Journal of Hydrology 568 (2019) 1093–1104

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Subjective modeling decisions can significantly impact the simulation of flood and drought events

Lieke A. Melsen[a,*], Adriaan J. Teuling[a], Paul J.J.F. Torfs[a], Massimiliano Zappa[b], Naoki Mizukami[c], Pablo A. Mendoza[d,e], Martyn P. Clark[c], Remko Uijlenhoet[a]

[a] Hydrology and Quantitative Water Management Group, Wageningen University, Wageningen, The Netherlands
[b] Swiss Federal Research Institute, WSL, Birmensdorf, Switzerland
[c] National Center for Atmospheric Research (NCAR), Boulder, CO, USA
[d] Advanced Mining Technology Center, Universidad de Chile, Santiago, Chile
[e] Department of Civil Engineering, Universidad de Chile, Santiago, Chile

ARTICLE INFO

This manuscript was handled by A. Bardossy, Editor-in-Chief, with the assistance of Peter Molnar, Associate Editor

Keywords:
Hydrological modeling
Hydrological extremes
Subjectivity
Modeling decisions
Model configuration

ABSTRACT

It is generally acknowledged in the environmental sciences that the choice of a computational model impacts the research results. In this study of a flood and drought event in the Swiss Thur basin, we show that modeling decisions during the model configuration, beyond the model choice, also impact the model results. In our carefully designed experiment we investigated four modeling 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 flood characteristics were mainly affected by the performance metric, whereas the drought characteristics were mainly affected by the calibration period. The results could be related to the processes that triggered the particular events studied. The impact of the modeling decisions on the simulations did, however, vary among the investigated sub-basins. In spite of the limitations of this study, our findings have important implications for the understanding and quantification of uncertainty in any hydrological or even environmental model. Modeling decisions during model configuration introduce subjectivity from the modeler. Multiple working hypotheses during model configuration can provide insights on the impact of such subjective modeling decisions.

1. Introduction

In jury sports, such as gymnastics, the jury is supposed to objectively evaluate the outcome of the competition. 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). This suggests 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 intercomparison studies (see e.g. Joussaume et al. (1999) on climate modeling, Holländer et al. (2009) and Clark et al. (2015) on hydrological modeling, Freni et al. (2009) on urban stormwater modeling and Bennett et al. (2013) on benchmarking environmental models). The modeler or expert acts as jury to determine the most appropriate model for the question at hand (Crout et al., 2009), while model intercomparison studies provide the modeler 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 intercomparison studies.

It should be noted, however, that expert knowledge is actually a mixture of opinion and knowledge (Krueger et al., 2012), also interestingly shown by the model-intercomparison of Holländer et al. (2009), where different modelers decided differently on which processes were relevant enough to represent in the model. The opinion-part of expert knowledge introduces subjectivity in the model choice, in the

* Corresponding author.
E-mail address: lieke.melsen@wur.nl (L.A. Melsen).

https://doi.org/10.1016/j.jhydrol.2018.11.046
Received 11 September 2017; Received in revised form 15 November 2018; Accepted 16 November 2018
Available online 23 November 2018

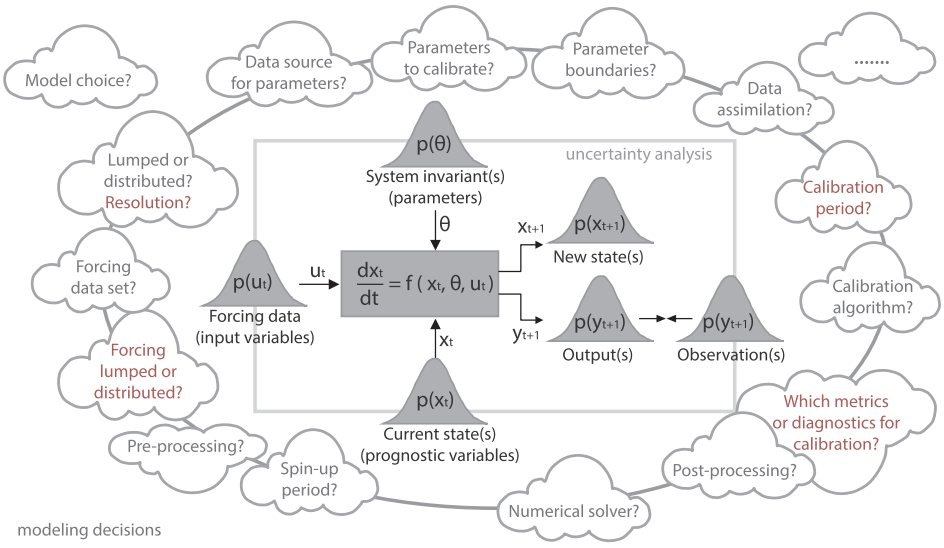

Fig. 1. Bayesian model evaluation framework for a hydrological model, based on Fig. 1 of Vrugt and Sadegh (2013), which explicitly recognizes uncertainty in parameters, forcing data, initial state, model structure, output and model state. The Bayesian framework is surrounded by decisions that a modeler has to make during model configuration. Note that the modeling decisions in this figure are non-exhaustive. The modeling decisions discussed in this study are highlighted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

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 modeler has to make during model configuration. The impact of those modeling decisions is currently overlooked in most, if not all, model intercomparison studies, and an assessment of their relative importance is lacking.

Several studies in different research disciplines have shown that individual modeling 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 modeling. This study explicitly validates a spin-up modeling decision in the NLDAS project. Pappenberger et al. (2006) explicitly study the effect of upstream boundary conditions and the way bridges are represented in the model on flood inundation predictions. It was demonstrated that the upstream boundary condition had significant impact on the model results. A recent study by Hauser et al. (2017) shows that, dependent on the adopted methodology and input data (besides the choice for a particular model), the European 2015 drought was *more likely*, *less likely*, or *unaffected* by anthropogenic forcing. This demonstrates the large uncertainties that are introduced by methodological choices, such as modeling decisions.

Different model configurations are, however, not always identified as 'modeling 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 modeler. Neal et al. (2010) compared three parallelization methods to model 2D flood inundations, where each method – i.e. modeling decision – had particular drawbacks. Neither Ettema et al. (2009) nor Neal et al. (2010) explicitly discuss their results as modeling decisions.

Some sources of uncertainty, for example concerning model choice (model intercomparison studies) or "optimal" model parameters (calibration studies), have been scrutinized in detail, whereas other sources of uncertainty, such as modeling 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 modeling 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 modeling 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., 2011; Ceola et al., 2015; Mendoza et al., 2016; Mendoza et al., 2015; Fenicia et al., 2016). The attention for this topic in the hydrological sciences is a logical extension of the ambition to improve realism in hydrological models (e.g. Beven, 1989; McDonnell et al., 2007; Clark et al., 2016), which are generally known for their conceptual nature, especially compared to other environmental sciences such as meteorology and oceanography.

In this study, we argue that the choice for a particular model is only one of several modeling decisions, and we illustrate the importance of modeling decisions during model configuration through an example from hydrology. In particular, we investigate the impacts of four modeling decisions on the simulation of a flood and drought event in the Swiss Thur basin. Our aim is to demonstrate the impact of multiple modeling decisions on model results, and to raise awareness to recognize the uncertainty introduced by modeling decisions. A novel contribution of this study is that we systematically investigate and quantify the statistical significance of multiple modeling decisions. Furthermore, the results of this study are particularly relevant because both floods and drought can have a strong societal and economic impact, which water managers aim to mitigate by model predictions.

2. Modeling decisions in hydrology

The sources of uncertainty in hydrological modeling have been an 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 Fig. 1. Hydrologic modeling, however, is also surrounded by modeling decisions, as illustrated in the ellipse in Fig. 1 and discussed in Clark et al. (2011, 2015). These modeling decisions do not only introduce uncertainty not incorporated in the Bayesian evaluation framework in Fig. 1, they also influence the uncertainty estimated with the framework. For example, they determine the prior in a Bayesian framework, or parameter uncertainty as affected by the parameter boundaries.

