Combining smart model diagnostics and effective data collection for snow catchments

Dominik Reusser

Published online at the
Institutional Repository of the University of Potsdam:
URL http://opus.kobv.de/ubp/volltexte/2011/5257/
URN urn:nbn:de:kobv:517-opus-52574
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus-52574

A dissertation submitted to the Faculty of Mathematics and Natural Sciences at the
University of Potsdam, Germany
for the degree of Doctor of Natural Sciences (Dr. rer. nat.) in Hydrology

Submitted	06.10.2010
Defended	31.03.2011
Published	31.05.2011

Referees

Prof. Dr. Erwin Zehe	University of Potsdam, Institute of Geoecology
Prof. Dr. Bruno Merz	University of Potsdam, Institute of Geoecology
Prof. Dr. Hoshin Gupta	The University of Arizona, Department of Hydrology and Water Resources

Contents

Abstract

Complete protection against flood risks by structural measures is impossible. Therefore flood prediction is important for flood risk management. Good explanatory power of flood models requires a meaningful representation of bio-physical processes. Therefore great interest exists to improve the process representation. Progress in hydrological process understanding is achieved through a learning cycle including critical assessment of an existing model for a given catchment as a first step. The assessment will highlight deficiencies of the model, from which useful additional data requirements are derived, giving a guideline for new measurements. These new measurements may in turn lead to improved process concepts. The improved process concepts are finally summarized in an updated hydrological model.

In this thesis I demonstrate such a learning cycle, focusing on the advancement of model evaluation methods and more cost effective measurements. For a successful model evaluation, I propose that three questions should be answered: 1) when is a model reproducing observations in a satisfactory way? 2) If model results deviate, of what nature is the difference? And 3) what are most likely the relevant model components affecting these differences? To answer the first two questions, I developed a new method to assess the temporal dynamics of model performance (or TIGER - TIme series of Grouped Errors). This method is powerful in highlighting recurrent patterns of insufficient model behaviour for long simulation periods. I answered the third question with the analysis of the temporal dynamics of parameter sensitivity (TEDPAS). For calculating TEDPAS, an efficient method for sensitivity analysis is necessary. I used such an efficient method called Fourier Amplitude Sensitivity Test, which has a smart sampling scheme. Combining the two methods TIGER and TEDPAS provided a powerful tool for model assessment.

With WaSiM-ETH applied to the Weisseritz catchment as a case study, I found insufficient process descriptions for the snow dynamics and for the recession during dry periods in late summer and fall. Focusing on snow dynamics, reasons for poor model performance can either be a poor representation of snow processes in the model, or poor data on snow cover, or both.

To obtain an improved data set on snow cover, time series of snow height and temperatures were collected with a cost efficient method based on temperature measurements on multiple levels at each location. An algorithm was developed to simultaneously estimate snow height and cold content from these measurements. Both, snow height and cold content are relevant quantities for spring flood forecasting.

Spatial variability was observed at the local and the catchment scale with an adjusted sampling design. At the local scale, samples were collected on two perpendicular transects of 60 m length and analysed with geostatistical methods. The range determined from fitted theoretical variograms was within the range of the sampling design for 80% of the plots. No patterns were found, that would explain the random variability and spatial correlation at the local scale.

At the watershed scale, locations of the extensive field campaign were selected according to a stratified sample design to capture the combined effects of elevation, aspect and land use. The snow height is mainly affected by the plot elevation. The expected influence of aspect and land use was not observed.

To better understand the deficiencies of the snow module in WaSiM-ETH, the same approach, a simple degree day model was checked for its capability to reproduce the data. The degree day model was capable to explain the temporal variability for plots with a continuous snow pack over the entire snow season, if parameters were estimated for single plots. However, processes described in the simple model are not sufficient to represent multiple accumulation-melt-cycles, as observed for the lower catchment. Thus, the combined spatio-

temporal variability at the watershed scale is not captured by the model. Further tests on improved concepts for the representation of snow dynamics at the Weißeritz are required. From the data I suggest to include at least rain on snow and redistribution by wind as additional processes to better describe spatio-temporal variability. Alternatively an energy balance snow model could be tested.

Overall, the proposed learning cycle is a useful framework for targeted model improvement. The advanced model diagnostics is valuable to identify model deficiencies and to guide field measurements. The additional data collected throughout this work helps to get a deepened understanding of the processes in the Weisseritz catchment.

Kurzfassung

Für einen effektiven Hochwasserschutz sind reine Infrastrukturmaßnahmen häufig ungenügend und müssen durch komplexe Modelle zur Hochwasservorhersage und –warnung, zu einem umfassenden Schutzkonzept vervollständigt werden. Derartige Modelle basieren auf einer bio-physikalisch Repräsentation der relevanten hydrologischen Prozesse, weshalb eine Verbesserungen in der Beschreibung dieser Prozesse, zuverlässigere Vorhersagen ermöglichen kann. Dabei markiert die zunächst kritische Beurteilung von bereits existierenden Modellen den Beginn zu einem erweiterten Systemverständnis. Weiterhin führen aufgedeckte Schwachstellen im Model häufig zu einer erneuten Datenerhebung, wobei die bei der Modelbeurteilung gewonnenen Erkenntnisse als Orientierungshilfe dienen können. Das daraus resultierende, vertiefte Verständnis kann zu einer verbesserten Beschreibung der hydrologischen Prozesse genutzt werden, gefolgt von einer Überarbeitung des Modells, wodurch ein Lernzyklus abgeschlossen wird.

