of hydrology*, North-Holland Pub. Co., Amsterdam, The Netherlands.

Blöschl, G., and M. Sivapalan (1995), Scale issues in hydrological modelling: a review, *Hydrol. Process.*, *9*, 251–290, doi: 10.1002/hyp.3360090305.

Blöschl, G., M. Sivapalan, T. Wagener, A. Viglione, and H. Savenije (Eds.) (2013), *Runoff Prediction in Ungauged Basins - Synthesis across Processes, Places and Scales*, Cambridge University Press.

Bonell, M., J. J. McDonnell, F. N. Scatena, J. Seibert, S. Uhlenbrook, and H. A. J. van Lanen (2006), HELPing FRIENDs in PUBs: Charting a course for synergies within international water research programmes in gauged and ungauged basins, *Hydrol. Process.*, *20*, 1867–1874, doi: 10.1002/hyp.6196.

Bourdeau, M. (2014), Auguste comte, *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/comte/, Accessed: 13 December 2016.

Boussinesq, J. (1877), Essai sur la théorie des eaux courantes, *Mem. Acad. Sci. Inst. Fr.*, *23*, 252–260.

Boyle, D., H. Gupta, S. Sorooshian, V. Koren, Z. Zhang, and M. Smith (2001), Towards improved streamflow forecasts: Value of semidistributed modeling, *Water Resour. Res.*, *37*(11), 2749–2759, doi: 10.1029/2000WR000207.

Brath, A., A. Montanari, and E. Toth (2004), Analysis of the effects of different scenarios of historical data availability on the calibration of a spatially-distributed hydrological model, *J. Hydrol.*, *291*, 232–253, doi: 10.1016/j.jhydrol.2003.12.044.

Brauer, C. C., A. J. Teuling, P. J. J. F. Torfs, and R. Uijlenhoet (2013), Investigating storage-discharge relations in a lowland catchment using hydrograph fitting, recession analysis and soil moisture data, *Water Resour. Res.*, *49*, 4257–4264, doi: 10.1002/wrcr.20320.

Brutsaert, W. (1988), *Evaporation into the atmosphere - theory, history, and applications*, D. Reidel Publishing Company, Dordrecht, The Netherlands.

Brutsaert, W. (2005), *Hydrology: An Introduction*, Cambridge University Press, Cambridge, UK.

Brutsaert, W., and J. L. Nieber (1977), Regionalized drought flow hydrographs from a mature glaciated plateau, *Water Resour. Res.*, *13*, 637–643, doi: 10.1029/WR013i003p00637.

Burnash, R., R. Ferral, and R. McGuire (1973), A generalized streamflow simulation system - conceptual modeling for digital computers, *Tech. rep.*, U.S. Department of Commerce, National Weather Service and State of California, Department of Water Resources.

Carnap, R. (1966), *Philosophical foundations of physics - an introduction to the philosophy of science*, Basic Books Inc, New York City (NY), USA.

Carpenter, T., and K. Georgakakos (2006), Intercomparison of lumped versus distributed hydrologic model ensemble simulations on operational forecast scales, *J. Hydrol.*, *329*, 174–185, doi: 10.1016/j.jhydrol.2006.02.013.

Chalmers, A. (1999), *What Is This Thing Called Science? An Assessment of the Nature and Status of Science and Its Methods*, Open University Press, Maidenhead, UK.

Chamberlin, T. (1890), The method of multiple working hypotheses, *Science (old series)*, *15*(92), available at: https://pdfs.semanticscholar.org/5502/5bd7729f1de950572f896ad3bf0e6b2b0984.pdf.

Chaney, N., J. Herman, P. Reed, and E. Wood (2015), Flood and drought hydrologic monitoring: the role of model parameter uncertainty, *Hydrol. Earth Syst. Sc.*, *19*, 3239–3251, doi: 10.5194/hess-19-3239-2015.

Clark, M., and D. Kavetski (2010), Ancient numerical daemons of conceptual hydrological modeling: 1. fidelity and efficiency of time stepping schemes, *Water Resour. Res.*, *46*, W10510, doi: 10.1029/2009WR008894.

Clark, M., D. Kavetski, and F. Fenicia (2011a), Pursuing the method of multiple working hypotheses for hydrological modeling, *Water Resour. Res.*, *47*, W09301, doi: 10.1029/2010WR009827.

Clark, M., H. McMillan, D. Collins, D. Kavetski, and R. Woods (2011b), Hydrological field data from a modeller's perspective: Part 2: process-based evaluation of model hypotheses, *Hydrol. Process.*, *25*, 523–543, doi: 10.1002/hyp.7902.

Clark, M., B. Nijssen, J. Lundquist, D. Kavetski, D. Rupp, R. Woods, J. Freer, E. Gutmann, A. Wood, L. D. Brekke, J. Arnold, D. Gochis, and R. Rasmussen (2015), A unified approach for process-based hydrologic modeling: 1. modeling concept, *Water Resour. Res.*, *51*, 2498–2514, doi: 10.1002/2015WR017198.

Clark, M., B. Schaefli, S. Schymanski, L. Samaniego, C. Luce, B. Jackson, J. Freer, J. Arnold, R. Moore, E. Istanbulluoglu, and S. Ceola (2016a), Improving the theoretical underpinnings of process-based hydrologic models, *Water Resour. Res.*, *52*, doi: 10.1002/2015WR017910.

Clark, M., R. Wilby, E. Gutmann, J. Vano, S. Gangopadhyay, A. Wood1, H. Fowler, C. Prudhomme, J. Arnold, and L. Brekke (2016b), Characterizing uncertainty of the hydrologic impacts of climate change, *Curr. Clim. Change Rep.*, *2*(2), 55–64, doi: 10.1007/s40641-016-0034-x.

Coron, L., V. Andréassian, C. Perrin, J. Lerat, J. Vaze, M. Bourqui, and F. Hendrickx (2012), Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments, *Water Resour. Res.*, *48*, W05552, doi: 10.1029/2011WR011721.

Cosgrove, B. A., D. Lohmann, K. Mitchell, P. Houser, E. Wood, J. Schaake, A. Robock, J. Sheffield, Q. Duan, L. Luo, R. W. Higgins, R. Pinker, and J. Tarpley (2003), Land surface model spin-up behavior in the North American

Land Data Assimilation System (NLDAS), *J. Geophys. Res.*, *108*, D22, doi: 10.1029/2002JD003316.

Costa-Cabral, M., S. Roy, E. Maurer, W. Mills, and L. Chen (2013), Snowpack and runoff response to climate change in Owens Valley and Mono Lake watersheds, *Climatic Change*, *116*, 97–109, doi: 10.1007/s10584-012-0529-y.

Crout, N., D. Tarsitano, and A. Wood (2009), Is my model too complex? evaluating model formulation using model reduction, *Environ. Modell. Softw.*, *24*, 1–7, doi: 10.1016/j.envsoft.2008.06.004.

Dai, A., K. Trenberth, and T. Qian (2004), A global data set of Palmer Drought Severity Index for 1870-2002: relationship with soil moisture and effects of surface warming, *J. Hydrometeorol.*, *5*, 117–1130, doi: 10.1175/JHM-386.1.

Demaria, E. M., B. Nijssen, and T. Wagener (2007), Monte Carlo sensitivity analysis of land surface parameters using the Variable Infiltration Capacity model, *J. Geophys. Res.*, *112*, D11113, doi: 10.1029/2006JD007534.

Dooge, J. (1959), Quantitative hydrology in the 17th century, *Houille Blanche*, *6*, 799–807, doi: 10.1051/lhb/1959063.

Dooge, J. (1986), *Reflections in Hydrology: Science and Practice - Scale problems in hydrology, Kiesel Memorial Lecture*, pp. 85–145, American Geophysical Union, Washington D.C., USA.

Dooge, J. (1988), Hydrology in perspective, *Hydrolog. Sci. J.*, *33*(1), 61–85, doi: 10.1080/02626668809491223.

Duan, Q., J. Schaake, V. Andréassian, S. Franks, G. Goteti, H. Gupta, Y. Gusev, F. Habets, A. Hall, L. Hay, T. Hogue, M. Huang, G. Leavesley, X. Liang, O. Nasonova, J. Noilhan, L. Oudin, S. Sorooshian, T. Wagener, and E. Wood (2006), Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, *J. Hydrol.*, *320*, 3–17, doi: 10.1016/j.jhydrol.2005.07.031.

Dunne, T. (1978), *Hillslope Hydrology*, chap. 7. Field studies of hillslope flow processes, pp. 227–293, Wiley & Sons, Chichester, UK.

Eagleson, P. (1991), Hydrologic science: A distinct geoscience, *Reviews of Geophysics*, *29*(2), 237–248, doi: 10.1029/90RG02615.

Easterbrook, S. (2014), Open code for open science?, *Nat. Geosci.*, *7*, 779–781, doi: 10.1038/ngeo2283.

Ettema, J., M. van den Broeke, E. van Meijgaard, W. J. van de Berg, J. Bamber, J. Box, and R. Bales (2009), Higher surface mass balance of the Greenland ice sheet revealed by high-resolution climate modeling, *Geophys. Res. Lett.*, *36*, L12501, doi: 10.1029/2009GL038110.

Famiglietti, J., M. Lo, S. Ho, J. Bethune, K. Anderson, T. Syed, S. Swenson, C. de Linage, and M. Rodell (2011), Satellites measure recent rates of groundwater depletion in California's Central Valley, *Geophys. Res. Lett.*, *38*(3), doi: 10.1029/2010GL046442.

FAO, IIASA, ISRIC, ISSCAS, and JRC (2012), Harmonized world soil database (version 1.2), *Tech. rep.*, AO, Rome, Italy and IIASA, Laxenburg, Austria.

Fearn, N. (2001), *Zeno and the Tortoise: How to think like a philosopher*, Atlantic Books, London, UK.

Feddes, R. (Ed.) (1995), *Space and Time scale variability and interdependencies in hydrological processes*, Cambridge University Press, Cambridge, UK.

Fenicia, F., D. Kavetski, H. Savenije, and L. Pfister (2016), From spatially variable streamflow to distributed hydrological models: Analysis of key modeling decisions, *Water Resour. Res.*, *52*, 1–35, doi: 10.1002/2015WR017398.

Ficchì, A., C. Perrin, and V. Andréassian (2016), Impact of temporal resolution of inputs on hydrological model performance: An analysis based on 2400 flood events, *J. Hydrol.*, *538*, 454–470, doi: 10.1016/j.jhydrol.2016.04.016.

Finnerty, B., M. Smith, D. Sea, V. Koren, and G. Moglen (1997), Space-time scale sensitivity of the Sacramento model to radar-gage precipitation inputs, *J. Hydrol.*, *203*, 21–38, doi: 10.1016/S0022-1694(97)00083-8.

Foresight Committee on Hydrological Science in The Netherlands (2005), *Turning the Water Wheel Inside Out - Foresight Study on Hydrological Science in The Netherlands*, Royal Netherlands Academy of Arts and Sciences (KNAW), Amsterdam, The Netherlands.

Fortak, H. (Ed.) (1982), *Meteorologie*, Dietrich Reimer, Berlin, Germany.

Fowler, K., M. Peel, A. Western, L. Zhang, and T. Peterson (2016), Simulating runoff under changing climatic conditions: Revisiting an apparent deficiency of conceptual rainfall-runoff models, *Water Resour. Res.*, *52*(3), 1820–1846, doi: 10.1002/2015WR018068.