Many modeling decisions are relevant during the process (sometimes referred to as *'the art'*) of modeling (Fig. 1). In this study we focus

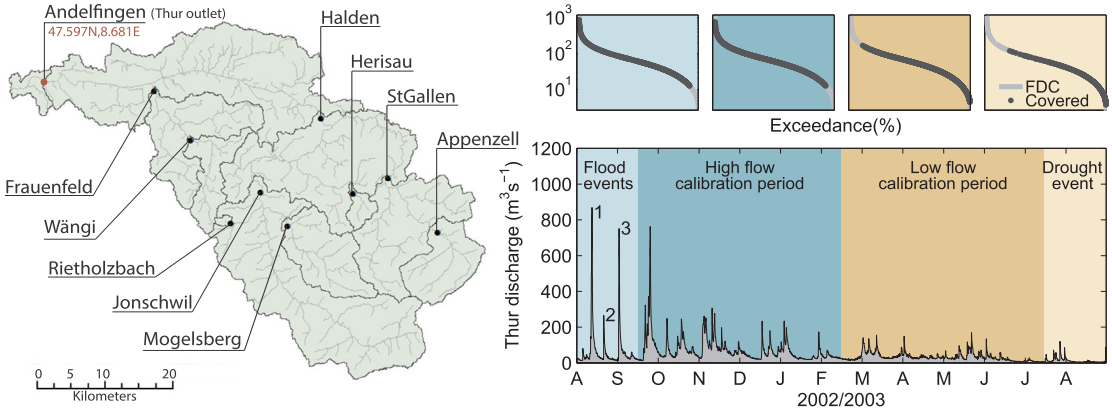

Fig. 2. Left panel: The Thur basin and the nine (nested) sub-basins (see also Table 1). The basins have been named after their gauge location, except for the Thur basin and the Rietholzbach basins. Right panel: Hourly discharge in the Thur basin, with the calibration period and validation period (flood 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.

on four modeling decisions for which the scientific literature provides ambiguous advice to the hydrologic modeler: the spatial resolution, the spatial representation of the forcing, the calibration period, and the performance metric. We aim to illustrate and rank the effects of these decisions on the simulation of a flood and a drought event in the Swiss Thur basin. We recognize that, since we focus on only four modeling decisions and only two events in one basin, the results of our study will be impacted by many other modeling decisions that we (as authors) made during model configuration, and on the specifics of the investigated events. This is further discussed in Section 5.

2.1. Spatial resolution

An important decision that modelers make when setting up a distributed hydrological model is how to represent the spatial distribution. The use of a constant grid is already a first modeling decision, leaving aside options such as hydrological response units (HRU's) or sub-basins. The second decision is 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 led to the call for large-scale hyper-resolution hydrological modeling (Wood et al., 2011). For the Variable Infiltration Capacity (VIC) model (Liang et al., 1994) it was shown that the spatial resolution applied in the scientific literature has increased over the years (Melsen et al., 2016a). Several studies have investigated the effect of spatial resolution (e.g. Haddeland et al., 2002; Liang et al., 2004; Troy et al., 2008; Melsen et al., 2016b), 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 Melsen et al. (2016b) found exactly the opposite for the same model: Both studies applied a different strategy to identify parameters of a distributed 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).

2.2. Spatial representation of forcing

Another important choice for distributed hydrological modeling is the spatial representation of the forcing data. In this paper, 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 subjectively 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 hydrologic models fed with spatially-distributed and uniform forcing data, starting with Wilson et al. (1979), followed by e.g. Beven and Hornberger (1982, 1994, 2008, 2013, 2014). None of the studies based on a large range of basins (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 basin where our study is conducted (see Section 3.1) 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 study, we use spatially interpolated (also a modeling decision!) data based on nine meteorological stations in and around the basin of interest (see Section 3.1). We compare uniformly applied (representing global datasets like WATCH and ERA-Interim) versus spatially-distributed (representing gauge networks or radars) forcing.

2.3. 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, 1986), in which the calibration and validation periods 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 sub-basin 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 modeler to decide on an appropriate calibration period. In this study we compare a high flow calibration period to a low flow calibration period, thus applying the differential split sample test (shown in Fig. 2). Note, however, that the length of the calibration period can also impact the modeling results (see amongst others Vaze et al., 2010; Melsen et al.,

2014). This point is further discussed in Section 5.

2.4. 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 allows for a better weighing of a correlation term, a bias term, and a measure of relative variability. Another approach is to use multiple criteria, e.g. in a Pareto optimization framework (e.g. Madsen, 2003). Since the call for a more process-based evaluation of hydrologic models (McDonnell et al., 2007; Gupta et al., 2008; Clark et al., 2016), hydrologic signatures have become more popular as performance metrics. Hydrologic signatures – e.g. the slope of the flow duration curve, or ecologically relevant streamflow indicators as in Pool et al. (2017) – 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), representing 'average' flow conditions.

3. Methods

In this section, the basin, the investigated extreme events, and the conducted analyses are discussed. A schematic overview of the analysis is provided in Fig. 3.

3.1. Basin and data description

This study has been conducted on the Thur basin (1703 km^2) and its nine (nested) sub-basins of various sizes (Fig. 2 and Table 1), in South-Eastern Switzerland. The Thur basin is characterized by strong topographic variations, with the highest point at the Säntis alpine peak of 2502 m a.s.l., and the lowest point at the outlet in Andelfingen at 356 m a.s.l. The large elevation difference causes orographic effects in the precipitation pattern, and temperature gradients within the basin. The climate in the Thur can be characterized as an alpine/pre-alpine climatic regime with long-term average precipitation varying from 2500 mm yr^{-1} in the Säntis region to 1000 mm yr^{-1} in the lower parts of the basin. In the winter season, some parts of the basin are covered with snow. Because the sub-basins are nested, the ten basins considered in this study are not completely independent. Five basins have upstream nested basins: Frauenfeld, St.Gallen, Jonschwil, Halden, and the Thur (see Fig. 2). The Rietholzbach, the smallest sub-basin of the Thur, is a research basin since 1976 (Seneviratne et al., 2012).

Table 1
Descriptors of the Thur basin and the nine sub-basins.

Basin	Area (km^2)	Mean elev. (m a.s.l.)	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

Hourly discharge observations were obtained for the ten basins for the period 2002–2003. The discharge observation record is short, which also results in short calibration and validation periods. Although a short observation record clearly is a disadvantage in a modelling exercise, we do believe that this short observation period is representative for many practical applications where limited data are available (Seibert and Beven, 2009; Tada and Beven, 2012).

Hourly forcing data for the same period have been obtained from nine meteorological stations in and around the basin. These were spatially interpolated using the WINMET tool (Viviroli et al., 2009). This provided us with spatially-distributed forcing data with a resolution of 1×1 km. In order to compare the distributed forcing data with uniform forcing data (Section 2.2), uniform forcing data have been obtained by spatially averaging the distributed forcing data for every time step. A more elaborate description of the discharge and forcing data used in this study can be found in Melsen et al. (2016b).

3.2. Extreme events

The period 1 August 2002–31 August 2003 is characterized by three flood events in the Thur basin (August, September 2002) as well as the severe 2003 drought (June, July, August 2003); see Fig. 2 and Supporting information S1. The rapid succession of these two contrasting hydrological events makes this period very interesting for our analysis.