In dieser Arbeit wird ein solcher Lernzyklus aufgegriffen, wobei der Schwerpunkt auf einer verbesserten Modellanalyse und kosteneffizienteren Messungen liegt. Für eine erfolgreiche Modellbeurteilung sind drei Fragen zu beantworten: 1) Wann reproduziert ein Modell die beobachteten Werte in einer zufriedenstellenden Art und Weise (nicht)? 2) Wenn Unterschiede bestehen, wie lassen sich die Abweichungen genau charakterisieren? und 3) welches sind die Modellkomponenten, die diese Abweichungen bedingen? Um die ersten beiden Fragen zu beantworten, wird eine neue Methode, genannt TIGER (TIme series of Grouped Errors), zur Beurteilung des zeitlichen Verlaufs der Modellgüte vorgestellt. Eine wichtige Stärke dieser neuen Methode liegt darin, dass wiederholende Muster ungenügender Modellgüte auch für lange Simulationsläufe einfach identifiziert werden können. Die dritte Frage wird durch die Analyse des zeitlichen Verlaufs der Parametersensitivität beantwortet, welche eine effiziente Sensitivitätsanalyse-Methode bedingt. In dieser Arbeit wird eine solche effiziente Methode namens Fourier-Amplituden-Sensitivitäts-Test verwendet, die den Parameterraum sehr effizient durchsucht. Eine Kombination der beiden Methoden zur Beantwortung aller drei Fragen stellt ein umfangreiches Werkzeug für die Analyse hydrologischer Modelle zur Verfügung.

Als Fallstudie wurde WaSiM-ETH verwendet, um das Einzugsgebiet der Wilden Weißeritz zu modellieren. Die Modellanalyse von WaSiM-ETH hat ergeben, dass die Schneedynamik und die Rezession während trockener Perioden im Spätsommer und Herbst, für eine Beschreibung der Prozesse an der Weißeritz nicht geeignet sind. Der Unterschied zwischen Modell und Simulation kann entweder von einer ungenügenden Beschreibung der Prozesse oder von Fehlern in den vorhandenen Daten oder aus beiden Quellen stammen. Der nächste Schritt im Lernzyklus beinhaltet die Erhebung zusätzlicher Daten, was am Beispiel der Schneedynamik aufgezeigt wird.

Detaillierte Daten über Schneetemperaturen und Schneehöhen wurden mit Hilfe eines neuen, preisgünstigen Verfahrens erhoben. Dazu wurde die Temperatur an jedem Standort mit unterschiedlichen Abständen zum Boden gemessen. Schließlich wurde ein Algorithmus entwickelt, der aus den Temperaturmessungen sowohl die Schneehöhe, als auch den Kältegehalt der Schneedecke berechnet. Die Schneehöhe und Kältegehalt sind wichtige Größen für die Vorhersage von Frühjahrshochwassern.

Die räumliche Variabilität der Schneedecke wurde mit einem zweistufigen Beprobungsplan sowohl kleinräumig, als auch auf der Einzugsgebietsskala erfasst. Auf der kleinräumigen Skala wurden Schneehöhen und -dichten auf zwei 60 m langen, rechtwinkligen Transekten gemessen und die Daten geostatistisch ausgewertet. Theoretische Variogramme wurden an die Daten angepasst, um die Korrelationslänge zu berechnen. Es stellte sich heraus, dass für 80% der beprobten Flächen die Korrelationslänge innerhalb der 60 m der Transek-

te lagen. Es konnten keine Faktoren gefunden werden, um Unterschiede in der Korrelationslänge, der Semivarianz und der Anisotropie zu erklären.

Auf der Einzugsgebietsskala wurden die Flächen entsprechend der Landnutzung, der Höhenzone und der Ausrichtung stratifiziert ausgewählt, um den Einfluss dieser drei Faktoren zu untersuchen, wobei lediglich der Einfluss der Höhe nachgewiesen werden konnte, während Ausrichtung und Landnutzung keinen statistisch signifikanten Einfluss hatten.