Francini, M., and M. Pacciani (1991), Comparative analysis of several conceptual rainfall-runoff models, *J. Hy-*

drol., *122*, 161–219, doi: 10.1016/0022-1694(91)90178-K.

Freni, G., G. Mannina, and G. Viviani (2009), Urban runoff modelling uncertainty: Comparison among Bayesian and pseudo-Bayesian methods, *Environ. Modell. Softw.*, pp. 1100–1111, doi: 10.1016/j.envsoft.2009.03.003.

Frigg, R., and S. Hartmann (2012), Models in science, *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/models-science/, Accessed: 4 November 2016.

Fundel, F., and M. Zappa (2011), Hydrological ensemble forecasting in mesoscale catchments: Sensitivity to initial conditions and value of reforecasts, *Water Resour. Res.*, *47*(W09520), doi: 10.1029/2010WR009996.

Fundel, F., S. Jörg-Hess, and M. Zappa (2013), Monthly hydrometeorological ensemble prediction of streamflow droughts and corresponding drought indices, *Hydrol. Earth Syst. Sc.*, *17*, 395–407, doi: 10.5194/hess-17-395-2013.

Gan, T. Y., E. M. Dlamini, and G. F. Biftu (1997), Effects of model complexity and structure, data quality, and objective functions on hydrologic modeling, *J. Hydrol.*, *192*, 81–103, doi: 10.1016/S0022-1694(96)03114-9.

Gelman, A., and D. Rubin (1992), Inference from iterative simulation using multiple sequences, *Stat. Sci.*, *7*(4), 457–472, doi: http://www.jstor.org/stable/2246093.

Gentine, P., T. Troy, B. Lintner, and K. Findell (2012), Scaling in surface hydrology: Progress and challenges, *J. Contemp. Water Res. Educ.*, *147*, 28–40, doi: 10.1111/j.1936-704X.2012.03105.x.

Gleeson, T., Y. Wada, and M. P. B. L. H. van Beek (2012), Water balance of global aquifers revealed by groundwater footprint, *Nature*, *488*, 197–200, doi: 10.1038/nature11295.

Gottschalk, L. (1992), *Advances in Theoretical Hydrology: A Tribute to James Dooge*, chap. 11. Quantitative hydrology in Scandinavia in the 18th century, pp. 177–189, Elsevier Amsterdam, The Netherlands.

Gray, W., A. Leijnse, R. Kolar, and C. Blain (1993), *Mathematical Tools for Changing Scale in the Analysis of Physical Systems*, CRC Press, Boca Raton, USA.

Grayson, R., and G. G. Blöschl (Eds.) (2000), *Spatial patterns in Catchment Hydrology*, Cambridge University Press, Cambridge, UK.

Green, W., and G. Ampt (1911), Studies of soil physics, part I - the flow of water and air through soils, *J. Agr. Sci.*, *4*, 1–24, doi: 10.1017/S0021859600001441.

Guillod, B., B. Orlowsky, D. Miralles, A. Teuling, and S. Seneviratne (2015), Reconciling spatial and temporal soil moisture effects on afternoon rainfall, *Nat. Comm.*, *6*(6443), doi: 10.1038/ncomms7443.

Gupta, H., and G. Nearing (2014), Debates - The future of hydrological sciences: A (common) path forward? Using models and data to learn: A systems theoretic perspective on the future of hydrological science, *Water Resour. Res.*, *50*, 5351–5359, doi: 10.1002/2013WR015096.

Gupta, H., S. Sorooshian, and P. Yapo (1998), Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, *Water Resour. Res.*, *34*(4), 751–763, doi: 10.1029/97WR03495.

Gupta, H., T. Wagener, and Y. Liu (2008), Reconciling theory with observations: elements of a diagnostic approach to model evaluation, *Hydrol. Process.*, *22*, 3802–3813, doi: 10.1002/hyp.6989.

Gupta, H., H. Kling, K. Yilmaz, and G. Martinez (2009), Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *J. Hydrol.*, *377*, 80–91, doi: 10.1016/j.jhydrol.2009.08.003.

Gupta, H., C. Perrin, G. Blöschl, A. Montanari, R. Kumar, M. Clark, and V. Andréassian (2014), Large-sample hydrology: a need to balance depth with breadth, *Hydrol. Earth Syst. Sc.*, *18*, 463–477, doi: 10.5194/hess-18-463-2014.

Gupta, V., I. Rodríguez-Iturber, and E. Wood (Eds.) (1986), *Scale problems in hydrology*, D. Reidel Publishing Company, Dordrecht, The Netherlands.

Gurtz, J., A. Baltensweiler, and H. Lang (1999), Spatially distributed hydrotope-based modelling of evapotranspiration and runoff in mountainous basins, *Hydrol. Process.*, *13*, 2751–2768, doi: 10.1002/(SICI)1099-1085(19991215)13:17<2751::AID-HYP897>3.0.CO;2-O.

Gurtz, J., M. Verbunt, M. Zappa, M. Moesch, F. Pos, and U. Moser (2003), Long-term hydrometeorological measurements and model-based analyses in the hydrological research catchment Rietholzbach, *J. Hydrol. Hydromech.*, *51*, 162–174.

Gutmann, E., T. Pruitt, M. Clark, L. Brekke, J. Arnold, D. Raff, and R. Rasmussen (2014), An intercomparison of

statistical downscaling methods used for water resource assessments in the United States, *Water Resour. Res.*, *50*, 7167–7186, doi: $10.1002/2014WR015559$.

Haddeland, I., B. Matheussen, and D. Lettenmaier (2002), Influence of spatial resolution on simulated stream-flow in a macroscale hydrologic model, *Water Resour. Res.*, *38*(7), 1124, doi: $10.1029/2001WR000854$.

Haddeland, I., D. Lettenmaier, and T. Skaugen (2006), Reconciling simulated moisture fluxes resulting from alternate hydrologic model time steps and energy budget closure assumptions, *J. Hydrometeorol.*, *7*, 355–370, doi: $10.1175/JHM496.1$.

Harman, C., and P. Troch (2014), What makes Darwinian hydrology "Darwinian"? Asking a different kind of question about landscapes, *Hydrol. Earth Syst. Sc.*, *18*, 417–433, doi: $10.5194/hess-18-417-2014$.

Harte, J. (2002), Toward a synthesis of Newtonian and Darwinian worldviews, *Phys. Today*, *55*(10), 29–34, doi: $10.1063/1.1522164$.

Hazenberg, P., Y. Fang, P. Broxton, D. Gochis, G. Niu, J. Pelletier, P. Troch, and X. Zeng (2015), A hybrid-3D hill-slope hydrological model for use in Earth system models., *Water Resour. Res.*, *10*, 8218–8239, doi: $10.1002/2014WR016842$.

Hisdal, H., L. M. Tallaksen, B. Clausen, E. Peters, and A. Gustard (2004), *Hydrological Drought - processes and estimation methods for streamflow and groundwater*, chap. 5. Hydrological Drought Characteristics, pp. 139–198, Elsevier Amsterdam, The Netherlands.

Hoekstra, A. (2014), Water scarcity challenges to business, *Nat. Clim. Change*, *4*(5), 318–320, doi: $10.1038/nclimate2214$.

Horton, R. (1931), The field, scope, and status of the science of hydrology, *Transactions, American Geophysical Union*, *12*(1), 189–202, doi: $10.1029/TR012i001p00189-2$.

Hurkmans, R. T. W. L., H. de Moel, J. C. J. H. Aerts, and P. A. Troch (2008), Water balance versus land surface model in the simulation of Rhine river discharges, *Water Resour. Res.*, *44*, W01418, doi: $10.1029/2007WR006168$.

Hutton, C., T. Wagener, J. Freer, D. Han, C. Duffy, and B. Arheimer (2016), Most computational hydrology is not reproducible, so is it really science?, *Water Resour. Res.*, *52*, 7548–7555, doi: $10.1002/2016WR019285$.

Iwema, J., R. Rosolem, R. Baatz, T. Wagener, and H. Bogena (2015), Investigating temporal field sampling strategies for site-specific calibration of three soil moisture-neutron intensity parameterisation methods, *Hydrol. Earth Syst. Sc.*, *19*(7), 3203–3216, doi: $10.5194/hess-19-3203-2015$.

Jakeman, A., and G. Hornberger (1993), How much complexity is warranted in a rainfall-runoff model?, *Water Resour. Res.*, *29*(8), 2637–2649, doi: $10.1029/93WR00877$.

Jasper, K., P. Calanca, D. Gyalistras, and J. Fuhrer (2004), Differential impacts of climate change on the hydrology of two Alpine river basins, *Clim. Res.*, *26*, 113–129, doi: $10.3354/cr026113$.

Jörg-Hess, S., S. Kempf, F. Fundel, and M. Zappa (2015), The benefit of climatological and calibrated reforecast data for simulating hydrological droughts in Switzerland, *Met. App.*, *22*, 444–458, doi: $10.1002/met.1474$.

Joussaume, S., K.E.Taylor, P. Braconnot, J.F.B.Mitchell, J.E.Kutzbach, S.P.Harrison, I.C.Prentice, A.J.Broccoli, A. Abe-Ouchi, P.J.Bartlein, C. Bonfils, B. Dong, J.Guiot, K. Henerich, C. D. Hewitt, D. Jolly, J. Kim, A. Kislov, A. Kitoh, M. F.Loutre, V. Masson, B.McAvaney, N. McFarlane, N. de Noblet, W.R.Peltier, J. Peterschmitt, D.Pollard, D. Rind, J. F. Royer, M. E. Schlesinger, J. Syktus, S. Thompson, P.Valdes, R.Vettoretti, R.S.Webb, and U. Wyputta (1999), Monsoon changes for 6000 years ago: Results of 18 simulations from the Paleoclimate Modeling Intercomparison Project (PMIP), *Geophys. Res. Lett.*, *26*(7), 859–862, doi: $10.1029/1999GL900126$.

Kalma, J., and M. Sivapalan (Eds.) (1995), *Advances in Hydrological Processes - Scale issues in Hydrological modelling*, Wiley & Sons, New York, USA.

Kampf, S., and S. Burges (2007), A framework for classifying and comparing distributed hillslope and catchment hydrologic models, *Water Resour. Res.*, *43*, W05423, doi: $10.1029/2006WR005370$.

Kauffeldt, A., S. Halldin, A. Rodhe, C.-Y. Xu, and I. Westerberg (2013), Disinformative data in large-scale hydrological modelling, *Hydrol. Earth Syst. Sc.*, *17*, 2845–2857, doi: $10.5194/hess-17-2845-2013$.

Kavetski, D., and M. Clark (2010), Ancient numerical daemons of conceptual hydrological modeling: 2. Impact of time stepping schemes on model analysis and prediction, *Water Resour. Res.*, *46*, W10511, doi: $10.1029/2009WR008896$.

Kavetski, D., F. Fenicia, and M. P. Clark (2011), Impact of temporal data resolution on parameter inference and model identification in conceptual hydrological modeling: Insights from an experimental catchment, *Water*

Resour. Res., 47, W05501, doi: 10.1029/2010WR009525.

Kileshye Onema, J. M., A. E. Taigbenu, and J. Ndiritu (2012), Classification and flow prediction in a data-scarce watershed of the equatorial Nile region, Hydrol. Earth. Syst. Sci., 16, 1435–1443, doi: 10.5194/hess-6-1435-2012.