Although a range of flood-triggering mechanisms can occur in the alpine/pre-alpine region (Parajka et al., 2010; Hall et al., 2014), the Thur flood events in 2002 were triggered by high rainfall (Zappa and Kan, 2007). The events have an estimated return period of 15 to 20 years. The rainfall was part of a larger system, a so-called VB event (Becker and Grünewald, 2003; Schmocker-Fackel and Naef, 2010)

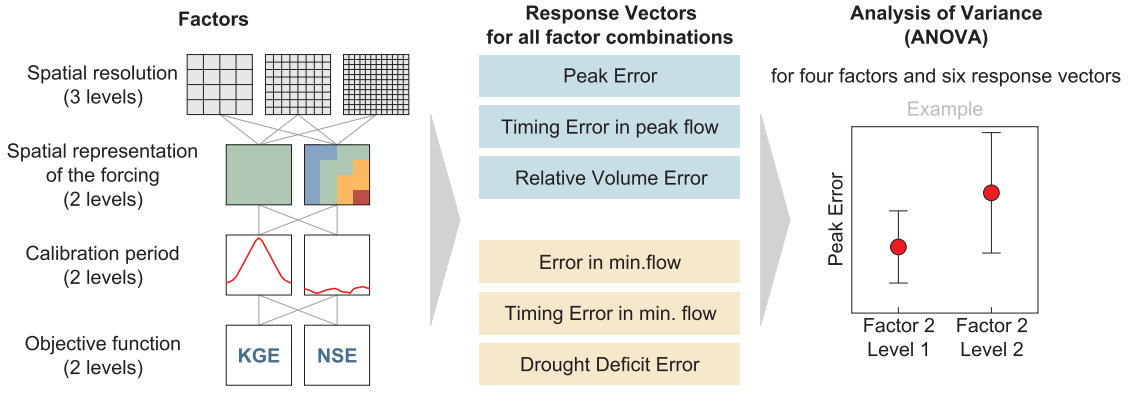

Fig. 3. Flow chart of the methodology. Hydrological models were constructed and calibrated to represent the different factors. Subsequently, response vectors were determined by obtaining the error between modeled and observed (events indicated in Fig. 2) for several flood and drought characteristics, 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.

where a low pressure system was travelling from the Atlantic south-east to the Mediterranean and from there north-east across the Alps. The same system was responsible for severe flooding in central Europe during this period (Becker and Grünewald, 2003).

Contrary to the wet 2002 autumn, the 2003 summer was extremely warm and dry in Western and Central Europe, with Switzerland being among the hottest and driest regions (Zappa and Kan, 2007). The 2003 hydrological drought (that is, anomalies in runoff) was not only caused by precipitation anomalies, but also by high evapotranspiration rates. Precipitation deficits already occurred in the early spring of 2003, thereby declining runoff. However, Seneviratne et al. (2012) demonstrate, based on lysimeter data from the Rietholzbach, that the onset of soil moisture deficit was only from June onwards, caused by evapotranspiration excess, which further declined runoff.

3.3. Model, routing and calibration

The model choice is probably one of the most important decisions a modeler makes. The goal of this study was, however, to show the impact of modeling decisions during model configuration, beyond the model choice. To illustrate this, the impact of modeling decisions has been tested for one widely used hydrological model.

Three Variable Infiltration Capacity (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. In Melsen et al. (2016b) it was shown that three months are enough to remove the effect of initial conditions (model spin-up). Total runoff was routed through the channel network using the MizuRoute routine (Mizukami et al., 2016). 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 analyze the drought event. Finally, six models were configured; three different spatial resolutions, with two different spatial representations of forcing.

To identify the most sensitive parameters in the VIC model, the Distributed Evaluation of Local Sensitivity Analysis (DELSA) method (Rakovec et al., 2014) has been applied on a selection of 28 parameters, including several soil-, vegetation- and snow parameters. For computational efficiency, the sensitivity analysis has been applied to a lumped version of the VIC model of the Thur. However, to investigate the effect of spatial resolution on parameter sensitivity, two lumped models of sub-basins of the Thur have also been subject to a sensitivity analysis: the Jonschwil sub-basin (495 km^2) and the Rietholzbach sub-basin (3.3 km^2). The most sensitive parameters have been identified based on the KGE(Q), the NSE(Q) and the NSE(logQ). Sensitivity analysis on the three lumped models revealed that parameter sensitivity did not change considerably over the assessed scales and objective functions. For the three lumped models, four parameters showed high sensitivity (the first four parameters in Table 2), although the relative sensitivity differed for different spatial scales and objective functions. The results of this sensitivity analysis closely resembled the results of Demaria et al. (2007), who conducted a sensitivity analysis of the VIC model for four basins in the United States. A fifth parameter was added to the selection because Demaria et al. (2007) found it to be highly sensitive (parameter number 5 in Table 2). Furthermore, two MizuRoute-parameters were added to the selection because they control the lateral exchange of water between grid cells. A more elaborate description of the sensitivity analysis, as well as an overview of the included parameters and their boundaries, can be found in Melsen et al. (2016b).

The seven selected parameters in Table 2 of the VIC model and the routing routine have been sampled 3150 times, using a Hierarchical Latin Hypercube Sample (Vořechovský, 2015). A Latin Hypercube sampling strategy was chosen because this is more efficient than random sampling strategies. The advantage of the hierarchical method is that the size of the sample can be extended step by step. Inherent to the Hierarchical Latin Hypercube Sample (Vořechovský, 2015) is that every sample extension is twice as large as the previous sample. The starting sample size was set at 350, sampled based on a space-filling criterion with a uniform prior. The next sample size was (350×2 plus the first sample, 350) 1,050 samples in total. With a Kolmogorov-Smirnov test, it was tested whether the cumulative distribution function (CDF) of the objective functions significantly changed with an increased sample size. It was shown that the CDF did not significantly change from 1050 samples to 3150 samples, indicating that 3150 samples were enough to cover the parameter space (Melsen et al., 2016b). The VIC model has been run with all 3150 parameter samples. The seven sampled parameters (Table 2) have been applied uniformly over the basin, 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, swisstopo license JA100118) and the Harmonized World Soil Database (FAO et al., 2012). A more elaborate description of these data can be found in Melsen et al. (2016b).

The different model configurations have been run with the full parameter sample over the model period (Fig. 2). 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.

Finally, 'calibration' (selection of behavioral parameter sets) is performed by identifying the best performing 1% (32 runs) of the 3,150 runs, for each case separately. The different cases consist of all the different combinations of the three investigated spatial resolutions with two different spatial representations of the forcing. The best 1% was selected based on the NSE(Q) for the flood events, and the NSE(logQ) for the drought event, and based on the KGE(Q) for both events to investigate the influence of the performance metric (one of the modeling decisions). Because the calibration period is also one of the modeling decisions investigated, the best performing 32 runs have been determined separately for the high flow calibration period and the low flow calibration period. The two calibration periods are indicated in Fig. 2.

An overview of the model performance, expressed in NSE(Q) and NSE(logQ) for the respective validation periods is given in Supporting information S2. Although direct comparison with obtained model performance in other studies is not fair due to different model periods,

Table 2
Sampled model parameters. The parameter boundaries are suggested by the developers of the VIC model (http://vic.readthedocs.io/en/develop/).

	Parameter	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 $d_{s,max}$ where non-linear base flow starts
3	d_m	mm d^{-1}	1.0	50	Maximum velocity of the base flow
4	$expt_2$	–	4.0	18.0	Exponent of the Brooks-Corey drainage equation for layer 2
5	$Depth_2$	m	$Depth_1$ + 0.1	$Depth_1$ + 3	Depth of soil layer 2
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

calibration strategies, and objective functions, it can provide context on how well our model is able to capture the basin dynamics. The highest achieved model performance in the Thur basin for the flood validation period in our set-up with the VIC model is NSE(Q) = 0.85, and for the drought validation period NSE(logQ) = 0.58. This compares with the model performance obtained with the SWAT model by Yang et al. (2007), with a NSE(Q) of 0.77 for the calibration period, and is for the drought-period lower than performances obtained with the PREVAH model by Zappa and Kan (2007), with a NSE(Q) and NSE(logQ) higher than 0.85 for both calibration and validation period. The highest model performance for the Rietholzbach basin for the validation period of this study is NSE(Q) = 0.53 and NSE(logQ) = 0.63. This is lower than model performances reported by Gurtz et al. (2003) for the Rietholzbach using the WaSiM-ETH model (NSE(Q) = 0.80, NSE(logQ) = 0.82) and the PREVAH model (NSE(Q) = 0.71,NSE(logQ) = 0.89) over the validation period 1981–1998. Melsen et al. (2014) applied a parsimonious stage-discharge-model to the Rietholzbach and reported NSE (Q) = 0.69, and NSE(logQ) = 0.74 for the validation period. Given the relatively limited part of the parameter space that was explored in this study compared to the calibration strategies applied in the studies cited, model performance was expected to be somewhat lower. The best performing runs are, however, in range with the literature and therefore seem to be able to mimic the behavior of the Thur basin and the Rietholzbach basin quite well.