Um die Defizite des WaSiM-ETH Schneemodules für die Beschreibung der Prozesse im Weißeritzeinzugsgebiets besser zu verstehen, wurde der gleiche konzeptionelle Ansatz als eigenständiges, kleines Modell benutzt, um die Dynamik in den Schneedaten zu reproduzieren. Während dieses Grad-Tag-Modell in der Lage war, den zeitlichen Verlauf für Flächen mit einer kontinuierlichen Schneedecke zu reproduzieren, konnte die Dynamik für Flächen mit mehreren Akkumulations- und Schmelzzyklen im unteren Einzugsgebiet vom Modell nicht abgebildet werden. Folglich können die raum-zeitlichen Schneemuster im Einzugsgebiet von diesem Modell nicht umfassend beschreiben werden. Eine Erweiterung des Modellkonzeptes für die Beschreibung der Schneedynamik an der Weißeritz ist deshalb erforderlich. Dabei scheint die Einbeziehung des Windtransportes und des Regeneinflusses auf die Schneedecke, in das Modell, notwendig zu sein. Alternativ könnte auch ein komplettes Energiebilanzmodel getestet werden.

Zusammenfassend hat sich das Lernzyklus-Konzept als nützlich erwiesen, um gezielt an einer Modellverbesserung zu arbeiten. Die differenzierte Modelldiagnose ist wertvoll, um Defizite im Modellkonzept zu identifizieren und die Planung von zusätzlichen Messungen zu unterstützen. Die während dieser Studie erhobenen Daten sind geeignet, um ein verbessertes Verständnis der Schnee-Prozesse an der Weißeritz zu erlangen.

Chapter 1

Introduction

1.1 Flood prediction as iterative learning process

Floods have a great potential for damage, effecting large efforts for flood risk management. Flood risk mitigation has two main approaches, structural and non-structural measures (Merz, 2006). Structural measures attempt to reduce the probability and impact of floodings by building dams, dikes and polders. However, even if the impossible was possible, it would be too cost intensive to provide complete protection based on structural measures (Merz, 2006). In addition, under climate and societal change, magnitudes, frequencies and impacts of floods are expected to change over time. Non-structural measures are more flexible and include flood risk oriented land use planing, establishment of warning schemas and emergency plans. Efficient planning and warning requires reliable flood predictions, which are generally based on hydrological models. The simulation of rare events will most likely require extrapolation from the historically observed catchment behavior. Also, the quality of measurements becomes questionable for extreme events. While extrapolation is critical from a scientific point of view, it is an intrinsic problem of flood prediction.

Hydrological models can be roughly separated into parsimonious, data driven models and complex, physically based models. Advantages and disadvantages of both is an ongoing discussion (e.g. Todini, 2007; Kirchner, 2006). A fundamental assumption in this thesis is, that confidence with respect to extrapolation beyond observed conditions may increase, if bio-physical functioning of a catchment is well represented in the model. With this approach, the model is considered to be the best available conceptualisation of the catchment functioning. Naturally, every model is necessarily a simplification of reality and cannot reproduce all the functions. To obtain a satisfying representation, which includes the relevant processes, a repeated adjustment to the model for the catchment under investigation is necessary. For the adjustment, the catchment dynamics are modelled and compared to observations.

Learning from the difference between a model and the observation requires the application of some diagnostic tools – in the simplest case plotting the two time series together. Ultimately, the diagnostic tool will help to determine which model components are insufficient to reproduce the catchment behaviour. However, diagnostics are complicated because differences can be caused by errors and gaps in the measurements or the model structure. In this thesis, model structure is defined to consists of the process conceptualisations (equations) for a spatial element and some scale reduction in space and time by describing the exchange between the spatial elements. In addition, we need estimates for the required parameter values.

To check for errors in the measurements and close existing information gaps, the differences between model and observation can be used to derive

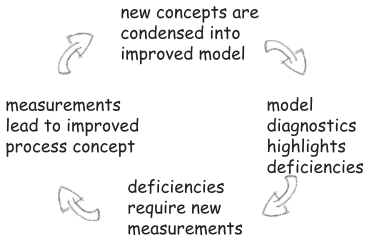


Figure 1.1: Model based learning cycle

a set of measurements to further improve our understanding. Since measurements always require a balance between amount of information gained and related costs, cost efficient measurements are preferred. The newly achieved insight from new data and new process descriptions eventually leads to improvements to the model.

Finally, this leads to a learning cycle as shown in Figure 1.1. In the Figure, the process described above is split into four consecutive steps. From a scientific point of view, learning is one of the main purposes of any kind of modelling, including data driven approaches.

The overall rational behind this thesis is to exemplify such an iterative learning cycle for the case of flood predictions:

a) use model diagnostics to provide information about missing understanding, b) obtain better data with affordable technology, c) find a parsimonious model to describe the new data, d) improve the model based on the new process understanding.