Kim, C. P., and J. N. M. Stricker (1996), Influence of spatially variable soil hydraulic properties and rainfall intensity on the water budget, Water Resour. Res., 32(6), 1699–1712, doi: 10.1029/96WR00603.

Kim, C. P., J. N. M. Stricker, and R. A. Feddes (1997), Impact of soil heterogeneity on the water budget of the unsaturated zone, Water Resour. Res., 33(5), 991–999, doi: 10.1029/97WR00364.

Kirchner, J. (2006), Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, Water Resour. Res., 42, W03S04, doi: 10.1029/2005WR004362.

Kirchner, J. W. (2009), Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward, Water Resour. Res., 45, W02429, doi: 10.1029/2008WR006912.

Klemeš, V. (1983), Conceptualization and scale in hydrology, J. Hydrol., 65, 1–23, doi: 10.1016/0022-1694(83)90208-1.

Klemeš, V. (1986a), Dilettantism in hydrology: Transition or destiny?, Water Resour. Res., 22(9), 177S–188S, doi: 10.1029/WR022i09Sp0177S.

Klemeš, V. (1986b), Operational testing of hydrological simulation models, Hydrolog. Sci. J., 31(1), 13–24, doi: 10.1080/02626668609491024.

Knutti, R., and J. Sedláček (2013), Robustness and uncertainties in the new CMIP5 climate model projections, Nat. Clim. Change, 3, 369–373, doi: 10.1038/NCLIMATE1716.

Knutti, R., D. Masson, and A. Gettelman (2013), Climate model genealogy: Generation CMIP5 and how we got there, Geophys. Res. Lett., 40, 1194–1199, doi: 10.1002/grl.50256.

König, P., and H. Lang (1994), On the runoff formation in the small pre-alpine research basin Rietholzbach, FRIEND: Flow Regimes from International Experimental and Network Data (Proceedings of the Braunschweig Conference October 1993), 221, 391–398.

Koskela, J. J., B. W. F. Croke, H. Koivusalo, A. J. Jakeman, and T. Kokkonen (2012), Bayesian inference of uncertainties in precipitation-streamflow modeling in a snow affected catchment, Water Resour. Res., 48, W11513, doi: 10.1029/2011WR011773.

Koster, R., S. Mahanama, B. Livneh, D. Lettenmaier, and R. Reichle (2010), Skill in streamflow forecasts derived from large-scale estimates of soil moisture and snow, Nat. Geosci., 3, 613–616, doi: 10.1038/NGEO944.

Krause, P., D. P. Boyle, and F. Bäse (2005), Comparison of different efficiency criteria for hydrological model assessment, Adv. Geosci., 5, 89–97, doi: 10.5194/adgeo-5-89-2005.

Krueger, T., T. Page, K. Hubacek, L. Smith, and K. Hiscock (2012), The role of expert opinion in environmental modelling, Environ. Modell. Softw., 36, 4–18, doi: 10.1016/j.envsoft.2012.01.011.

Kumar, R., L. Samaniego, and S. Attinger (2013), Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations, Water Resour. Res., 49, 360–379, doi: 10.1029/2012WR012195.

Kustas, W. P., A. Rango, and R. Uijlenhoet (1994), A simple energy budget algorithm for the snowmelt runoff model, Water Resour. Res., 30(5), 1515–1527, doi: 10.1029/94WR00152.

Lahoz, W., and G. M. De Lannoy (2014), Closing the gaps in our knowledge of the hydrological cycle over land: Conceptual problems, Surv. Geophys., 35, 623–660, doi: 10.1007/s10712-013-9221-7.

Lambert, A. (1972), Catchment models based on ISO-functions, J. Inst. Water Eng., 26(8), 413–422.

Lehner, I., and S. I. Seneviratne (2010), Mosnang-Rietholz - Rietholzbach basin, Switzerland, IAC-ETH Poster.

Lhomme, J. (1997), An examination of the Priestley-Taylor equation using a convective boundary layer model, Water Resour. Res., 33(11), 2571–2578, doi: 10.1029/97WR01897.

Li, C. Z., L. Zhang, H. Wang, Y. Q. Zhang, F. L. Yu, and D. H. Yan (2012), The transferability of hydrological models under nonstationary climatic conditions, Hydrol. Earth Syst. Sc., 16, 1239–1254, doi: 10.5194/hess-16-1239-2012.

Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J. Burges (1994), A simple hydrologically based model of land surface water and energy fluxes for general circulation models, J. Geophys. Res., 99(D7), 14,415–14,458, doi: 10.1029/94JD00483.

Liang, X., E. F. Wood, and D. P. Lettenmaier (1996), Surface soil moisture parameterization of the VIC-2L model: Evaluation and modification, *Global Planet. Change*, *13*, 195–206, doi: 10.1016/0921-8181(95)00046-1.

Liang, X., J. Guo, and L. Leung (2004), Assessment of the effects of spatial resolutions on daily water flux simulations, *J. Hydrol.*, *298*, 287–310, doi: 10.1016/j.jhydrol.2003.07.007.

Littlewood, I., and B. Croke (2008), Data time-step dependency of conceptual rainfall-streamflow model parameters: an empirical study with implications for regionalisation, *Hydrolog. Sci. J.*, *53*(4), 685–695, doi: 10.1623/hysj.53.4.685.

Littlewood, I., and B. Croke (2013), Effects of data time-step on the accuracy of calibrated rainfall-streamflow model parameters: practical aspects of uncertainty reduction, *Hydrolog. Res.*, *44*(3), 430–440, doi: 10.2166/nh.2012.099.

Liu, H., F. Tian, H. Hu, H. Hu, and M. Sivapalan (2013), Soil moisture controls on patterns of grass green-up in Inner Mongolia: an index based approach, *Hydrol. Earth Syst. Sc.*, *17*, 805–815, doi: 10.5194/hess-17-805-2013.

Liu, Y., and H. V. Gupta (2007), Uncertainty in hydrologic modeling: Towards an integrated data assimilation framework, *Water Resour. Res.*, *43*, W07401, doi: 10.1029/2006WR005756.

Lobligeois, F., V. Andréassian, C. Perrin, P. Tabary, and C. Loumagne (2014), When does higher spatial resolution rainfall information improve streamflow simulation? An evaluation using 3620 flood events, *Hydrol. Earth Syst. Sc.*, *18*, 575–594, doi: 10.5194/hess-18-575-2014.

Lohmann, D., R. Nolte-Holube, and E. Raschke (1996), A large-scale horizontal routing model to be coupled to land surface parameterization schemes, *Tellus*, *48A*, 708–721, doi: 10.1034/j.1600-0870.1996.t01-3-00009.x.

Maier, H., and G. Dandy (2000), Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications, *Environ. Modell. Softw.*, *15*, 101–124, doi: 10.1016/S1364-8152(99)00007-9.

Maurer, E., A. Wood, J. Adam, D. Lettenmaier, and B. Nijssen (2002), A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States, *J. Climate*, *15*, 3237–3251, doi: http://dx.doi.org/10.1175/JCLI-D-12-00508.1.

Maxwell, R., and S. Kollet (2008), Quantifying the effects of three-dimensional subsurface heterogeneity on Hortonian runoff processes using a coupled numerical, stochastic approach, *Adv. Water Resour.*, *31*(5), 807–817, doi: 10.1016/j.advwatres.2008.01.020.

McDonnell, J., and K. Beven (2014), Debates - the future of hydrological sciences: A (common) path forward? a call to action aimed at understanding velocities, celerities and residence time distributions of the headwater hydrograph, *Water Resour. Res.*, *50*, 5342–5350, doi: 10.1002/2013WR015141.

McDonnell, J., M. Sivapalan, K. Vaché, S. Dunn, G. Grant, R. Haggerty, C. Hinz, R. Hooper, J. Kirchner, M. L. Roderick, J. Selker, and M. Weiler (2007), Moving beyond heterogeneity and process complexity: A new vision for watershed hydrology, *Water Resour. Res.*, *43*, W07301, doi: 10.1029/2006WR005467.

McIntyre, N. R., and H. S. Wheater (2004), Calibration of an in-river phosphorus model: prior evaluation of data needs and model uncertainty, *J. Hydrol.*, *290*, 100–116, doi: 10.1016/j.jhydrol.2003.12.003.

Melsen, L., P. Torfs, R. Uijlenhoet, and A. Teuling (2017), Comment on "Most computational hydrology is not reproducible, so is it really science?", *Water Resour. Res.*, *accepted*.

Mendoza, P., M. Clark, N. Mizukami, E. Gutmann, J. Arnold, L. Brekke, and B. Rajagopalan (2015a), How do hydrologic modeling decisions affect the portrayal of climate change impacts?, *Hydrol. Process.*, *30*(7), 1071–1095, doi: 10.1002/hyp.10684.

Mendoza, P., M. Clark, N. Mizukami, A. Newman, M. Barlage, E. Gutmann, R. Rasmussen, B. Rajagopalan, L. Brekke, and J. Arnold (2015b), Effects of hydrologic model choice and calibration on the portrayal of climate change impacts, *J. Hydrometeorol.*, *16*, 762–780, doi: 10.1175/JHM-D-14-0104.1.

Mendoza, P. A., M. Clark, M. Barlage, B. Rajagopalan, L. Samaniego, G. Abramowitz, and H. Gupta (2015c), Are we unnecessarily constraining the agility of complex process-based models?, *Water Resour. Res.*, *51*, 716–728, doi: 10.1002/2014WR015820.

Merz, R., J. Parajka, and G. Blöschl (2011), Time stability of catchment model parameters: Implications for climate impact analyses, *Water Resour. Res.*, *47*, W02531, doi: 10.1029/2010WR009505.

Miller, E., and R. Miller (1956), Physical theory for capillary flow phenomena, *J. Appl. Phys.*, *27*, 324–332, doi:

10.1063/1.1722370.

Milly, P., K. Dunne, and A. Vecchia (2005), Global pattern of trends in streamflow and water availability in a changing climate, *Nature*, *438*, 347–350, doi: 10.1038/nature04312.

Milly, P., J. Betancourt, M. Falkenmark, R. Hirsch, Z. Kundzewicz, D. Lettenmaier, and R. Stouffer (2008), Stationarity is dead: Whither water management?, *Science*, *319*, 573–574, doi: 10.1126/science.1151915.

Miralles, D., A. Teuling, C. van Heerwaarden, and J. Vilà-Guerau de Arellano (2014), Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation, *Nat. Geosci.*, *7*, 345–349, doi: 10.1038/ngeo2141.

Mithen, S., and E. Black (Eds.) (2011), *Water, Life and Civilisation - Climate, Environment and Society in the Jordan Valley*, Cambridge University Press, Cambridge, UK.

Mizukami, N., M. Clark, E. Gutmann, P. Mendoza, A. Newman, B. Nijssen, B. Livneh, L. Hay, J. Arnold, and L. Brekke (2015), Implications of the methodological choices for hydrologic portrayals of climate change over the contiguous United States: statistically downscaled forcing data and hydrologic models, *J. Hydrometeorol.*, *17*, 73–98, doi: 10.1175/JHM-D-14-0187.1.

Mizukami, N., M. P. Clark, K. Sampson, B. Nijssen, Y. Mao, H. McMillan, R. J. Viger, S. L. Markstrom, L. E. Hay, R. Woods, J. R. Arnold, and L. D. Brekke (2016), mizuRoute version 1: a river network routing tool for a continental domain water resources applications, *Geosci. Model Dev.*, *9*, 2223–2238, doi: 10.5194/gmd-9-2223-2016.