By selecting the best performing 1% of the runs, all model configurations have an equally-sized set of runs defined as behavioral. This improves the robustness of the statistical test (Analysis of Variance). An implication of this approach 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 modeling decisions, rather than evaluating the model performance directly.

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). Finally, the calibrated model configurations were validated for events that were not included in the calibration periods: three flood events and one drought event (Fig. 2).

3.4. Flood and drought characteristics

To investigate the effect of subjective modeling decisions on extreme hydrological events, the error in flood and drought characteristics between observations and simulations were investigated for different model configurations (Fig. 3). The flood characteristics have been validated for three flood events (Fig. 2), and the drought characteristics have been validated for one drought event (Fig. 2).

The three main characteristics of a flood event are the peak height, the timing, and the volume (Lobligeois et al., 2014). For each behavioral 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:

$$\Delta Q_p = Q^p_{\text{sim}} - Q^p_{\text{obs}}, \tag{1}$$

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

$$\Delta t_p = t(Q^p_{\text{sim}}) - t(Q^p_{\text{obs}}), \tag{2}$$

where $t(Q^p_{\text{sim}})$ is the timing of the modeled 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 modeled discharge:

$$\text{RVE} = \frac{\sum (Q_{\text{sim}} - Q_{\text{obs}})}{\sum Q_{\text{obs}}}, \tag{3}$$

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 (modeled) 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:

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

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 (Fig. 3). 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 (Supporting information S1). 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

$$\Delta Q_{min} = Q_{min,\text{sim}} - Q_{min,\text{obs}}, \tag{5}$$

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;

$$\Delta t_{min} = t(Q^{min}_{\text{sim}}) - t(Q^{min}_{\text{obs}}). \tag{6}$$

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:

$$d(t) = \begin{cases} \tau(t) - Q(t) & \text{if } Q(t) < \tau(t) \\ 0 & \text{if } Q(t) \geqslant \tau(t). \end{cases}$$

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

$$D = \sum_{t=1}^{T} d(t)\cdot\Delta t. \tag{7}$$

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}:

$$\Delta D = D_{\text{sim}} - D_{\text{obs}}. \tag{8}$$

3.5. Analysis of variance

After computing the error in flood and drought characteristics for all the behavioral runs and for the different model configurations as shown in Fig. 3, 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 modeling decisions), and has been applied to six response vectors (the errors in flood and drought characteristics), as shown in Fig. 3. If the probability $p < 0.05$, the factor was assumed to have significant impact on the response vector.

The aim of this study was to demonstrate that modeling decisions significantly impact the simulation of two hydrological extremes, for a case-study in the Thur and its nine sub-basins. 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 modeling 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 modeling decisions, it was also investigated how the decisions impact the parameter distribution, using ANOVA.

4. Results

4.1. Flood characteristics

In this section we focus on three flood events (Fig. 2). Fig. 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. Fig. 4 also shows to what extent the impact of modeling decisions on the error in characteristics of the three flood events is significant, using ANOVA (panels c, f and i).

Fig. 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 modeling decisions on the timing error (Fig. 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 (Fig. 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 modeling decisions. A summary of the results is given in Table 3.

4.2. Drought characteristics

Fig. 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 and the nine sub-basins. The results show that the calibration period has a large impact on the error in drought characteristics. Fig. 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

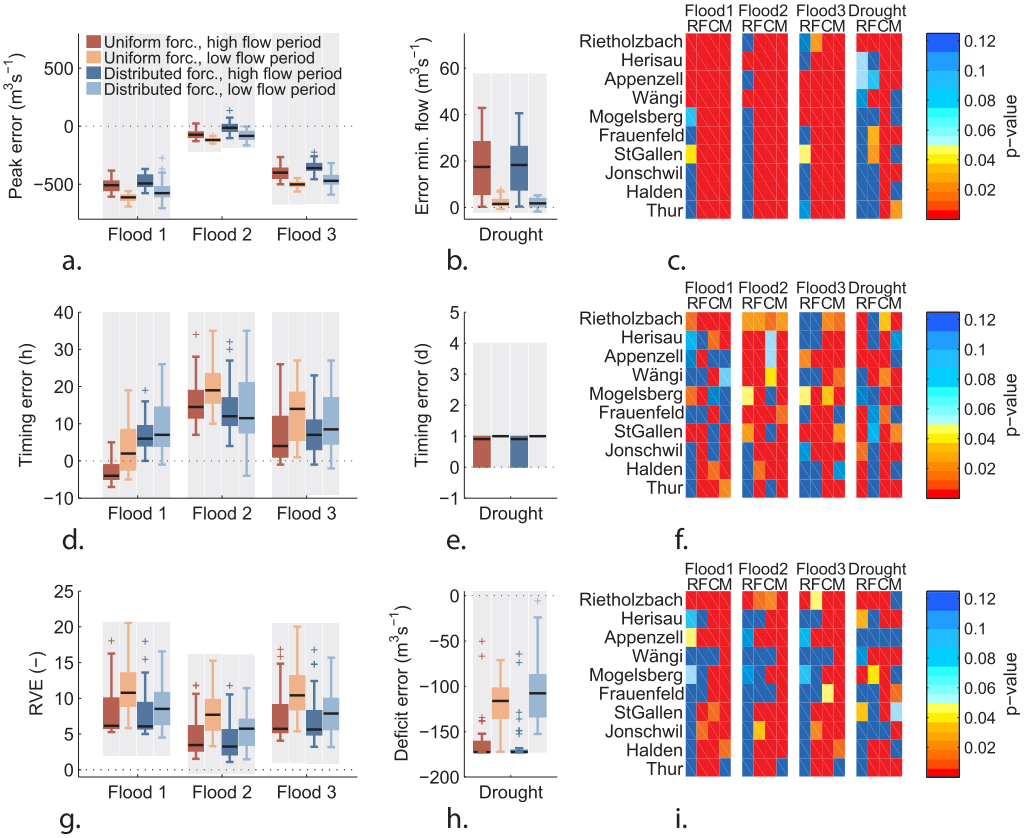

Fig. 4. The impact of the investigated modeling decisions on the error in the three flood characteristics (peak error, timing error, relative volume error denoted as RVE) and the three drought characteristics (error in minimum flow, timing error in minimum flow, and deficit). Panel a, d, g: The distribution of the behavioral sets (best 1% expressed in NSE(Q)) for the error in flood characteristics for three flood events in the Thur (1×1 km resolution, NSE(Q) as objective function). Panels b, e, h: The distribution of the behavioral sets (best 1% expressed in NSE(logQ)) for the error in drought characteristics for the drought event in the Thur (1×1 km resolution, NSE(logQ) as objective function). The dashed line indicates the optimum (no difference between modeled 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 basin size (area).

Table 3
Number of basins (out of 10 in total) for which the modeling 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 ($p < 0.05$). Note that for the error in flood characteristics, the average for the three flood events is given, since the number of basins for which the modelling decisions significantly impacted the flood characteristics could differ per event.

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	

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 (Fig. 4f); the impact of the modeling decisions 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 (Fig. 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 3 reveals that the timing error in the minimum flow experiences most impact from the investigated subjective modeling decisions.

4.3. Impact on parameter distribution

Table 4 provides an overview of the percentage of basins for which the distribution of the sampled parameters (Table 2) was significantly affected by any of the four modeling 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 $Depth_2$, the depth of soil layer 2, and C, the velocity parameter of the routing scheme.

For the drought event, calibration period is by far the most important modeling decision that determines the parameter distribution.

Table 4
Number of basins (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), or performance metric (M).