I will put the main focus on two steps of the learning cycle. The first core topic are diagnostic tools. Diagnostic tools are most often applied to entire simulation periods providing information about the average performance, since objective functions are generally information aggregation functions. However, depending on the conditions, different processes of the hydrological cycle are relevant and much information could be gained when assessing performance seperately for different processes. We refer to the conditions as the hydrological context. For example, the hydrology of the catchment may be a) dominated by either mass input or energy input. b) Thresholds may alter the functioning of the catchment; for example snow influenced periods occur when temperatures drop below snow melt temperature. The catchment may spend most of its time c) either close to or far from equilibria. The processes during various contexts are generally conceptualised in different model components. Therefore, examination of average performance of a model is only a first order assessment. Improving hydrological models in a more targeted way requires time dependent performance measures as different processes dominate in different contexts. Based on time dependent performance measures we may assess when model components fail. Methodologies for such a context or temporally resolved analysis require further development. The topic is further introduced in section 1.2 and a method for diagnostics is presented in chapter 2, chapter 3 and chapter 4.

The second core topic is cost effective snow measurements. From the model diagnostic, we learn that the model WaSiM-ETH, which we use as conceptualisation of the Weisseritz catchment, consistently produces overestimates during the snow melt periods. Therefore, we conclude that the snow module using a temperature index approach is not sufficient to describe dynamics in the Weisseritz catchment. In particular a single degree day factor appears not to be a sufficient description of the processes. Thus we need additional data

about the snow cover, since the two existing meteorological stations are not sufficient to capture the influence of topography and aspect in the catchment. In order to meet budget constraints of the research project, we focuse on cost efficient observations. The new methods may also be attractive for measurement networks in regions with constraint research budgets*(van de Giesen et al., 2009a,b). More details about snow hydrology and our measurement approach are introduced in section 1.3 and are presented in chapter 5 and chapter 6.

In section 1.4 the Weisseritz as the study area is introduced and some encompassing issues concerning knowledge management in science is discussed in section 1.5. Throughout the introduction, a number of guiding questions are formulated, that are summarized in section 1.6 and answered throughout this thesis.

1.2 Model diagnostic

Development of temporally dynamic or context dependent model diagnostics is the first focus and constitutes the first step of the learning cycle. To start I will give a more precise description of what I mean with model diagnostic. The development of a model is alway (explicitly or implicitly) driven by a purpose, i.e. some characteristics to be reproduced or a number of questions to be answered using this model. I define model diagnostic as the test whether a model is able to reproduce the characteristics as required by the model purpose. For rainfall-runoff hydrology, the model purpose is to reproduce amounts and timing of discharge. It also includes separating different mechanisms of runoff generation. Thus, a hydrological model is expected to reproduce different processes in nature: fast formation of runoff due to saturation or infiltration excess during rainfall events, interactions of the unsaturated zone with groundwater, formation of discharge from groundwater and interflow during dry periods, withdrawal of water by evaporation and transpiration and storage and runoff due to snow and ice processes.

Model diagnostics should be developed in a way to also work for complex models because, with increasing process understanding and spatial resolution, models tend to increase in complexity. On the one hand, this means that in general we have more parameters. On the other hand, new measurement techniques make available more observations against which to test the model. In addition more a priory knowledge about catchment properties (soil patterns, layering etc) can be used to reduce the degrees of freedom of our parameter estimation process.

Independent of model complexity, integrative objective functions are most commonly used, such as mean squared error or the Nash Sutcliffe coefficient of efficiency (Nash and Sutcliffe, 1970) between observed and modelled discharge. The advantage of such objective functions is, that they summarize a lot of information and can be perceived much faster than long time series. Also, they are often used for automatic calibration procedures, which fostered further development and understanding of objective functions (Gupta et al., 1998, 2009, e.g.). But there may be more informative ways to assess a model. This leads to the first guiding question: How to assess (poor) model performance?

As outlined above, depending on the contexts (rainfall driven, energy driven, snow dominated) different processes are driving the hydrological response. This is the fundamental reason for the fact that global performance measures are not sufficient. Different model components are developed to mimic catchment functions in different contexts and we need to better understand what the errors are depending on context and sensitive parameters. Temporally resolved model diagnostic, which is implicitly used during visual inspection, is able to enhance the understanding about context dependent performance. Temporally resolved model diagnostics did not undergo the same formaliza-

*Nick van de Giesen: Trans-African Hydro-Meteorological Observatory; http://www.tahmo.org/

tion process as objective functions for calibration purposes. Enhancing objectivity of temporally resolved model diagnostics is the goal of the second guiding question: How is it possible to identify temporal patterns and context dependence in model performance?

For diagnostics of a hydrological model including context dependence, I propose to assess 1) when a model is performing acceptably/poorly, 2) of what the deviations are and 3) whether the relevant process conceptualisations are active. This is visualized in figure 1.2. The hydrological model depicted in the center consists of three components that represent different processes. The model output (bottom) is tested for deviations from the observation with respect to the first two questions. To answer the third question, the right model components (center) need to be active.