Mohanty, B., and T. Skaggs (2001), Spatio-temporal evolution and time-stable characteristics of soil moisture within remote sensing footprints with varying soil, slope, and vegetation, *Adv. Water Resour.*, *24*, 1051–1067, doi: 10.1016/S0309-1708(01)00034-3.

Montaldo, N., and J. Albertson (2003), Temporal dynamics of soil moisture variability: 2. Implications for land surface models, *Water Resour. Res.*, *39*(10), 1275, doi: 10.1029/2002WR001618.

Moore, G. E. (1965), Cramming more components onto integrated circuits, *Electronics*, *April*, 114–117.

Mulvany, T. (1851), On the use of self-registering rain and flood gauges in making observations of the relations of rainfall and of flood discharges in a given catchment, *Proc. Inst. Civil Eng. Ireland*, *4*(2), 18–33.

Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models: Part 1. A discussion of principles, *J. Hydrol.*, *10*, 282–290, doi: 10.1016/0022-1694(70)90255-6.

National Research Council (1991), *Opportunities in the Hydrologic sciences*, National Academy Press, Washington, USA.

National Weather Service (2002), II.3-SAC-SMA conceptualization of the sacramento soil moisture accounting model, *Tech. rep.*, National Oceanic and Atmospheric Administration (NOAA).

Neal, J., T. Fewtrell, P. Bates, and N. Wright (2010), A comparison of three parallelisation methods for 2D flood inundation models, *Environ. Modell. Softw.*, *25*, 398–411, doi: 10.1016/j.envsoft.2009.11.007.

Nearing, G., Y. Tian, H. Gupta, M. Clark, K. Harrison, and S. Weijse (2016), A philosophical basis for hydrological uncertainty, *Hydr. Sci. J.*, *61*(9), 1666–1678, doi: 10.1080/02626667.2016.1183009.

Newman, A., M. Clark, K. Sampson, A. Wood, L. Hay, A. Bock, R. Viger, D. Blodgett, L. Brekke, J. Arnold, T. Hopson, and Q. Duan (2015), Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance, *Hydrol. Earth Syst. Sc.*, *19*, 209–223, doi: 10.5194/hess-19-209-2015.

Nicótina, L., E. Alessi Celegon, A. Rinaldo, and M. Marani (2008), On the impact of rainfall patterns on the hydrologic response, *Water Resour. Res.*, *44*, W12401, doi: 10.1029/2007WR006654.

Nijssen, B., G. O'Donnell, and D. Lettenmaier (2001), Predicting the discharge of global rivers, *J. Climate*, *14*, 3307–3323, doi: 10.1175/1520-0442(2001)014<3307:PTDOGR>2.0.CO;2.

Obled, C., J. Wendling, and K. Beven (1994), The sensitivity of hydrological models to spatial rainfall patterns: an evaluation using observed data, *J. Hydrol.*, *159*, 305–333, doi: 10.1016/0022-1694(94)90263-1.

O'Grady, P. (2016), Thales of Miletus (c. 620 B.C.E.-c. 546 B.C.E.), *The Internet Encyclopedia of Philosophy*, http://www.iep.utm.edu/, Accessed: Tuesday 31 May 2016.

Oreskes, N., K. Shrader-Frechette, and K. Belitz (1994), Verification, validation, and confirmation of numerical models in the Earth Sciences, *Science*, *263*(5147), 641–646, doi: 10.1126/science.263.5147.641.

Orlanski, I. (1975), A rational subdivision of scales for atmospheric processes, *Bull. Am. Meteorol. Soc.*, *56*, 527–530.

Orth, R., and S. I. Seneviratne (2013), Predictability of soil moisture and streamflow on sub-seasonal timescales: A case study, *J. Geophys. Res.*, *118*(19), 10,963–10,979, doi: 10.1002/jgrd.50846.

Ott, R., and M. Longnecker (2010), *An introduction to statistical methods and data analysis*, sixth edition ed., Brooks/Cole, Cengage Learning.

Pan, M., E. Wood, R. Wójcik, and M. McCabe (2008), Estimation of regional terrestrial water cycle using multi-sensor remote sensing observations and data assimilation, *Remote Sens. Environ.*, *112*, 1282–1294, doi: 10.1016/j.rse.2007.02.039.

Paniconi, C., and M. Putti (2015), Physically based modeling in catchment hydrology at 50: Survey and outlook, *Water Resour. Res.*, *51*, 7090–7129, doi: 10.1002/2015WR017780.

Paola, C., E. Foufoula-Georgiou, W. Dietrich, M. Hondzo, D. Mohrig, G. Parker, M. Power, I. Rodriguez-Iturbe, V. Voller, and P. Wilcock (2006), Toward a unified science of the Earth's surface: Opportunities for synthesis among hydrology, geomorphology, geochemistry, and ecology, *Water Resour. Res.*, *42*, doi: 10.1029/2005WR004336, w03S10.

Parajka, J., R. Merz, and G. Blöschl (2007), Uncertainty and multiple objective calibration in regional water balance modelling: case study in 320 Austrian catchments, *Hydrol. Process.*, *21*, 435–446, doi: 10.1002/hyp.6253.

Perrin, C., L. Oudin, V. Andréassian, C. Rojas-Serna, C. Michel, and T. Mathevet (2007), Impact of limited streamflow data on the efficiency and the parameters of rainfall-runoff models, *Hydrolog. Sci. J.*, *52*, 131–151, doi: 10.1623/hysj.52.1.131.

Pfister, L., and H. Savenije (2006), Leonardo da Vinci's scriptures as benchmark papers in hydrology, *Hydrol. Process.*, *20*, 1653–1655, doi: 10.1002/hyp.6207.

Popper, K. (1968), *The logic of scientific discovery*, Hutchinson of London, London, UK.

Psillos, S. (2011), Living with the abstract: realism and models, *Synthese*, *180*, 3–17, doi: 10.1007/s11229-009-9563-3.

Rakovec, O., M. C. Hill, M. P. Clark, A. H. Weerts, A. J. Teuling, and R. Uijlenhoet (2014), Distributed evaluation of local sensitivity analysis (DELSA), with application to hydrologic models, *Water Resour. Res.*, *50*, 409–426, doi: 10.1002/2013WR014063.

Rakovec, O., R. Kumar, S. Attinger, and L. Samaniego (2016), Improving the realism of hydrologic model functioning through multivariate parameter estimation, *Water Resour. Res.*, *52*(10), 7779–7792, doi: 10.1002/2016WR019430.

Rebetez, M., H. Mayer, O. Dupont, D. Schindler, K. Gartner, J. Kropp, and A. Menzel (2006), Heat and drought 2003 in Europe: a climate synthesis, *Ann. For. Sci.*, *63*, 569–577, doi: 10.1051/forest:2006043.

Reed, R. D., and R. J. Marks II (1999), *Neural Smithing - Supervised Learning in Feedforward Artificial Neural Networks*, chap. 10. Classical Optimization Techniques, The MIT press, Cambridge, USA.

Reggiani, P., M. Sivapalan, and S. Hassanizadeh (1998), A unifying framework for watershed thermodynamics: balance equations for mass, momentum, energy and entropy, and the second law of thermodynamics, *Adv. Water. Resour.*, *22*(4), 367–398, doi: 10.1016/S0309-1708(98)00012-8.

Rios Gaona, M., A. Overeem, H. Leijnse, and R. Uijlenhoet (2015), Measurement and interpolation uncertainties in rainfall maps from cellular communication networks, *Hydrol. Earth Syst. Sc.*, *19*, 3571–3584, doi: 10.5194/hess-19-3571-2015.

Roderick, M., P. Greve, and G. Farquhar (2015), On the assessment of aridity with changes in atmospheric CO_2, *Water Resour. Res.*, *51*, 5450–5463, doi: 10.1002/2015WR017031.

Rosero, E., Z. Yang, T. Wagener, L. Gulden, S. Yatheendradas, and G. Niu (2010), Quantifying parameter sensitivity, interaction, and transferability in hydrologically enhanced versions of the Noah land surface model over transition zones during the warm season, *J. Geophys. Res.*, *115*, D03106, doi: 10.1029/2009JD012035.

Roudier, P., J. Andersson, C. Donnelly, L. Feyen, W. Greuell, and F. Ludwig (2016), Projections of future floods and hydrological droughts in Europe under a +2°C global warming, *Climatic Change*, *135*(2), 341–355, doi: 10.1007/s10584-015-1570-4.

Rupp, D. E., and J. S. Selker (2006), On the use of the Boussinesq equation for interpreting recession hydrographs from sloping aquifers, *Water Resour. Res.*, *42*, W12421, doi: 10.1029/2006WR005080.

Samaniego, L., R. Kumar, and S. Attinger (2010), Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale, *Water Resour. Res.*, *46*, W05523, doi: 10.1029/2008WR007327.

Savenije, H. (2009), HESS opinions: The art of hydrology, *Hydrol. Earth Syst. Sc.*, *13*, 157–161, doi: $10.5194/\mathrm{hess}\text{-}13\text{-}157\text{-}2009$.

Sawicz, K., T. Wagener, M. Sivapalan, P. A. Troch, and G. Carrillo (2011), Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA, *Hydrol. Earth Syst. Sc.*, *15*, 2895–2911, doi: $10.5194/\mathrm{hess}\text{-}15\text{-}2895\text{-}2011$.

Schaap, J., P. Lehmann, A. Kaestner, P. Vontobel, R. Hassanein, G. Frei, G. de Rooij, E. Lehmann, and H. Flühler (2008), Measuring the effect of structural connectivity on the water dynamics in heterogeneous porous media using speedy neutron tomography, *Adv. Water Resour.*, *31*(9), 1233–1241, doi: $10.1016/\mathrm{j.advwatres.}$ $2008.04.014$.

Schaefli, B., and H. Gupta (2007), Do Nash values have value?, *Hydrol. Process.*, *21*(15), 2075–2080, doi: $10.$ $1002/\mathrm{hyp.6825}$.

Schmocker-Fackel, P., and F. Naef (2010), More frequent flooding? Changes in flood frequency in Switzerland since 1850, *J. Hydrol.*, *381*, 1–8, doi: $10.1016/\mathrm{j.jhydrol.2009.09.022}$.

Schymanski, S., M. Sivapalan, M. Roderick, L. Hutley, and J. Beringer (2009), An optimality-based model of the dynamic feedbacks between natural vegetation and the water balance, *Water Resour. Res.*, *45*(W01412), doi: $10.1029/2008\mathrm{WR006841}$.

Seibert, J., and K. J. Beven (2009), Gauging the ungauged basin: how many discharge measurements are needed?, *Hydrol. Earth. Syst. Sci.*, *13*, 883–892, doi: $10.5194/\mathrm{hess}\text{-}13\text{-}883\text{-}2009$.

Seibert, J., K. Bishop, A. Rodhe, and J. J. McDonnell (2003), Groundwater dynamics along a hillslope: A test of the steady state hypothesis, *Water Resour. Res.*, *39*(1), 1014, doi: $10.1029/2002\mathrm{WR001404}$.

Seneviratne, S. (2012), Historical drought trends revisited, *Nature*, *491*, 338–339, doi: $10.1038/491338\mathrm{a}$.