Parameter	NSE(Q) (hourly time step)				NSE(logQ) (daily time step)			
	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
$expt_2$	1	2	9	6	2	3	8	4
$Depth_2$	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

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 modeling decisions.

The affected parameters differ for flood and drought events. For the flood events, mainly the parameters impacting the response time are influenced by the modeling decisions, whereas for the low flows the infiltration and base flow parameter are mostly affected by the decisions. For the low flows, the calibration period is by far the most important decision, whereas spatial resolution only plays a very minor role. For high flows the calibration period is most important, followed closely by the performance metric, and the spatial representation of the forcing.

5. Discussion

The main point of this study was to demonstrate that subjective modeling decisions, beyond the model choice, affect the simulation of a flood and drought event in the Swiss Thur basin. In Section 5.1, we aim to relate the results to relevant hydrological processes during the investigate flood and drought event. Section 5.2 discusses several decisions that we, authors and modelers, made during the design of this study.

5.1. Relation between results and hydrological processes

Preferably, we would be able to couple the impact of the investigated modeling decisions on our model simulations to the hydrological functioning of our system, or to particular hydrological processes. This could help to substantiate certain modeling decisions. Although this study is a case-study with only one flood and drought event/type investigated for ten nested basins with comparable climate and land-use, several links can still be identified.

The clearest example is the limited impact of the spatial representation of the forcing on the error in drought characteristics. The studied hydrological drought was partly caused by a lack of precipitation (Seneviratne et al., 2012). 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 basin. The results seem to follow the line of expectation. However, Seneviratne et al. (2012) also demonstrates that high evapotranspiration (ET) rates played an important role in the onset of the 2003 drought (at least in terms of soil moisture). Gurtz et al. (1999) emphasize that the spatial resolution of the model is very important when modelling ET-rates in mountainous regions. For the Thur, Gurtz et al. (1999) recommend no coarser resolution than 2×2 km and Zappa (2002) even recommends a resolution of max 1×1 km in the hilly sub-regions of the Thur, which can explain why the modelling decision on the resolution of the model had more impact on the drought simulations compared to the flood simulations and compared to the spatial representation of the forcing for the drought simulation. The calibration period and the performance metric had most influence on the drought simulation, which stresses the importance of modelling decisions concerning calibration strategy. For other hydrological drought typologies, for example a cold snow season drought (below-average temperature at the end of the snow season causing a delay in snow melt; Vann Loon and Van Lanen, 2012), ET-rates might be less relevant and therefore spatial resolution of the model might be less important. On the other hand, a high spatial resolution might still be needed to capture highly heterogeneous snow melt processes.

The three high flow events that were studied in the Thur basin and the nine subbasins were caused by rainfall from a large low pressure system, although the precipitation still displayed a high spatial variation (see Fig. 6). Therefore, it can be understood that the spatial representation of the forcing (i.e., lumped or distributed) did have substantial impact on the flood simulations in most basins. On the other hand, the spatial resolution of the model only had minor influence on

the flood simulations. With a high spatial resolution, the model can better capture spatial variability in soil moisture which can influence surface-runoff processes, but perhaps spatial variability in soil moisture was limited because the wet conditions extended over the complete basin. For other flood typologies, such as flash floods that usually have a very local character, the spatial resolution might play a more important role. As for the drought simulations, calibration period and performance metric also had most impact on the flood simulations in this study, underlining the important role of a deliberate calibration in rainfall-runoff modeling.

Although the results of this study can be explained through process-reasoning, this section does show that this case study does not yet provide enough insights to draw robust conclusions to substantiate hydrological modelling decisions concerning drought and floods. Different flood and drought typologies, but also a different climate or spatial variation in elevation or land use (given the variation in results among the sub-basins), could lead to different decisions that have most impact on the simulations. More research is needed to provide insights into the impact of modeling decisions on hydrological extremes.

5.2. Subjectivity in our study-design

The results of this study also depend on model decisions that we, as modelers and authors, made for the experimental set-up. We only investigated the effect of four modeling 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, the choice for spatially interpolated station data as forcing, and the selection of the best 1% of the model runs as 'behavioral'. Further, we also made important decisions on the parameters included in the sensitivity analysis, their boundaries and the sampling strategy adopted.

The uniform application of the sampled parameters can decrease the effect of spatial resolution. Most likely, the spatial resolution will become a more important modeling 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 for the VIC model. Therefore, sampling parameters in a (semi-)distributed fashion was out of reach in our current set-up.

The length of the calibration period was fixed to five months. For the Rietholzbach basin this was shown to be a sufficient period to obtain stable parameters for a parsimonious model (Melsen et al., 2014). In this study, the calibration period is one of the most influencing decisions investigated, which implies that, in this case-study, a five-month calibration period is not sufficient to obtain stable parameters. This effect can be even stronger when shorter calibration periods are explored. Therefore, the analysis as shown in Fig. 3 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. Fig. 5 shows that the investigated modeling decisions still have significant impact on the error in characteristics of the hydrological extremes for a shorter calibration period. Most modeling 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 20% (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.

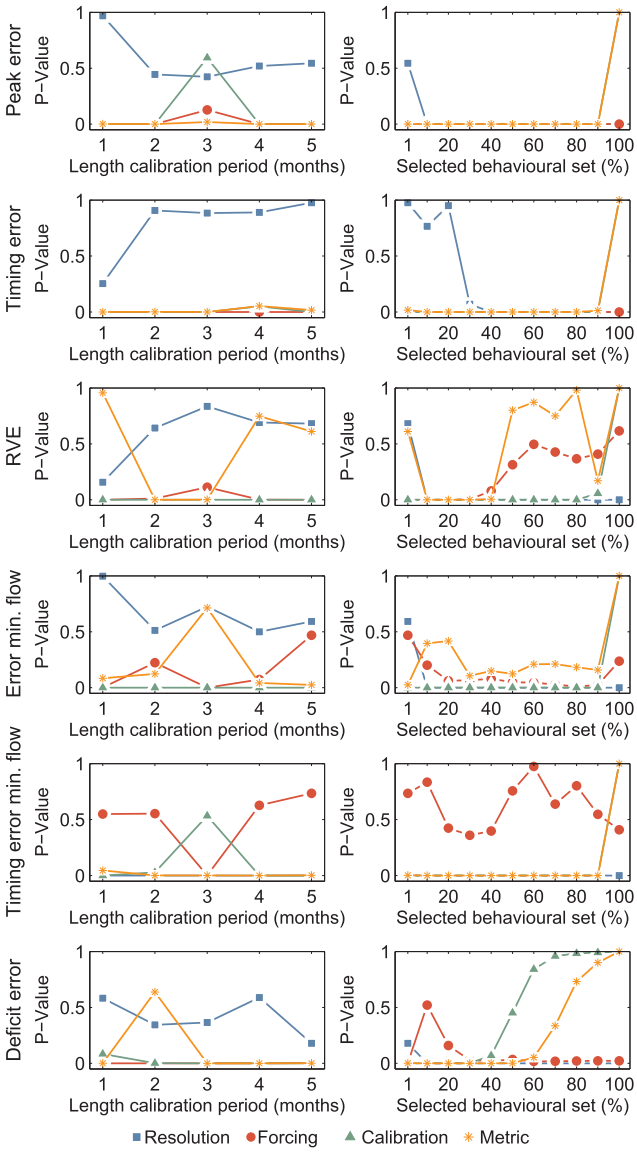

Fig. 5. ANOVA p-value versus length of the calibration period (left panels) and size of the sample selected as 'behavioral' (right panels) for the four investigated modeling decisions. For clarity, only the results for the Thur basin are shown, and for the error in flood characteristics only the results for Flood event 1 (see Fig. 2).

The selection of the best 1% of the sample as 'behavioral' is not so much a modeling decision as it is a decision in the research set-up. To investigate the effect of this choice, the analysis as shown in Fig. 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%. Fig. 5 shows that choosing a larger sample affects the results, but in most cases it increases the significance level of the modeling 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 'behavioral'. 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 modeling decisions on hydrological extremes could change if other modeling decisions or experimental configurations would be adopted. Nevertheless, the conclusion that subjective modeling decisions significantly impact the simulation of hydrological extremes remains valid.