Before giving a short overview of the fundamental principles of well established approaches, I would like to highlight that recent work also targets at more informative diagnostics. With their diagnostic approach to model evaluation, Gupta et al. (2008) suggest to use a multi-objective approach during which it is checked whether relevant signatures are reproduced by the model. They argue that with objective functions as currently used in hydrology we are often aggregating too much, thereby loosing important pieces of information. Much in the same direction is the outcome from a workshop of the PUB initiative titled 'Uncertainty Analysis and Model Diagnostics' (Wagener et al., 2006). For example, (Liu and Others, 2010) suggest to use temporal variation of optimal parameter sets (dynamic parameter analysis) for model diagnostic. Identification of information rich periods in a data set based on information theory is presented by (Jackson et al., 2010).

A glossary related to model diagnostics together with a list of available software tools is provided by Matott et al. (2009). Note that their broader definition of model diagnostics also includes data analysis, parameter estimation, multi-model analysis and Bayesian networks. However, these meth-

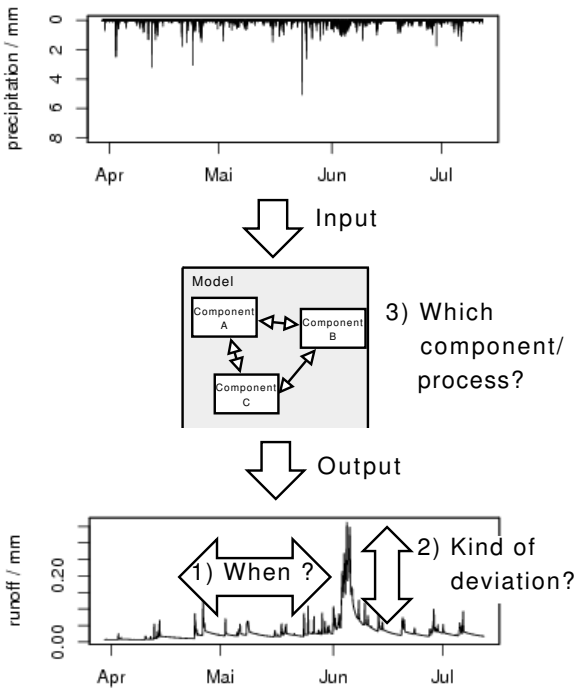


Figure 1.2: Sketch of a model showing three questions to be answered from model diagnostics

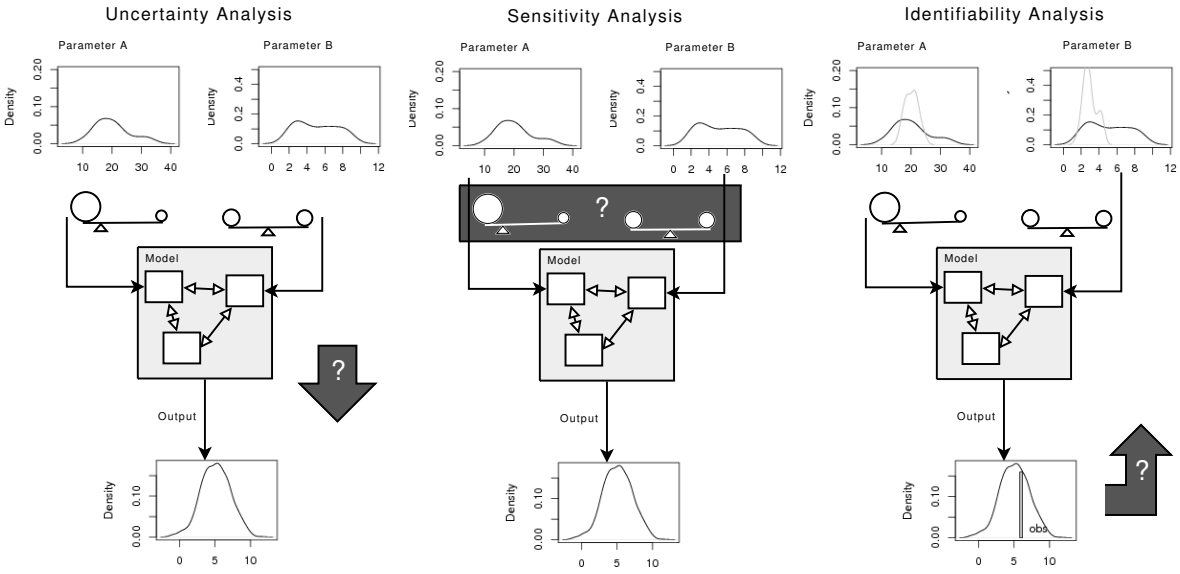


Figure 1.3: Relationship between uncertainty, sensitivity and identifiability analysis. The model in the center has parameters (and other inputs) about which some uncertainty (density functions) exists. Depending on the sensitivity of the model (balances) on these parameters, the uncertainty is propagated through the model and makes the model output uncertain.

ods are not methods for model diagnostic, as it was defined above, i.e to test whether a model can fulfill its purpose.

Existing approaches for diagnostics are generally based on one of three related concepts (Figure 1.3): Uncertainty, sensitivity and identifiability. Each one may serve as diagnostics for either time series of model outputs or for an aggregating objective function.