Seneviratne, S. I., T. Corti, E. Davin, M. Hirschi, E. Jaeger, I. Lehner, B. Orlowsky, and A. Teuling (2010), Investigating soil moisture-climate interactions in a changing climate: A review, *Earth-Sci. Rev.*, *99*, 125–161, doi: $10.1016/\mathrm{j.earscirev.2010.02.004}$.

Seneviratne, S. I., I. Lehner, J. Gurtz, A. J. Teuling, H. Lang, U. Moser, D. Grebner, L. Menzel, K. Schroff, T. Vitvar, and M. Zappa (2012), Swiss prealpine Rietholzbach research catchment and lysimeter: 32 year time series and 2003 drought event, *Water Resour. Res.*, *48*, W06526, doi: $10.1029/2011\mathrm{WR011749}$.

Shaw, S. B., and S. J. Riha (2012), Examining individual recession events instead of a data cloud: Using a modified interpretation of $\mathrm{d}Q/\mathrm{dt}{-}Q$ streamflow recession in glaciated watersheds to better inform models on low flow, *J. Hydrol.*, *434-435*, 46–54, doi: $10.1016/\mathrm{j.jhydrol.2012.02.034}$.

Sheffield, J., E. Wood, and M. Roderick (2012), Little change in global drought over the past 60 years, *Nature*, *491*, 435–438, doi: $10.1038/\mathrm{nature11575}$.

Sheffield, J., A. Barrett, B. Colle, D. Fernando, R. Fu, K. Geil, Q. Hu, J. Kinter, S. Kumar, B. Langenbrunner, K. Lombardo, L. Long, E. Maloney, A. Mariotti, J. Meyerson, K. Mo, J. Neelin, S. Nigam, Z. Pan, T. Ren, A. Ruiz-Barradas, Y. Serra, A. Seth, J. Thibeault, J. Stroeve, Z. Yang, and L. Yin (2013), North American climate in CMIP5 experiments. Part I: Evaluation of historical simulations of continental and regional climatology, *J. Climate*, *26*, 9209–9244, doi: $10.1175/\mathrm{JCLI}\text{-}\mathrm{D}\text{-}12\text{-}00592.1$.

Shrestha, P., M. Sulis, C. Simmer, and S. Kollet (2015), Impacts of grid resolution on surface energy fluxes simulated with an integrated surface-groundwater flow model, *Hydrol. Earth Syst. Sc.*, *19*, 4317–4326, doi: $10.5194/\mathrm{hess}\text{-}19\text{-}4317\text{-}2015$.

Sivapalan, M. (2003), Prediction in ungauged basins: a grand challenge for theoretical hydrology, *Hydrol. Process.*, *17*, 316303,170, doi: $10.1002/\mathrm{hyp.5155}$.

Sivapalan, M., K. Takeuchi, S. W. Franks, V. K. Gupta, H. Karambiri, V. Lakshmi, X. Liang, J. J. McDonnell, E. M. Mendiondo, P. E. O'Connell, T. Oki, J. W. Pomeroy, D. Schertzer, S. Uhlenbrook, and E. Zehe (2003), IAHS Decade on Predictions in Ungauged Basins (PUB), 2003-2012: Shaping an exciting future for the hydrological sciences, *Hydrolog. Sci. J.*, *48*(6), 857–880, doi: $10.1623/\mathrm{hysj.48.6.857.51421}$.

Sivapalan, M., S. Thompson, C. Hartman, N. Basu, and P. Kumar (2011), Water cycle dynamics in a changing environment: improving predictability through synthesis, *Water Resour. Res.*, *47*(W00J01), doi: $10.1029/$ $2011\mathrm{WR011377}$.

Sood, A., and V. Smakhtin (2015), Global hydrological models: a review, *Hydrolog. Sci. J.*, *60*(4), 549–565, doi: $10.1080/02626667.2014.950580$.

Sorooshian, S. (1983), Evaluation of maximum likelihood parameter estimation techniques for conceptual

rainfall-runoff models: Influence of calibration data variability and length on model credibility, *Water Resour. Res.*, *19*, 251–259, doi: 10.1029/WR019i001p00251.

Sposito, G. (Ed.) (1998), *Scale dependence and scale invariance in hydrology*, Cambridge University Press, Cambridge, UK.

Steele, K., and C. Werndl (2015), Model tuning in engineering: uncovering the logic, *J. Strain Analysis*, *March*, 1–9, doi: 10.1177/0309324715575445.

Stillwell, J. (1989), *Mathematics and its History*, Springer, New York, USA.

Stommel, H. (1963), Varieties of oceanographic experience, *Science*, *139*(3555), 572–576, doi: 10.1126/science.139.3555.572.

Sulis, M., C. Paniconi, and M. Camporese (2011), Impact of grid resolution on the integrated and distributedr esponse of a coupled surface-subsurface hydrological model for the des Anglais catchment, Quebec, *Hydrol. Process.*, *25*, 1853–1865, doi: 10.1002/hyp.7941.

Sutanudjaja, E., L. van Beek, S. de Jong, F. van Geer, and M. Bierkens (2014), Calibrating a large-extent high-resolution coupled groundwater-land surface model using soil moisture and discharge data, *Water Resour. Res.*, *50*, 687–705, doi: 10.1002/2013WR013807.

Tada, T., and K. J. Beven (2012), Hydrological model calibration using a short period of observations, *Hydrol. Process.*, *26*, 883–892, doi: 10.1002/hyp.8302.

Ten Bos, R. (2014), *Water*, Boom Amsterdam, The Netherlands.

Tetzlaff, D., J. Buttle, S. Carey, K. McGuire, H. Laudon, and C. Soulsby (2015), Tracer-based assessment of flow paths, storage and runoff generation in northern catchments: a review, *Hydrol. Process.*, *29*, 3475–3490, doi: 10.1002/hyp.10412.

Teuling, A. J., I. Lehner, J. W. Kirchner, and S. I. Seneviratne (2010), Catchments as simple dynamical systems: Experience from a Swiss prealpine catchment, *Water Resour. Res.*, *46*, W10502, doi: 10.1029/2009WR008777.

Teutschbein, C., T. Grabs, R. H. Karlsen, H. Laudon, and K. Bishop (2015), Hydrological response to changing climate conditions: Spatial streamflow variability in the Boreal region, *Water Resour. Res.*, *51*(12), 9425–9446, doi: 10.1002/2015WR017337.

Todini, E. (1988), Rainfall-runoff modeling - past, present and future, *J. Hydrol.*, *100*, 341–352, doi: 10.1016/0022-1694(88)90191-6.

Todini, E. (2007), Hydrological catchment modelling: Past, present and future, *Hydrol. Earth Syst. Sc.*, *11*(1), 468–482, doi: 10.5194/hess-11-468-2007.

Troch, P. A., A. Berne, P. Bogaart, C. Harman, A. G. J. Hilberts, S. W. Lyon, C. Paniconi, V. R. N. Pauwels, D. E. Rupp, J. S. Selker, A. J. Teuling, R. Uijlenhoet, and N. E. C. Verhoest (2013), The importance of hydraulic groundwater theory in catchment hydrology: The legacy of Wilfried Brutsaert and Jean-Yves Parlange, *Water Resour. Res.*, *49*, 1–18, doi: 10.1002/wrcr.20407.

Troy, T. J., E. F. Wood, and J. Sheffield (2008), An efficient calibration method for continental-scale land surface modeling, *Water Resour. Res.*, *44*, W09411, doi: 10.1029/2007WR006513.

Uchida, T., Y. Asano, T. Mizuyama, and J. J. McDonnell (2004), Role of upslope soil pore pressure on lateral subsurface storm flow dynamics, *Water Resour. Res.*, *40*, W12401, doi: 10.1029/2003WR002139.

Uhlenbrook, S., J. Seibert, C. Leibundgut, and A. Rodhe (1999), Prediction uncertainty of conceptual rainfall-runoff models caused by problems in identifying model parameters and structure, *Hydrolog. Sci. J.*, *44*(5), 779–797.

Van Loon, A., T. Gleeson, J. Clark, A. Van Dijk, K. Stahl, J. Hannaford, G. Di Baldassarre, A. Teuling, L. Tallaksen, R. Uijlenhoet, D. Hannah, J. Sheffield, M. Svoboda, B. Verbeiren, T. Wagener, S. Rangecroft, N. Wanders, and H. Van Lanen (2016), Drought in the Anthropocene, *Nat. Geosci.*, *9*, 89–91, doi: 10.1038/ngeo2646.

Van Loon, A. F., E. Tijdeman, N. Wanders, H. A. J. Van Lanen, A. J. Teuling, and R. Uijlenhoet (2014), How climate seasonality modifies drought duration and deficit, *J. Geophys. Res-Atmos.*, *119*, 4640–4656, doi: 10.1002/2013JD020383.

Van Vliet, M. T. H., J. R. Yearsly, F. Ludwig, S. Vögele, D. P. Lettenmaier, and P. Kabat (2012), Vulnerability of US and European electricity supply to climate change, *Nat. Clim. Change*, *2*, 676–681, doi: 10.1038/nclimate1546.

Vaze, J., D. Post, F. Chiew, J. Perraud, N. Viney, and J. Teng (2010), Climate non-stationarity - validity of calibrated rainfall-runoff models for use in climate change studies, *J. Hydrol.*, *394*(3-4), 447–457, doi: 10.1016/

j.jhydrol.2010.09.018.

Viviroli, D., M. Zappa, J. Gurtz, and R. Weingartner (2009a), An introduction to the hydrological modelling system PREVAH and its pre- and post-processing-tools, *Environ. Modell. Softw.*, *24*, 1209–1222, doi: 10.1016/j.envsoft.2009.04.001.

Viviroli, D., M. Zappa, J. Schwanbeck, J. Gurtz, and R. Weingartner (2009b), Continuous simulation for flood estimation in ungauged mesoscale catchments of Switzerland - Part I: Modelling framework and calibration results, *J. Hydrol.*, *377*, 191–207, doi: 10.1016/j.jhydrol.2009.08.023.

Vivoni, E., V. Ivanov, R. Bras, and D. Entekhabi (2005), On the effects of triangulated terrain resolution on distributed hydrologic model response, *Hydrol. Process.*, *19*, 2101–2122, doi: 10.1002/hyp.5671.

Vořechovský, M. (2015), Hierarchical refinement of latin hypercube samples, *Comput.-aided Civ. Inf.*, *30*(5), 394–411, doi: 10.1111/mice.12088.

Vořechovský, M., and D. Novák (2009), Correlation control in small-sample Monte Carlo type simulations I: A simulated annealing approach, *Probabilistic Eng. Mech.*, *24*, 452–462, doi: 10.1016/j.probengmech.2009.01.004.

Vörösmarty, C., B. Moore III, A. Grace, and M. Gildea (1989), Continental scale models of water balance and fluvial transport: an application to South-America, *Glob. Biogeochem. cy.*, *3*(3), 241–265, doi: 10.1029/GB003i003p00241.

Vrugt, J., and M. Sadegh (2013), Toward diagnostic model calibration and evaluation: Approximate Bayesian computation, *Water Resour. Res.*, *49*, 4335–4345, doi: 10.1002/wrcr.20354.

Vrugt, J. A., C. J. F. ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson (2008), Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov Chain Monte Carlo simulation, *Water Resour. Res.*, *44*, W00B09, doi: 10.1029/2007WR006720.