6. Summary and conclusion

Computational models in Earth and environmental sciences have to deal with uncertainty, which is partially augmented by subjective modeling decisions (e.g., model choice, performance metric selection). The impact of model choice on model results is generally acknowledged, whereas the uncertainty introduced by modeling decisions during the configuration of the model is often neglected. In this study we show, with an example for a flood and drought event in the Swiss Thur basin, that modeling decisions, beyond the model choice, affect model results significantly.

We investigated four modeling 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 for a flood and drought event in the Swiss Thur basin. Both extreme events were mainly impacted by the two calibration-decisions: the calibration period and the performance metric. For the flood events, also the spatial representation of the forcing was important, which could be explained by the spatial variability of the precipitation during the studied flood events. For the drought event, the role of spatial resolution of the model could be related to evapotranspiration processes that played a role during the onset of the studied drought event. However, extreme events such as floods and drought can be triggered by different processes, and a different typology of any of the events could therefore lead to a different impact of the modeling decisions on the model simulations. Furthermore, the impact of the investigated modeling decisions differed from (sub-)basin to (sub-)basin, even though the compared basins had much in common in terms of climate and land-use. Therefore, more research is needed to provide insights into the impact of subjective modeling decisions on model simulations. However, in spite of the limitations of this study, our results do undeniably show that modeling 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 modeler.

A better understanding of the uncertainty in hydrological model results can improve the robustness of water management decisions (McMillan et al., 2017). Many model studies therefore already include some form of uncertainty analysis, by comparing several models or several parameter sets. Modeling decisions, however, are hardly ever included in these analyses, whereas this study has shown that their impact on the results is significant. An evaluation of the impact of modeling decisions helps in estimating the value of model results. This paper provides an example on how to conduct such an assessment for a specific hydrologic application. Further, it is critical to constrain the spectrum of options (or hypotheses) for a particular modeling 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 behavior of environmental systems. Uncertainty associated to a particular modeling decision can be characterized through multiple working hypotheses (Clark et al., 2011; Beven et al., 2012), carefully selected to avoid over-confident portrayals of environmental processes.

Conflict of interest

None.

Acknowledgments

We like to thank the Swiss Federal Office for the Environment (FOEN) for providing the discharge data for the Thur basin and eight sub-basins. We would also like to thank Martin Hirschi and Dominic Michel from ETH Zürich for providing the discharge data for the Rietholzbach. The required forcing data (precipitation, incoming shortwave radiation, temperature, vapor pressure, wind) have been kindly provided by the Swiss Federal Office for Meteorology and Climatology (MeteoSwiss).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.jhydrol.2018.11.046.

References

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

Becker, A., Grünewald, U., 2003. Flood risk in Central Europe. Science 300, 1099. https://doi.org/10.1126/science.1083624.

Bennett, N., Croke, B., Guariso, G., Guillaume, J., Hamilton, S., Jakeman, A., Marsili-Libelli, S., Newham, L., Norton, J., Perrin, C., Pierce, S., Robson, B., Seppelt, R., Voinov, A., Fath, B., Andreassian, V., 2013. Characterising preformance of environmental models. Env. Modell. Softw. 40, 1–20. https://doi.org/10.1016/j.envsoft.2012.09.011.

Beven, K., 1989. Changing ideas in hydrology – the case of physically-based models. J. Hydrol. 105, 157–172. https://doi.org/10.1016/0022-1694(89)90101-7.

Beven, K., Binley, A., 1992. The future of distributed models: model calibration and uncertainty prediction. Hydraul. Process. 6, 279–298. https://doi.org/10.1002/hyp.3360060305.

Beven, K., Hornberger, G., 1982. Assessing the effect of spatial pattern of precipitation in modeling stream flow hydrographs. Water Res. Bul. 18, 823–829. https://doi.org/10.1111/j.1752-1688.1982.tb00078.x.

Beven, K., Smith, P., Westerberg, I., Freer, J., 2012. Comment on "Pursuing the method of multiple working hypothesesfor hydrological modeling by M.P. Clark et al. Water Resour. Res. 48. https://doi.org/10.1029/2012WR012282.

Ceola, S., Arheimer, B., Baratti, E., Blöschl, G., Capell, R., Castellarin, A., Freer, J., Han, D., Hrachowitz, M., Hundecha, Y., Hutton, C., Lindström, G., Montanari, A., Nijzink, R., Parajka, J., Toth, E., Viglione, A., Wagener, T., 2015. Virtual laboratories: new opportunities for collaborative water science. Hydr. Earth Syst. Sci. 19, 2101–2117. https://doi.org/10.5194/hess-19-2101-2015.

Clark, M., Kavetski, D., 2010. Ancient numerical daemons of conceptual hydrological modeling: 1. Fidelity and efficiency of time stepping schemes. Water Resour. Res. 46. https://doi.org/10.1029/2009WR008894.

Clark, M., Kavetski, D., Fenicia, F., 2011. Pursuing the method of multiple working hypotheses for hydrological modeling. Water Resour. Res. 47. https://doi.org/10.1029/2010WR009827.

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

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

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

Cosgrove, B.A., Lohmann, D., Mitchell, K., Houser, P., Wood, E., Schaake, J., Robock, A., Sheffield, J., Duan, Q., Luo, L., Higgins, R.W., Pinker, R., Tarpley, J., 2003. Land surface model spin-up behavior in the North American Land Data Assimilation System (NLDAS). J. Geophys. Res. 108. https://doi.org/10.1029/2002JD003316.

Crout, N., Tarsitano, D., Wood, A., 2009. Is my model too complex? Evaluating model formulation using model reduction. Env. Modell. Softw. 24, 1–7. https://doi.org/10.1016/j.envsoft.2008.06.004.