The differences between the concepts should become clear from the following, brief introduction together with figure 1.3. The figure shows three repetitions of the same sketch from left to right. In the center of each sketch, a model with different components is shown. The model has some parameters (top) and the resulting model output (bottom) shows different sensitivities towards these parameters (depicted by the balance). In the example, the model is relatively insensitive for parameter A, i.e. large changes of the parameter cause only small changes in the output. The focus of the various methods is indicated by the dark shading. Beside a short description of the method, I will also briefly highlight how results are generally presented with respect to temporal resolution.

Uncertainty analysis (Figure 1.3, left) determines the amount of uncertainty in the output of a model (often as cumulative density function of the output variable) that is caused by various sources of uncertainty (Montanari et al., 2009). Sources of uncertainty are classically input variables, model structure, model parameters, and data. Uncertainty analysis is a very active field of research, thus it is not surprising that different ways exist to assign (informal) measures of likelihood for the quantification of uncertainty (e.g. Montanari et al., 2009; Beven and Freer, 2001). Also, no agreement about the exact meaning of uncertainty analysis exists in the literature (Montanari, 2007). The example presented in figure 1.3 shows parameter uncertainty of the model, other sources of uncertainty can be included (e.g. Kavetski et al., 2006a; Clark et al., 2008). Splitting uncertainty according to sources is an important goal in hydrology and two approaches are presented by Beven et al. (2010). This always requires additional information about the relationship between the model and the modelled system, e.g. some error model. Despite all the variety in uncertainty analysis, it is very common to represent uncertainties evolving with time - often as a confidence-like band around the modelled discharge.

Sensitivity analysis (SA – Figure 1.3, middle) searches for the parameters affecting the model output the most. A model output is said to be sensitive for a certain parameter if a small change in the parameter value causes a large change in the model output. SA is related to the concept of uncertainty because sensitive parameters cause large uncertainty in the model output if the parameter itself is uncertain. It is quite common that SA is performed for objective functions, aggregating over time and I found only two studies that consider temporal variability of sensitivity (Sieber and Uhlenbrook, 2005; Cloke et al., 2008). SA is much related to the third question as depicted in figure 1.2 and throughout this thesis I would like to further investigate the topic answering the guiding question: Can we identify relevant model components (for computationally expensive models)?

The goal of identifiability analysis (Figure 1.3, right) is to determine how far a set of unobservable parameters may be constrained by minimizing the difference between a given set of observations and the corresponding model output. If parameters are well identifiable, the remaining uncertainty in the model output from these parameters will naturally be small. Parameter sensitivity is a necessary but not sufficient condition for identifiability. Temporal analysis of model identifiability was established by Wagener et al. (2003) with the dynamic identifiability analysis. In a similar way, Choi and Beven (2007) showed with their model conditioning procedure that performance measures calculated on a seasonal scale give some additional indication about parameter identifiability and model structure deficiencies when compared to global performance measures. Similarly, Shamir et al. (2005) were

able to improve identifiability of model parameters when looking at model performance on different time scales.

In this thesis, I develop and present a complementary model diagnostic approach targeted at context dependent model diagnostics of complex models. As stated before, successful model diagnostics should answer the three questions (Figure 1.2): 1) during which periods the model is or is not reproducing observed quantities and dynamics; 2) What is the nature of the error in times of poor model performance, and 3) which components of the model are causing this error.

An approach called TIGER (Time series of grouped errors) answering the the first two questions is presented in chapter 2. Temporal dynamics of parameter sensitivity (TEDPAS) is useful to answer the third question and is presented in chapter 3, while the combination to a full diagnostic tool is demonstrated in chapter 4. The procedure was developed using WaSiM-ETH, Catflow and LARSIM as possible representations for the Weisseritz catchment. Results will be mainly presented for WaSiM-ETH throughout this thesis. With the diagnostic tool, I will answer the third guiding question: What are the limitations of WaSiM-ETH as representation of the Weißeritz catchment? The answer to the guiding questions will then lead to the next step in the learning cycle, which is to identify additional measurements from the deficiencies.

1.3 Snow hydrology

The deficiencies identified with the model diagnostics guide the collection of additional data. The model diagnostics revealed discharge to be consistently too high during snow melt events for a large parameter range (Chapter 4). I used this deficiency as an example for the next step of the learning process. Additional, distributed observations on snow state and height are required for four reasons: a) Spatial variability is not sufficiently captured by existing stations. Variation of land use and topography causes inhomogeneous snow processes throughout the catchment. b) Deficiencies may also be caused by missing processes such as interception, evaporation and sublimation. c) Data is necessary to test effects of a simple model improvement, which is based on distributed snow melt parameters compared to a single catchment wide parameter. d) A distributed multi-response test of the model may be performed with the distributed data. The simplest way for such a test is to correlate the ranks of the distributed observations with the ranks of the simulated snow cover at these locations. A high correlation would indicate that the major processes causing spatial variability are represented in the model, while a low correlation would indicate that there are still missing processes.