Wada, Y., and M. Bierkens (2014), Sustainability of global water use: past reconstruction and future projections, *Environ. Res. Lett.*, *9*, 104003, doi: 10.1088/1748-9326/9/10/104003.

Wada, Y., D. Wisser, and M. F. P. Bierkens (2014), Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources, *Earth Syst. Dynam.*, *5*, 15–40, doi: 10.5194/esd-5-15-2014.

Wagener, T., and H. Gupta (2005), Model identification for hydrological forecasting under uncertainty, *Stoch. Environ. Res. Risk Assess.*, *19*, 378–387, doi: 10.1007/s00477-005-0006-5.

Wagener, T., and H. S. Wheater (2006), Parameter estimation and regionalization for continuous rainfall-runoff models including uncertainty, *J. Hydrol.*, *320*, 132–154, doi: 10.1016/j.jhydrol.2005.07.015.

Wagener, T., H. Wheater, and H. Gupta (Eds.) (2004), *Rainfall-Runoff Modelling in Gauged and Ungauged Catchments*, Imperial College Press.

Wagener, T., M. Sivapalan, P. Troch, B. McGlynn, C. Harman, H. Gupta, P. Kumar, P. Rao, N. Basu, and J. Wilson (2010), The future of hydrology: An evolving science for a changing world, *Water Resour. Res.*, *46*, W05301, doi: 10.1029/2009WR008906.

Wang, Y., B. He, and K. Takase (2009), Effects of temporal resolution on hydrological model parameters and its impact on prediction of river discharge, *Hydrolog. Sci. J.*, *54*(5), 886–898, doi: 10.1623/hysj.54.5.886.

Weiler, M., and K. Beven (2015), Do we need a Community Hydrological Model?, *Water Resour. Res.*, *51*, 7777–7784, doi: 10.1002/2014WR016731.

Wen, Z., X. Liang, and S. Yang (2012), A new multiscale routing framework and its evaluation for land surface modeling applications, *Water Resour. Res.*, *48*, W08528, doi: 10.1029/2011WR011337.

Wenger, S., C. Luce, A. Hamlet, D. Isaak, and H. Neville (2010), Macroscale hydrologic modeling of ecologically relevant flow metrics, *Water Resour. Res.*, *46*, W09513, doi: 10.1029/2009WR008839.

Western, A., S. Zhou, R. Grayson, T. McMahona, G. Blöschl, and D. Wilson (2004), Spatial correlation of soil moisture in small catchments and its relationship to dominant spatial hydrological processes, *J. Hydrol.*, *286*, 113–134, doi: 10.1016/j.jhydrol.2003.09.014.

Whitaker, S. (1999), *Theory and Applications of Transport in Porous Media: The Method of Volume Averaging*, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Williams, C., M. Reichstein, N. Buchmann, D. Baldocchi, C. Beer, C. Schwalm, G. Wohlfahrt, N. Hasler, C. Bernhofer, T. Foken, D. Papale, S. Schymanski, and K. Schaefer (2012), Climate and vegetation controls on the surface water balance: Synthesis of evapotranspiration measured across a global network of flux towers,

Water Resour. Res., *48*(W06523), doi: 10.1029/2011WR011586.

Wilson, C., J. Valdes, and I. Rodriguez-Iturbe (1979), On the influence of the spatial distribution of rainfall on storm runoff, *Water Resour. Res.*, *15*(2), 321–328, doi: 10.1029/WR015i002p00321.

Wood, E., M. Sivapalan, K. Beven, and L. Band (1988), Effects of spatial variability and scale with implications to hydrologic modeling, *J. Hydrol.*, *102*, 29–47, doi: 10.1016/0022-1694(88)90090-X.

Wood, E., D. Lettenmaier, and V. Zartarian (1992), A land-surface hydrology parameterization with subgrid variability for general circulation models, *J. Geophys. Res.*, *97*(D3), 2717–2728, doi: 10.1029/91JD01786.

Wood, E., J. Roundy, T. Troy, L. van Beek, M. P. Bierkens, E. Blyth, A. de Roo, P. Döll, M. Ek, J. Famiglietti, D. Gochis, N. van de Giesen, P. Houser, P. Jaffé, S. Kollet, B. Lehner, D. Lettenmaier, C. Peters-Lidard, M. Sivapalan, J. Sheffield, A. Wade, and P. Whitehead (2011), Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water, *Water Resour. Res.*, *47*, W05301, doi: 10.1029/2010WR010090.

Yang, J., P. Reichert, and K. Abbaspour (2007), Bayesian uncertainty analysis in distributed hydrologic modeling: A case study in the Thur river basin (Switzerland), *Water Resour. Res.*, *43*, W10401, doi: 10.1029/2006WR005497.

Yapo, P. O., H. V. Gupta, and S. Sorooshian (1996), Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data, *J. Hydrol.*, *181*, 23–48, doi: 10.1016/0022-1694(95)02918-4.

Yevjevich, V. (1968), Misconceptions in hydrology and their consequences, *Water Resour. Res.*, *4*(2), 225–232, doi: 10.1029/WR004i002p00225.

Zappa, M. (2002), *Multiple-Response Verification of a Distributed Hydrological Model at Different Spatial Scales*, chap. 4. The sensitivity of distributed hydrological simulations to the spatial resolution of physiographic data, pp. 35–51, 14895, ETH Zürich, doi: 10.3929/ethz-a-004529728.

Zappa, M., and C. Kan (2007), Extreme heat and runoff extremes in the Swiss Alps, *Nat. Hazards Earth Syst. Sc.*, *7*, 375–389, doi: 10.5194/nhess-7-375-2007.

Zhao, F., L. Zhang, F. Chiew, J. Vaze, and L. Cheng (2013), The effect of spatial rainfall variability on water balance modelling for south-eastern Australian catchments, *J. Hydrol.*, *493*, 16–29, doi: 10.1016/j.jhydrol.2013.04.028.

Zhao, R., Y. Zuang, L. Fang, X. Liu, and Q. Zhang (1980), The Xinanjiang model, *Hydrological forecasting - Prévisions hydrologiques (Proceedings of the Oxford Symposium, April 1980)*, *129*, 351–356.

Statement of authorship contribution

The introduction (Chapter 1) and the synthesis (Chapter 8) have completely been written by Lieke Melsen only. An overview of the contribution from co-authors for the other chapters is provided below.

AN	=	Andrew Newman (NCAR)	NM	=	Naoki Mizukami (NCAR)
AT	=	Adriaan Teuling (WUR)	PM	=	Pablo Mendoza (NCAR)
LM	=	Lieke Melsen (WUR)	PT	=	Paul Torfs (WUR)
MC	=	Martyn Clark (NCAR)	RU	=	Remko Uijlenhoet (WUR)
MZ	=	Massimiliano Zappa (WSL)	SB	=	Sonja van Berkum (WUR)
NA	=	Nans Addor (NCAR)			

Chapter 3

Conception or design of the work:	AT, LM, PT, SB
Data collection:	LM, SB
Data analysis and interpretation:	LM in consultation with AT, PT, RU
Drafting the article:	LM
Critical revision of the article:	AT, RU
Approval of the version to be published:	LM, AT, SB, PT, RU

Chapter 4

Conception or design of the work:	LM, AT, PT, NM
Data collection:	LM, MZ
Data analysis and interpretation:	LM in consultation with AT, PT, RU, MC
Drafting the article:	LM
Critical revision of the article:	AT, RU, MZ, MC
Approval of the version to be published:	LM, AT, PT, MZ, NM, MC, RU

Chapter 5

Conception or design of the work:	LM, AT, PT, PM, NM
Data collection:	LM, MZ
Data analysis and interpretation:	LM in consultation with PT, AT, RU
Drafting the article:	LM
Critical revision of the article:	AT, RU, PM, MC
Approval of the version to be published:	LM, AT, PT, MZ, NM, PM,MC, RU

Chapter 6

Conception or design of the work:	LM, AT, MC, NA
Data collection:	LM, AN, NM, NA
Data analysis and interpretation:	LM in consultation with AT, PT, RU, MC, NA, NM, AN
Drafting the article:	LM
Critical revision of the article:	AT, NA, RU
Approval of the version to be published:	LM, AT, PT, RU, MC, NA, NM, AN

Chapter 7

Conception or design of the work:	LM
Data collection:	LM
Data analysis and interpretation:	LM, AT, PT, NM, MC, RU
Drafting the article:	LM
Critical revision of the article:	AT, RU, MC
Approval of the version to be published:	LM, AT, PT, RU, NM, MC

Statement of code and data availability

Input and evaluation data

- Meteorological and catchment data for the Rietholzbach and the Thur can be requested from ETH Zürich and WSL in Birmensdorf, respectively.

- Discharge data can be requested through the website of the Swiss Federal Office of the Environment (FOEN), see: http://www.bafu.admin.ch/wasser/.

- For the 605 basins, all USGS discharge and Daymet forcing data can be downloaded from:
  http://ral.ucar.edu/projects/hap/flowpredict/subpages/modelvar.php

- Catchment characteristics for the 605 basins are currently only available upon request at NCAR, but will be publicly available soon.

Models

- The VIC model can be downloaded from:
  http://vic.readthedocs.io/en/master/

- The SAC model can be downloaded from:
  http://www.nws.noaa.gov/iao/iao_hydroSoftDoc.php

- The employed HBV model was an R-package available at:
  https://cran.r-project.org/web/packages/TUWmodel/index.html

- Code for Kirchners 'Catchments as simply dynamical systems' implementation for the Rietholzbach is available upon request from Lieke Melsen and Adriaan Teuling.

Output data

The output data from all studies are available upon request from Lieke Melsen. This for example includes 26.500 model runs with a daily time step for the period 1980-2008 and 2065-2100 per basin for the 605 basins in CONUS (3 hydrological models, run with a large parameter sample, forced with 5 different statistically downscaled and bias corrected GCM outputs).

Code

All code (mainly Matlab) used to perform analyses and construct figures is available upon request from Lieke Melsen, including the distributed VIC implementation for the Thur basin.

Summary

Six steps can be distinguished in the process of hydrological modelling: the perceptual model (deciding on the processes), the conceptual model (deciding on the equations), the procedural model (get the code to run on a computer), calibration (identify the parameters), evaluation (confronting output with observations), and uncertainty analysis (estimate uncertainty in the model and its output). An engineer conducts these steps different than a scientist, because the goal of an engineer is to solve practical problems, while the goal of a scientist is to increase the understanding of the system. The difference between scientists and engineers is most pronounced in the perceptual modelling step. However, in many of the current hydrologic sciences studies, engineering and scientific approaches are mixed. As a scientist, three common philosophies of science can be adopted: verificationism, falsificationism, and Bayesianism. It was demonstrated that verificationism most closely resembles engineering in the modelling steps, while falsificationism and Bayesianism call for a different practice.

In this thesis, several of the modelling steps have been investigated in more detail. In order to investigate these modelling steps, we applied widely used hydrological models (**Chapter 2**). These models vary in complexity, and have been applied to catchments with varying temporal and spatial scales.

In **Chapter 3** three parameter identification methods and their data requirements were compared for a small (3.3 km^2) catchment using a parsimonious two-parameter model. Two methods based on discharge data were employed, Bayesian based automatic calibration (DREAM) and recession analysis, and one physics-based method was employed, Boussinesq theory. Automatic calibration and recession analysis both required five months of discharge data in order to obtain stable parameter estimations. Boussinesq theory, which allows a-priori parameter estimation based on catchment characteristics, showed to lead to highly uncertain parameters due to uncertainty in the catchment characteristics.