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

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

2009GL038110.
FAO, IIASA, ISRIC, ISSCAS, JRC, 2012. Harmonized World Soil Database (version 1.2). Technical Report. AO, Rome, Italy and IIASA, Laxenburg, Austria. doi: http://www.fao.org/soils-portal/soil-survey/soil-maps-and-databases/harmonized-world-soil-database-v12/en/.
Fenicia, F., Kavetski, D., Savenije, H., Pfister, L., 2016. From spatially variable streamflow to distributed hydrological models: analysis of key modeling decisions. Water Resour. Res. 52, 1–35. https://doi.org/10.1002/2015WR017398.
Freni, G., Mannina, G., Viviani, G., 2009. Urban runoff modelling uncertainty: comparison among Bayesian and pseudo-Bayesian methods. Environ. Modell. Softw. 1100–1111. https://doi.org/10.1016/j.envsoft.2009.03.003.
Gupta, H., Kling, H., Yilmaz, K., Martinez, G., 2009. Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. J. Hydrol. 377, 80–91.
Gupta, H., Wagener, T., Liu, Y., 2008. Reconciling theory with observations: elements of a diagnostic approach to model evaluation. Hydraul. Process. 22, 3802–3813. https://doi.org/10.1002/hyp.6989.
Gurtz, J., Baltensweiler, A., Lang, H., 1999. Spatially distributed hydrotope-based modelling of evapotranspiration and runoff in mountainous basins. Hydrol. Process. 13, 2751–2768.
Gurtz, J., Zappa, M., Jasper, K., Lang, H., Verbunt, M., Badoux, A., Vitvar, T., 2003. A comparative study in modelling runoff and its components in two mountainous catchments. Hydraul. Process. 17, 297–311. https://doi.org/10.1002/hyp.1125.
Haddeland, I., Matheussen, B., Lettenmaier, D., 2002. Influence of spatial resolution on simulated streamflow in a macroscale hydrologic model. Water Resour. Res. 38. https://doi.org/10.1029/2001WR000854.
Hall, J., Arheimer, B., Borga, M., Brázdil, R., Claps, P., Kiss, A., Kjeldsen, T., Kriaučiuniene, J., Kundzewicz, Z., Lang, M., Llasat, M., Macdonald, N., McIntyre, N., Mediero, L., Merz, B., Merz, R., Molnar, P., Montanari, A., Neuhold, C., Parajka, J., ao, R.P., Plavcová1, L., Rogger, M., Salinas, J., Sauquet, E., Schär, C., Szolgay, J., Viglione, A., Blöschl, G., 2014. Understanding flood regime changes in Europe: a state-of-the-art assessment. Hydrol. Earth Syst. Sci. 18, 2735–2772.. https://doi.org/10.5194/hess-18-2735-2014.
Hauser, M., Gudmundsson, L., Orth, R., Jézéquel, A., Haustein, K., Vautard, R., van Oldenborgh, G., Wilcox, L., Seneviratne, S., 2017. Methods and model dependency of extreme event attribution: the 2015 european drought. Earth's Future 5, 1034–1043. https://doi.org/10.1002/2017EF000612.
Hisdal, H., Tallaksen, L.M., Clausen, B., Peters, E., Gustard, A., 2004. Hydrological drought processes and estimation methods for streamflow and groundwater. Elsevier Science. Chapter in Hydrological Drought Characteristics. pp. 139–198.
Holländer, H., Blume, T., Bormann, H., Buytaert, W., Chirico, G., Exbrayat, J., Gustafsson, D., Hölzel, H., Kraft, P., Stamm, C., Stoll, S., Blöschl, G., Flühler, H., 2009. Comparative predictions of discharge from an artificial catchment (Chicken Creek) using sparse data. Hydrol. Earth Syst. Sci. 13, 2069–2094. https://doi.org/10.5194/hess-13-2069-2009.
Joussaume, S., Taylor, K., Braconnot, P., Mitchell, J., Kutzbach, J., Harrison, S., Prentice, I., Broccoli, A., Abe-Ouchi, A., Bartlein, P., Bonfils, C., Dong, B., Guiot, J., Henerich, K., Hewitt, C.D., Jolly, D., Kim, J., Kislov, A., Kitoh, A., Loutre, M., Masson, V., McAvaney, B., McFarlane, N., de Noblet, N., Peltier, W., Peterschmitt, J., Pollard, D., Rind, D., Royer, J., Schlesinger, M., Syktus, J., Thompson, S., Valdes, P., Vettoretti, G., Webb, R., Wyputta, U., 1999. Monsoon changes for 6000 years ago: results of 18 simulationsfrom the Paleoclimate Modeling Intercomparison Project (PMIP). Geophys. Res. Lett. 26, 859–862. https://doi.org/10.1029/1999GL900126.
Kavetski, D., Clark, M., 2010. Ancient numerical daemons of conceptual hydrological modeling: 2. Impact of time stepping schemes on model analysis and prediction. Water Resour. Res. 46. https://doi.org/10.1029/2009WR008896.
Klemeš, V., 1986. Operational testing of hydrological simulation models. Hydrol. Sci. J. 31, 13–24. https://doi.org/10.1080/02626668609491024.
Krueger, T., Page, T., Hubacek, K., Smith, L., Hiscock, K., 2012. The role of expert opinion in environmental modelling. Environ. Modell. Softw. 36, 4–18. https://doi.org/10.1016/j.envsoft.2012.01.011.
Li, C.Z., Zhang, L., Wang, H., Zhang, Y.Q., Yu, F.L., Yan, D.H., 2012. The transferability of hydrological models under nonstationary climatic conditions. Hydrol. Earth Syst. Sci. 16, 1239–1254. https://doi.org/10.5194/hess-16-1239-2012.
Liang, X., Guo, J., Leung, L., 2004. Assessment of the effects of spatial resolutions on daily water flux simulations. J. Hydrol. 298, 287–310. https://doi.org/10.1016/j.jhydrol.2003.07.007.
Liang, X., Lettenmaier, D.P., Wood, E.F., Burges, S.J., 1994. A simple hydrologically based model of land surface water and energy fluxes for general circulation models. J. Geophys. Res. 99, 14415–14458.
Liu, Y., Gupta, H.V., 2007. Uncertainty in hydrologic modeling: towards an integrated data assimilation framework. Water Resour. Res. 43. https://doi.org/10.1029/2006WR005756.
Lobligeois, F., Andréassian, V., Perrin, C., Tabary, P., Loumagne, C., 2014. When does higher spatial resolution rainfall information improve streamflow simulation? An evaluation using 3620 flood events. Hydrol. Earth Syst. Sci. 18, 575–594. https://doi.org/10.5194/hess-18-575-2014.
Madsen, H., 2003. Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. Adv. Water Resour. 26, 205–216. https://doi.org/10.1016/S0309-1708(02)00092-1.
Maier, H., Dandy, G., 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. https://doi.org/10.1016/S1364-8152(99)00007-9.
McDonnell, J., Sivapalan, M., Vaché, K., Dunn, S., Grant, G., Haggerty, R., Hinz, C., Hooper, R., Kirchner, J., Roderick, M.L., Selker, J., Weiler, M., 2007. Moving beyond heterogeneity and process complexity: a new vision for watershed hydrology. Water Resour. Res. 43. https://doi.org/10.1029/2006WR005467.
McMillan, H., Seibert, J., Petersen-Overleir, A., Lang, M., White, P., Snelder, T., Rutherford, K., Krueger, T., Mason, R., Kiang, J., 2017. How uncertainty analysis of streamflow data can reduce costs and promote robust decisions in water management applications. Water Resour. Res. 53, 5220–5228. https://doi.org/10.1002/2016WR020328.
Melsen, L., Teuling, A., Torfs, P., Uijlenhoet, R., Mizukami, N., Clark, M., 2016a. HESS opinions: the need for process-based evaluation of large-domain hyper-resolution models. Hydrol. Earth Syst. Sci. 20, 1069–1079. https://doi.org/10.5194/hess-20-1069-2016.
Melsen, L., Teuling, A., Torfs, P.J., Zappa, M., Mizukami, N., Clark, M., Uijlenhoet, R., 2016b. Representation of spatial and temporal variability in large-domain hydrological models: case study for a mesoscale prealpine basin. Hydrol. Earth Syst. Sci. 20, 2207–2226. https://doi.org/10.5194/hess-20-2207-2016.
Melsen, L., Teuling, A., van Berkum, S., Torfs, P., Uijlenhoet, R., 2014. Catchments as simple dynamical systems: a case study on methods and data requirements for parameter identification. Water Resour. Res. 50, 5577–5596. https://doi.org/10.1002/2013WR014720.
Mendoza, P., Clark, M., Mizukami, N., Gutmann, E., Arnold, J., Brekke, L., Rajagopalan, B., 2016. How do hydrologic modeling decisions affect the portrayal of climate change impacts? Hydraul. Process. 30, 1071–1095. https://doi.org/10.1002/hyp.10684.
Mendoza, P., Clark, M., Mizukami, N., Newman, A., Barlage, M., Gutmann, E., Rasmussen, R., Rajagopalan, B., Brekke, L., Arnold, J., 2015. Effects of hydrologic model choice and calibration on the portrayal of climate change impacts. J. Hydrometeorol. 16, 762–780. https://doi.org/10.1175/JHM-D-14-0104.1.
Merz, R., Parajka, J., Blöschl, G., 2011. Time stability of catchment model parameters: Implications for climate impact analyses. Water Resour. Res. 47. https://doi.org/10.1029/2010WR009505.
Mizukami, N., Clark, M.P., Sampson, K., Nijssen, B., Mao, Y., McMillan, H., Viger, R.J., Markstrom, S.L., Hay, L.E., Woods, R., Arnold, J.R., Brekke, L.D., 2016. mizuRoute version 1: a river network routing tool for a continental domain water resources applications. Geosci. Model Dev. 9, 2223–2238. https://doi.org/10.5194/gmd-9-2223-2016.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models, I. A discussion of principles. J. Hydrol. 10, 282–290.
Neal, J., Fewtrell, T., Bates, P., Wright, N., 2010. A comparison of three parallelisation methods for 2D flood inundation models. Environ. Modell. Softw. 25, 398–411. https://doi.org/10.1016/j.envsoft.2009.11.007.
Nicótina, L., Alessi Celegon, E., Rinaldo, A., Marani, M., 2008. On the impact of rainfall patterns on the hydrologic response. Water Resour. Res. 44. https://doi.org/10.1029/2007WR006654.
Obled, C., Wendling, J., Beven, K., 1994. The sensitivity of hydrological models to spatial rainfall patterns: an evaluation using observed data. J. Hydrol. 159, 305–333. https://doi.org/10.1016/0022-1694(94)90263-1.
Oreskes, N., Shrader-Frechette, K., Belitz, K., 1994. Verification, validation, and confirmation of numerical models in the Earth Sciences. Science 263, 641–646. https://doi.org/10.1126/science.263.5147.641.
Ott, R., Longnecker, M., 2010. An introduction to statistical methods and data analysis, sixth ed. Brooks/Cole, Cengage Learning.
Pappenberger, F., Matgen, P., Beven, K., Henry, J., Pfister, L., Fraipont, P., 2006. Influence of uncertain boundary conditions and model structure on flood inundation predictions. Adv. Water Resour. 29, 1430–1449. https://doi.org/10.1016/j.advwatres.2005.11.012.
Parajka, J., Kohnová, S., Bálint, G., Barbuc, M., Borga, M., Claps, P., Cheval, S., Dumitrescu, A., Gaume, E., Hlavcŏvá, K., Merz, R., Pfaundler, M., Stancalie, G., Szolgay, J., Blöschl, G., 2010. Seasonal characteristics of flood regimes across the Alpine-Carpathian range. J. Hydrol. 394, 78–89. https://doi.org/10.1016/j.jhydrol.2010.05.015.
Pool, S., Vis, M., Knight, R., Seibert, J., 2017. Streamflow characteristics from modeled runoff time series – importance of calibration criteria selection. Hydrol. Earth Syst. Sc. 21, 5443–5457. https://doi.org/10.5194/hess-21-5443-2017.
Rakovec, O., Hill, M.C., Clark, M.P., Weerts, A.H., Teuling, A.J., Uijlenhoet, R., 2014. Distributed evaluation of local sensitivity analysis (DELSA), with application to hydrologic models. Water Resour. Res. 50, 409–426. https://doi.org/10.1002/2013WR014063.
Samaniego, L., Kumar, R., Attinger, S., 2010. Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. Water Resour. Res. 46. https://doi.org/10.1029/2008WR007327.
Sawicz, K., Wagener, T., Sivapalan, M., Troch, P.A., Carrillo, G., 2011. Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA. Hydrol. Earth Syst. Sci. 15, 2895–2911. https://doi.org/10.5194/hess-15-2895-2011.
Schaefli, B., Gupta, H., 2007. Do Nash values have value? Hydraul. Process. 21, 2075–2080. https://doi.org/10.1002/hyp.6825.
Schmocker-Fackel, P., Naef, F., 2010. More frequent flooding? Changes in flood frequency in Switzerland since 1850. J. Hydrol. 381, 1–8. https://doi.org/10.1016/j.jhydrol.2009.09.022.
Seibert, J., Beven, K.J., 2009. Gauging the ungauged basin: how many discharge measurements are needed? Hydrol. Earth. Syst. Sci. 13, 883–892. https://doi.org/10.5194/hess-13-883-2009.
Seneviratne, S.I., Lehner, I., Gurtz, J., Teuling, A.J., Lang, H., Moser, U., Grebner, D., Menzel, L., Schroff, K., Vitvar, T., Zappa, M., 2012. Swiss prealpine Rietholzbach research catchment and lysimeter: 32 year time series and 2003 drought event. Water Resour. Res. 48. https://doi.org/10.1029/2011WR011749.
Tada, T., Beven, K.J., 2012. Hydrological model calibration using a short period of