The remaining chapter will 1) give a short introduction of relevant processes for snow hydrology 2) look at spatio-temporal variability and how it can be measured 3) modelling approaches for spatio-temporal snow processes 4) summarize how snow campaigns will provide data for the development of better process concepts.

1.3.1 Relevant processes for snow patterns

Snow dynamics is complex and includes many influencing factors. Three major phases can be distinguished: accumulation, metamorphosis and ablation (Dingman, 2002). Figure 1.4 shows a conceptualisation of the three phases, depicting the main processes, that are active during each phase.

Accumulation phase includes snow fall and redistribution during snow fall events. The amount of snow is strongly affected by the amount of water in the atmosphere available for precipitation. Compared to rain, measurement of snow (and other solid precipitation) has larger errors due to wind. Depending on exposition, the error may be up to 34% (Richter, 1995).

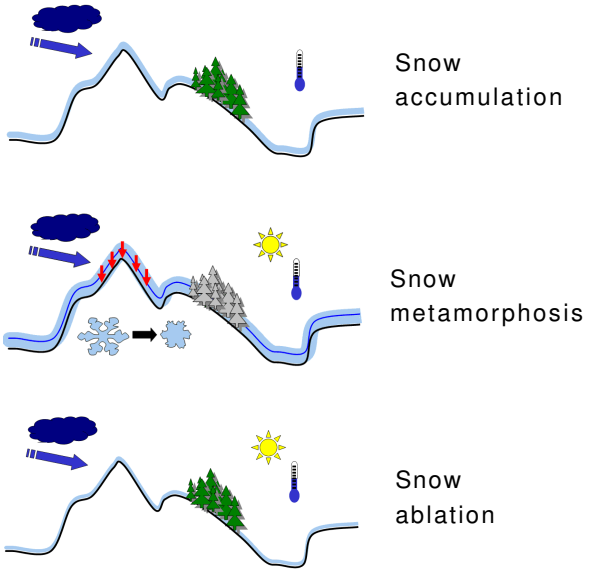


Figure 1.4: Snow processes. Depending on the phase (accumulation, metamorphosis and ablation), different main factors exist: precipitation, wind, land use, topography, temperature, radiation and snow metamorphosis, (copied with permission from Eckart, 2008)

Snow fall is also strongly affected by the thermal gradient of temperature, which causes higher snow covers at higher elevations. The transition between rain and snow occurs between 4 and -2 ° C (Braun, 1985) and is dependent on various factors related to energy availability in the atmosphere during the snow fall process (Bourgouin, 2000; Stewart, 1985). Redistribution by wind causes higher accumulation in wind protected (depressions, edge of the wood), compared to exposed areas (MacDonald et al., 2009).

Forest and aspect are reported to have a similar influence in magnitude, which is relatively small compared to the effect of elevation (Jost et al., 2007, 2009). The effect of forest is complicated and has many influencing factors, such as forest type (Winkler et al., 2005) and canopy density measured by sky view factor (López-Moreno and Latron, 2008). Interception, affecting snow amount and distribution, is highly dependent on meteorological conditions such as wind speed and snowfall amount (Hedstrom and Pomeroy, 1998) and conditions affecting radiation (clear and overcast days) (Hardy et al., 2004). A model for accumulation and ablation in forested environments has recently been presented by Andreadis et al. (2009) based on extensive snow lysimeter data (Storck et al., 2002). A factor affecting heat exchange by wind (snow roughness length) and the maximum interception capacity are reported as highly influential parameters (Andreadis et al., 2009). In general redistribution in the forest are affected by lower wind speeds and higher spatial variability of the wind.

The main factors during the accumulation phase are thus precipitation, temperature and wind and effects of topography on wind (Figure 1.4).

Metamorphosis phase is also affected by wind transport (Deems et al., 2006; Winstral et al., 2002).

Existing gradients from the scale of crystals to the entire snow cover cause restructuring of the

snow cover. Very locally, thermodynamic gradients cause water to be transported as vapor from convex surfaces with small radii to less convex surfaces in the crystal, causing snow flakes to convert to round grains (destructive metamorphism). With increasing abilities for visualization, understanding of these processes gets better (Schneebeli and Sokratov, 2004). Along with the metamorphosis, physical properties of snow are affected, for example increasing density and thermal conductivity (Sturm et al., 1997; Schneebeli and Sokratov, 2004)

Thermal conductivity is of interest because it strongly affects the thermodynamic gradients within the snow cover and to the atmosphere and thus determines to a great extend the speed of heat exchange processes. Measurements of thermal conductivity are described and discussed in the literature (Sturm et al., 1997; Brandt and Warren, 1997; Satyawali and Singh, 2008; Singh, 1999; Aggarwal et al., 2009; Fukusako, 1990)

Constructive metamorphism includes both, sintering and formation of depth hoar. Sintering describes the formation of connections between touching snow grains by deposition of water. If large thermal gradients exist across the snow cover, water gets transported from warmer to colder regions building grains with facets that can be cup-shaped and that are up to 10 mm in diameter, called depth hoar. Depth hoares adhere loosely to each other, greatly increasing the risk for avalanches. Thus elaborate models for snow metamorphosis have been developed (e.g Bartelt and Lehning, 2002; Lehning et al., 2002).