Chapter 4 deals with the transferability (and thus sensitivity) of parameters across spatial and temporal resolutions in the Thur catchment (1703 km^2). It was shown that parameters were hardly sensitive to the spatial resolution (a high transferability), while the parameters were very sensitive to the temporal resolution (especially from hourly/daily to a monthly time step). This indicates that the spatial variability is substantially underestimated. In this study we adopted common practice for hyper-resolution models applied at a large domain. The results therefore provide a strong motivation to further investigate and improve the representation of spatial and temporal variability in large-domain hydrological models.

Chapter 5 shows that decisions during model configuration, basically subjective decisions of the modeller, significantly impact the simulation of hydrological extremes in the Thur basin (1703 km^2). We explored four modelling decisions; the spatial resolution

of the model, the spatial representation of forcing, the calibration period, and the performance metric, and investigated if these decisions influenced the simulated flood and drought characteristics in the Thur basin. It was shown that for the flood characteristics, the performance metric was the most influential decision, and for the drought characteristics, the calibration period was most important. Subjective modelling decisions introduce uncertainty in the modelling process. Working with multiple hypotheses of model implementations could help in providing insight in this uncertainty.

In **Chapter 6** we explore three sources of uncertainty in a hydrological climate change impact assessment for the period 2070–2100 for 605 basins throughout the contiguous United States; parameter uncertainty, hydrological model structural uncertainty, and uncertainty in General Circulation Model (GCM) forcing. It was demonstrated that the uncertainty introduced by any of the three sources can be thus large that even the sign of the change is unknown in many basins (i.e., an increase or decrease in average annual runoff compared to the period 1985–2008). This uncertainty could be attributed to the snow parameterization in the hydrological models and disagreement among the GCMs on the change in precipitation. Furthermore, it was demonstrated that processes related to aridity and intermittent flow behaviour are not yet well captured in the investigated hydrological models.

In **Chapter 7** it was shown, based on a literature study of 192 peer-reviewed publications, that the spatial resolution at which the Variable Infiltration Capacity (VIC) model is applied has increased, while the calibration and validation time interval has remained unchanged. It is argued that the calibration and validation time interval should keep pace with the spatial resolution of the model in order to resolve the processes relevant at the applied spatial resolution. Different processes are relevant at different spatial scales; the calibration and validation time interval should reflect the temporal scale of these processes in order to estimate the credibility of the model.

The results from all the studies can be summarized in three points: Not only the model choice, but also the configuration of the model determines the outcome of the model; sufficient data are needed to constrain and evaluate a model; and the large uncertainty in modelling studies provides a strong motivation to increase our understanding - i.e., to focus on science rather than on engineering. In order to establish this, models should be related to theories (hypotheses), which should be tried to falsified. The model set-up should be considered an element of the tested theory. Novel observation technologies provide the opportunity to test and falsify these theories, and can lead to the formulation of new theories.

Samenvatting

Er zijn zes stappen te onderscheiden in het proces van hydrologisch modelleren: het opstellen van het perceptuele model (de relevante processen), vervolgens het conceptuele model (de wiskundige vergelijkingen die het proces beschrijven), het procedurele model (de code laten werken op een computer), kalibratie (het schatten van de parameterwaarden), evaluatie (modelresultaten vergelijken met observaties), en onzekerheidsanalyse (inschatting maken van onzekerheid in het model en de modelresultaten). Een ingenieur zal deze stappen op een andere manier uitvoeren dan een wetenschapper, omdat het doel van een ingenieur is om praktische problemen op te lossen, terwijl het doel van een wetenschapper is om het systeem beter te begrijpen. Het verschil tussen beide benaderingen komt het sterkst naar voren in het perceptuele model. In veel hydrologische studies worden ingenieurs- en wetenschappelijke benaderingen echter door elkaar gebruikt. De drie meest gangbare wetenschapsfilosofieën zijn verificatie, falsificatie, en Bayes' theorie. Verificatie ligt het dichtst bij de ingenieursaanpak, terwijl falsificatie en Bayes' theorie duidelijk om een andere aanpak vragen.

In dit proefschrift zijn verschillende modelleerstappen in meer detail onderzocht. Om deze stappen te onderzoeken, hebben we gebruik gemaakt van verschillende veel gebruikte hydrologische modellen (**Hoofdstuk 2**). Deze modellen variëren sterk in complexiteit en zijn toegepast op verschillende tijd- en ruimteschalen.

In **Hoofdstuk 3** vergelijken we drie parameteridentificatiemethoden en de bijbehorende data-vereisten in een klein stroomgebied (3.3 km^2), waarbij we gebruik maken van een simpel 2-parameter-model. In twee methoden worden afvoerdata gebruikt voor parameteridentificatie: een Bayesiaans automatisch kalibratie-algoritme (DREAM) en recessie-analyse. De derde methode is gebaseerd op fysica, de theorie van Boussinesq, waardoor de parameters van het model bepaald kunnen worden op basis van stroomgebiedskarakteristieken. Zowel automatische kalibratie als recessie-analyse hadden vijf maanden aan uurlijkse data nodig om stabiele parameters te kunnen identificeren. De parameters verkregen met Boussinesqs theorie bleken uiterst onzeker door onzekerheid in stroomgebiedskarakteristieken.

Hoofdstuk 4 behandelt de overdraagbaarheid (en daarmee gevoeligheid) van parameterwaarden tussen verschillende ruimtelijke en temporele resoluties van het Variable Infiltration Capacity (VIC) model toegepast op het stroomgebied van de Thur in Zwitserland (1703 km^2). We hebben aangetoond dat de parameterwaarden nauwelijks gevoelig zijn voor de ruimtelijke resolutie van het model, terwijl ze zeer gevoelig zijn voor de temporele resolutie (vooral van uur/dag naar maand). Dit laat zien dat de ruimtelijke variatie niet goed in het model meegenomen wordt. In deze studie is gebruik gemaakt van standaardmethoden voor hoge-resolutiemodellen die toegepast worden op grote gebieden. De resultaten zijn daarom aanleiding om de temporele en ruimtelijke variabiliteit in modellen ontwikkeld voor en toegepast op grote domeinen verder te onderzoeken en verbeteren.

In **Hoofdstuk 5** is aangetoond dat subjectieve beslissingen van de modelleur tijdens het opzetten van een hydrologisch model significant invloed hebben op gesimuleerde hydrologische extremen in het stroomgebied van de Thur. We hebben vier verschillende modelbeslissingen onderzocht: de ruimtelijke resolutie van het model, de ruimtelijke resolutie van de modelforcering, de kalibratieperiode en het evaluatiecriterium. De invloed van deze modelbeslissingen op verschillende karakteristieken van piekafvoeren en droogte is onderzocht. De simulatie van piekafvoeren bleek het meest gevoelig voor het evaluatiecriterium, bij droogtesimulatie bleek de kalibratieperiode het meest invloedrijk. Subjectieve modelbeslissingen zorgen dus voor onzekerheid in het modelleerproces. Werken met verschillende modelimplementaties verschaft inzicht in deze onzekerheid.

In **Hoofdstuk 6** onderzochten we drie bronnen van onzekerheid in een hydrologische effectbeoordeling van klimaatverandering voor de periode 2070-2100 voor 605 stroomgebieden verspreid over de Verenigde Staten. Deze drie bronnen zijn: onzekerheid in parameters, onzekerheid in de hydrologische modelstructuur en onzekerheid in de neerslag die gesimuleerd wordt door algemene circulatiemodellen (GCMs). Aangetoond is dat elk van de drie bronnen kan leiden tot een dergelijke mate van onzekerheid dat zelfs de richting van de verandering in afvoer (toename of afname in gemiddelde jaarlijkse afvoer t.o.v. 1985-2008) onbekend is. Deze onzekerheid kon specifiek toegeschreven worden aan de parametrisatie van sneeuw in hydrologische modellen en inconsistentie tussen de verschillende GCMs over de verandering in neerslag. Daarnaast is aangetoond dat processen gerelateerd aan droogte niet goed beschreven zijn in de onderzochte hydrologische modellen.

Hoofdstuk 7 is gebaseerd op een literatuurstudie van 192 peer-reviewed publicaties waarin het VIC model gebruikt wordt. Uit de literatuurstudie bleek dat de ruimtelijke resolutie waarop het model wordt toegepast, is toegenomen in de tijd, terwijl het tijdsinterval van de kalibratie en validatie van het model niet veranderd is. Gesteld wordt dat het kalibratie- en validatietijdsinterval gelijke tred moeten houden met de ruimtelijke resolutie van het model, omdat bij een andere ruimtelijke resolutie ook andere processen relevant worden. Het kalibratie- en validatietijdsinterval moet de tijdschaal van deze processen reflecteren.

De resultaten die gepresenteerd zijn in de verschillende hoofdstukken kunnen samengevat worden in drie punten: niet alleen de modelkeuze maar ook de modelconfiguratie bepaalt de resultaten van het model; voldoende data zijn nodig om modellen te kalibreren en evalueren; en de grote onzekerheid in modelstudies is een sterke motivatie om ons begrip van het hydrologisch systeem te vergroten (meer focus op de wetenschappelijke benadering). Om dit te bereiken, moeten modellen weer gekoppeld worden aan hun onderliggende theorie. De modelconfiguratie moet beschouwd worden als onderdeel van de theorie die getest wordt met het model. Nieuwe observatietechnieken bieden de mogelijkheid om de theorieën te falsificeren en om nieuwe theorieën te formuleren.

Dankwoord

Het dankwoord is bij uitstek het bewijs dat het aantal lezers van bepaalde pagina's in dit proefschrift waarschijnlijk omgekeerd evenredig is aan de tijd die ik besteed heb aan het schrijven ervan. Het maakt het dankwoord in ieder geval niet minder belangrijk; zonder de hulp van velen was dit proefschrift er nooit geweest.

Allereerst wil ik natuurlijk mijn promotor, Remko Uijlenhoet, bedanken: voor de vele discussies, voor het eindeloos nakijken van mijn manuscripten, voor de ondersteuning bij het (soms moeizame) publicatie-proces, en voor de gezellige borrels en congresbezoeken. Daarnaast ook veel dank voor mijn dagelijks begeleider en co-promotor, Ryan Teuling: bedankt dat ik altijd onaangekondigd binnen kon lopen (waar ik veel gebruik van heb gemaakt), de discussies onder het genot van een goede koffie of chai latte, en je eindeloze commentaar op mijn figuren. Ik begin haast te geloven dat ik figuren maken een klein beetje onder de knie begin te krijgen. Furthermore, I would like to thank my other co-promotor, Martyn Clark, especially for your talent to provide suggestions to look at my work from a slightly different angle, which always improved the interpretation of the results. En, hoewel niet officieel co-promotor voelt het toch zo: Paul Torfs, voor al je advies wat betreft statistiek, interpretatie, en programmeerwerk, maar ook voor onze bezoekjes aan lezingen van onder anderen Peter Sloterdijk en Martha Nussbaum, om de hersenen ook eens te trainen op niet-hydrologische vraagstukken. Henny van Lanen, bedankt dat je me als student-assistent geïntroduceerd hebt in de wetenschappelijke wereld - daar lag zeker de bron om verder te gaan in het onderzoek. Professor Savenije (TU Delft), bedankt dat u als opponent aanwezig wilt zijn bij mijn verdediging!