observations. Hydrol. Process. 26, 883–892. https://doi.org/10.1002/hyp.8302.
Troy, T.J., Wood, E.F., Sheffield, J., 2008. An efficient calibration method for continental-scale land surface modeling. Water Resour. Res 44. https://doi.org/10.1029/2007WR006513.
Van Loon, A., Van Lanen, H., 2012. A process-based typology of hydrological drought. Hydrol. Earth Syst. Sci. 16, 1915–1946. https://doi.org/10.5194/hess-16-1915-2012.
Van Loon, A.F., Tijdeman, E., Wanders, N., Van Lanen, H.A.J., Teuling, A.J., Uijlenhoet, R., 2014. How climate seasonality modifies drought duration and deficit. J. Geophys. Res. Atmos. 119, 4640–4656. https://doi.org/10.1002/2013JD020383.
Vaze, J., Post, D., Chiew, F., Perraud, J., Viney, N., Teng, J., 2010. Climate non-stationarity – validity of calibrated rainfall-runoff models for use in climate change studies. J. Hydrol. 394, 447–457. https://doi.org/10.1016/j.jhydrol.2010.09.018.
Viviroli, D., Zappa, M., Gurtz, J., Weingartner, R., 2009. An introduction to the hydrological modelling system PREVAH and its pre- and post-processing-tools. Environ. Modell. Softw. 24, 1209–1222. https://doi.org/10.1016/j.envsoft.2009.04.001.
Vořechovský, M., 2015. Hierarchical refinement of latin hypercube samples. Comput.-Aided Civil Infrastruct. Eng. 30, 394–411. https://doi.org/10.1111/mice.12088.
Vrugt, J., Sadegh, M., 2013. Toward diagnostic model calibration and evaluation: approximate Bayesian computation. Water Resour. Res. 49, 4335–4345. https://doi.org/10.1002/wrcr.20354.
Wagener, T., Gupta, H., 2005. Model identification for hydrological forecasting under uncertainty. Stoch. Environ. Res. Risk Assess. 19, 378–387. https://doi.org/10.1007/s00477-005-0006-5.
Wagener, T., Sivapalan, M., Troch, P., McGlynn, B., Harman, C., Gupta, H., Kumar, P., Rao, P., Basu, N., Wilson, J., 2010. The future of hydrology: an evolving science for a changing world. Water Resour. Res. 46. https://doi.org/10.1029/2009WR008906.
Wilson, C., Valdes, J., Rodriguez-Iturbe, I., 1979. On the influence of the spatial distribution of rainfall on storm runoff. Water Resour. Res. 15, 321–328.
Wood, E., Roundy, J., Troy, T., van Beek, L., Bierkens, M.P., Blyth, E., de Roo, A., Döll, P., Ek, M., Famiglietti, J., Gochis, D., van de Giesen, N., Houser, P., Jaffé, P., Kollet, S., Lehner, B., Lettenmaier, D., Peters-Lidard, C., Sivapalan, M., Sheffield, J., Wade, A., Whitehead, P., 2011. Hyperresolution global land surface modeling: meeting a grand challenge for monitoring Earth's terrestrial water. Water Resour. Res. 47. https://doi.org/10.1029/2010WR010090.
Yang, J., Reichert, P., Abbaspour, K., 2007. Bayesian uncertainty analysis in distributed hydrologic modeling: a case study in the Thur river basin (Switzerland). Water Resour. Res. 43. https://doi.org/10.1029/2006WR005497.
Zappa, M., 2002. Multiple-Response Verification of a Distributed Hydrological Model at Different Spatial Scales. ETH Zürich. Chapter 4. The sensitivity of distributed hydrological simulations to the spatial resolution of physiographic data. 14895, pp. 35–51. doi:https://doi.org/10.3929/ethz-a-004529728.
Zappa, M., Kan, C., 2007. Extreme heat and runoff extremes in the Swiss Alps. Nat. Hazards Earth Syst. Sci. 7, 375–389. https://doi.org/10.5194/nhess-7-375-2007.
Zhao, F., Zhang, L., Chiew, F., Vaze, J., Cheng, L., 2013. The effect of spatial rainfall variability on water balance modelling for south-eastern Australian catchments. J. Hydrol. 493, 16–29. https://doi.org/10.1016/j.jhydrol.2013.04.028.