Uiteraard wil ik ook al mijn andere collega's en oud-collega's bij de hydrologie-vakgroep bedanken, voor de gezellige pauzes (en het proefschrift-bashen, ben ik nu aan de beurt?), en de mooie congresbezoekjes. Special thanks to my two lovely roomies and collegazo's Manuel and Tjitske, for providing such a good vibe in our room. Manuel, for always sharing your chocolate with your ladies. En Tjits, bedankt dat jij me als paranimf bij wilt staan! Roel, Henny, Ryan, Anne, Tjitske, Lotte, Victor, Claudia, en Remko; voor de mooie tripjes naar IJsland en de gesmeerde samenwerking aldaar, en de huiselijke gezelligheid bij het beroemde Hupsel practicum. Olda and Joris, my highly appreciated former roomies, thanks for your useful guidance at the start of my project.

Furthermore, I would like to thank all my co-authors from abroad. Massimiliano Zappa, thanks for your contribution to our Thur studies, and for having me over to WSL, where I enjoyed the Swiss snow and of course the Swiss cheese fondue. Thanks to Naoki, Andy, Nans, and Pablo, for the run-beer-pizza-raffle nights, but especially for teaching me so much! Thanks also to all my other NCAR colleagues who did make sure that I really enjoyed my stay in Boulder.

Ik heb het geluk dat veel van mijn vrienden in vergelijkbare disciplines werken, waardoor inhoudelijke discussies goed te combineren waren met de nodige ontspanning, zoals

tripjes naar Norwich, Skye, Princeton (Niko en Karin) en Bristol (Yvonne en Joost). Bedankt voor jullie feedback en commentaar, maar bovenal voor de gezelligheid. Niko, bedankt voor de gastvrijheid op je computer. Yvonne, bedankt dat je me bij wilt staan als paranimf!

Ontspanning zonder inhoudelijke discussies, of eigenlijk, met inhoudelijke discussies maar over andere onderwerpen, was ook zeer noodzakelijk voor het tot-een-goed-einde brengen van dit proefschrift. Astrid, voor de mooie avonturen, o.a. in Wenen. Al mijn (Eks)AkCie vriendjes en vriendinnetjes; Boyen, Daniel, Tobias, Nauta, Iris, Lisanne, (etc.). Arjan en de familie Van Eijk, voor alle warmte en support. De familie Schaap, voor de gezelligheid en de goede zorgen. En last but not least: de Plakkies. Juud, Mayo en Ellis, bedankt voor jullie onvoorwaardelijke vriendschap. Het geeft mij altijd veel vertrouwen te weten dat ik vrienden heb waar ik altijd op terug kan en mag vallen, dat we altijd aan een half woord genoeg hebben, ongeacht hoelang we elkaar niet gezien hebben en hoe verschillend onze paden inmiddels zijn. Thnx matties!

En natuurlijk wil ik mijn lieve familie bedanken. Jasper, Diede, Simone, voor jullie goede humor en gezellige drukte. Pap en mam, voor jullie eindeloze steun (jullie verwennen je kinderen te veel!). En Joris, bedankt voor alle cappuccino's (sterker nog; voor mij koffie leren drinken), afbakbroodjes, pasta pesto's, mooie reizen, en het belangrijkste: dat je er voor me bent, door dik en dun. Ik kijk uit naar de toekomst!

Lieke

Curriculum Vitae
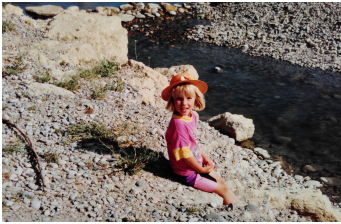
Lieke Anna Melsen
7 November 1988, Amsterdam

2001-2007	Pre-university education (VWO), VeenLanden College, Mijdrecht Profile *Nature and Technology* | *Nature and Health*
2007-2010	BSc Soil, Water, Atmosphere, Wageningen University Minor *Applied Philosophy*
2010-2012	MSc Hydrology and Water Quality, Wageningen University Specialization *Hydrology and Quantitative Water Management* Internship at Nepal Engineering College, Kathmandu, Nepal
2013-2017	PhD candidate, Wageningen University Chair Group *Hydrology and Quantitative Water Management* Including two-month visit to the National Center for Atmospheric Research (NCAR), Boulder (CO), USA

List of publications

Peer-reviewed publications

Huza, J., A. Teuling, I. Braud, J. Grazioli, **L. Melsen**, G. Nord, T. Raupach, and R. Uijlenhoet (2014), Precipitation, soil moisture and runoff variability in a small river catchment (Ardèche, France) during HyMeX special observation period 1, *J. Hydrol.*, *516*, 330–342, doi: 10.1016/j.jhydrol.2014.01.041.

Melsen, L., A. Teuling, S. van Berkum, P. Torfs, and R. Uijlenhoet (2014), Catchments as simple dynamical systems: A case study on methods and data requirements for parameter identification, *Water Resour. Res.*, *50*, 5577–5596, doi: 10.1002/2013WR014720.

Melsen, L., A. Teuling, P. Torfs, R. Uijlenhoet, N. Mizukami, and M. Clark (2016), HESS opinions: The need for process-based evaluation of large-domain hyper-resolution models, *Hydrol. Earth Syst. Sc.*, *20*, 1069–1079, doi: 10.5194/hess-20-1069-2016.

Melsen, L., A. Teuling, P. Torfs, M. Zappa, N. Mizukami, M. Clark, and R. Uijlenhoet (2016), Representation of spatial and temporal variability in large-domain hydrological models: Case study for a mesoscale prealpine basin, *Hydrol. Earth Syst. Sc.*, *20*, 2207–2226, doi: 10.5194/hess-20-2207-2016.

Melsen, L., P. Torfs, R. Uijlenhoet, and A. Teuling (2017), Comment on "Most computational hydrology is not reproducible, so is it really science?", *Water Resour. Res.*, *accepted*.

Melsen, L., A. Teuling, P. Torfs, M. Zappa, N. Mizukami, P. Mendoza, M. Clark, and R. Uijlenhoet (2017), Subjective modeling decisions significantly impact the simulation of hydrological extremes, *J. Hydrol*, *in review*.

Melsen, L., N. Addor, N. Mizukami, A. Newman, P. Torfs, M. Clark, R. Uijlenhoet, and A. Teuling (2017), Deciphering (un)certainties in hydrologic projections, *Sci. Rep.*, *submitted*.

Sada, R., A. Shrestha, A. Shukla, and **L. Melsen** (2014), People's experience and facts of changing climate: impacts and responses, *Int. J. Clim. Change Str.*, *6*(1), 47–62.

Stahl, K., I. Kohn, V. Blauhut, J. Urquijo, L. De Stefano, V. Acacio, S. Dias, J. H. Stagge, L. M. Tallaksen, E. Kampragou, A. F. Van Loon, L. J. Barker, **L. Melsen**, C. Bifulco, D. Musolino, A. de Carli, A. Massarutto, D. Assimacopoulos, and H. A. J. Van Lanen (2016), Impacts of European drought events: insights from an international database of text-based reports, *Nat. Hazards Earth Syst. Sc.*, *16*, 801–819, doi: 10.5194/nhess-16-801-2016.

Teuling, A., C. Taylor, J. Meirink, **L. Melsen**, D. Miralles, C. van Heerwaarden, R. Vautard, A. Stegehuis, G. Nabuurs, and J. Vilà-Guerau de Arellano (2017), Observational evidence for cloud cover enhancement over western European forests, *Nat. Comm.*, *8*(14065), 1–7, doi: 10.1038/ncomms14065.

Book chapters

Sada, R., A. Shukla, A. Shrestha, and **L. Melsen** (2016), Water security in Peri-urban South Asia (Eds: V. Narain and A. Prakash), Ch. 4: Growing Urbanization, Changing Climate and Adaptation in Peri-urban Kathmandu: Emerging shapes of Water and Livelihood Security, *Oxford University Press*, Delhi, India.

Melsen, L. (2017), Rood-groene politiek voor de 21e eeuw - duurzame waarden voor de sociaal-democratie (Eds: H. de Coninck, M. Hurenkamp, **L. Melsen**, and J. Opschoor), Ch. 6: De Waarde van Water, *Wiardi Beckman Stichting publicatie*, Van Gennep Amsterdam, The Netherlands.

Netherlands Research School for the
Socio-Economic and Natural Sciences of the Environment

D I P L O M A

For specialised PhD training

The Netherlands Research School for the
Socio-Economic and Natural Sciences of the Environment
(SENSE) declares that

Lieke Anna Melsen

born on 7 November 1988 in Amsterdam, The Netherlands

has successfully fulfilled all requirements of the
Educational Programme of SENSE.

Wageningen, 19 April 2017

the Chairman of the SENSE board

the SENSE Director of Education

Prof. dr. Huub Rijnaarts

Dr. Ad van Dommelen

The SENSE Research School has been accredited by the Royal Netherlands Academy of Arts and Sciences (KNAW)

K O N I N K L I J K E   N E D E R L A N D S E
A K A D E M I E   V A N   W E T E N S C H A P P E N

The SENSE Research School declares that **Ms Lieke Melsen** has successfully fulfilled all requirements of the Educational PhD Programme of SENSE with a work load of 45.4 EC, including the following activities:

SENSE PhD Courses

o   Environmental research in context (2013)
o   Research in context activity: 'Contribution to the review of an IPCC chapter under coordination of the Netherlands Environmental Assessment Agency (PBL)' (2013)
o   Bayesian statistics (2014)
o   Pitch and presentation skills (2014)

Other PhD and Advanced MSc Courses

o   Model building, inference and hypothesis testing,
    Gabriel Lippmann Institute, Luxembourg (2013)
o   Fundamental and advanced topics in hydrogeology and hydrological simulation,
    University of Cagliary, Italy (2014)
o   Improving the theoretical underpinnings of hydrologic models,
    University of Bologna, Italy (2016)

External training at a foreign research institute

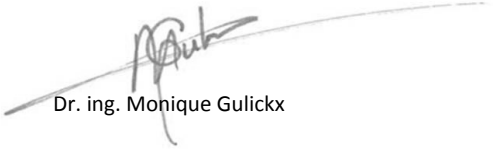o   Two months research visit, National Center for Atmospheric Research (NCAR), Boulder (CO), United States (2015)

Management and Didactic Skills Training

o   Supervising six BSc students with their BSc thesis (2013-2015)
o   Assisting field practicals of the BSc course 'Field practical hydrology, water quality, and meteorology' (2013-2016)
o   Lecturing, assisting computer practicals, and assisting the excursion of the MSc course 'Catchment hydrology' (2013-2016)
o   Supervising four MSc student with their MSc thesis (2014-2016)
o   Member of the WIMEK PhD-council (2015-2016)
o   Reviewer for the journals 'Hydrological Processes', 'Water Resources Research', and 'Hydrology and Earth System Sciences' (2014-2016)
o   Involved in the Young Hydrologic Society (YHS) by (co-)convening several sessions organized by YHS at conferences (2015-2016)

Selection of Oral Presentations

o   *Linking time and space scales in distributed hydrological modelling: A case study for the VIC model.* European Geosciences Union (EGU) General Assembly, 12-17 April 2015, Vienna, Austria
o   *Model output: fact or artefact?* European Geosciences Union (EGU) General Assembly, 12-17 April 2015, Vienna, Austria

SENSE Coordinator PhD Education

Dr. ing. Monique Gulickx