Recurrent melt and freezing cycles (melt metamorphism) that often occur towards the end of the snow season strongly affect the structure of the snow cover. Pores are filled by melting snow and subsequent refreezing results in larger aggregates of ice. A similar effect may occur during rain on snow events. Repeated melt and freezing cycles will eventually lead to firn.

Pressure by wind and the mass of the snow pack itself compact the snow cover, resulting in higher densities. The main factors during the metamorphosis phase are thus energy sources (temperature, radiation) causing thermal gradients, as well as wind, which strongly affects heat exchange between the atmosphere and the snow cover (Figure 1.4).

Snow ablation consists of removal of water to the atmosphere by sublimation and evaporation, as well as snow melt. For both, energy input to the snow cover is of importance. Important factors are exposition to direct sun light, snow albedo, temperature and the thermal conductivity of the snow cover.

Snow age is very important for energy input since albedo is much lower for aged snow, causing higher energy input. However, comparison of commonly used albedo models with albedo from remote sensing data show that snow age alone is not sufficient to estimate albedo (Molotch and Bales, 2006). For energy balances, distinction into short wave (direct sun light) and long wave radiation (diffuse radiation) is important since the two radiation types behave differently with respect to the atmosphere, the snow surface and vegetation. While short wave radiation is lower in forests, the long wave radiation may be higher, especially at night if the snow cover is shielded from the clear sky (which acts as heat sink) (Stähli et al., 2009). Melt rates are reported to increase close to vegetation, because of the increased energy input by radiation (Pomeroy et al., 2004; Liston, 1999).

While sublimation and evaporation require higher energy inputs to remove the same amount of SWE compared to melt, especially sublimation occurs also at low temperatures. Sublimation in connection with wind transport is a very important process and has been reported to be responsible for removal of up to 60%, depending on the location (MacDonald et al., 2009, 2010). Lower snow cover in forests compared to fields are mainly the result of sublimation and evaporation from intercepted snow (Pomeroy et al., 1998).

No snow melt occurs as long as the snow cover is below freezing temperatures. The term cold content is used to describe the energy input required to heat the snow cover to freezing temperature. Consequently, snow melt starts when the cold content is zero. For above freezing temperatures, heat transport is mainly caused by sensible and latent heat flux. The sensible heat flux is caused by temperature differences between air and the surface. Due to the low heat capacity of air, large masses of air are necessary for a considerable energy input. Latent heat flux occurs due to the energy released by condensing humidity. Because of the large condensation energy, large energy amounts are released.

Spatial variability of snow melt can be related to SWE and depends on the spatial scale (Pomeroy et al., 2004; Faria et al., 2000). On the small scale, melt rates are high at locations with little snow as the probability for exposure of vegetation and thus additional absorption of radiation is higher (negative correlation). Similarly, according to DeBeer and Pomeroy (2010) the cold content of thick snow packs is larger, thus the onset of the melt is strongly delayed compared to thin snow covers. At the catchment scale, high melt rates are reported for locations with high SWE (Pomeroy et al., 2004). Jost et al. (2007) report strong influence of elevation and exposition and lower melt rates in forests compared to clear cuts early in the melt season.

Melting water gets stored in the snow cover because of the free pore space. The retention capacity describes the volumetric fraction that the snow cover can store as melted water. It is generally low on the order of 3% (Dingman, 2002). Storage and refreezing of melt water as well as lateral transport of the melting water cause time delays between energy input into the snow cover and observation of melted water in rivers. Especially critical and sudden releases of the stored water occurs if rain enters a (nearly) saturated snow cover.

The main factors during the ablation phase can be summarized to be temperature, radiation, wind and land use (Figure 1.4).

1.3.2 Measurement of spatio-temporal variability

In some cases, missing processes may be identified directly from differences between observations and model results. However, analysis and understanding of spatial variability may be helpful to find missing processes, since factors affecting various processes are heterogeneous in land scape. Therefore possibilities to measure spatio-temporal variability are presented. To simultaneously observe variability of snow with high resolution in space and time is very challenging. Thus, observations are generally highly resolved either in time or in space.

Spatial variability: For snow, the same fundamental questions relevant for all spatial data are of importance. Most important, the relevant scale that fits the available data and the required model purpose needs to be determined (Clark et al., 2008). This in turn leads to a distinction between processes that have to be represented spatially explicit and processes that have to be represented spatially implicitly (Clark et al., 2008). Large biases in variance and correlation lengths occur if process scale, measurement scale and model scale do not agree (Blöschl, 1999). For snow, spatially resolved information are generally obtained from labor intensive measurement campaigns or from remote sensing techniques.

Several studies report large data sets from measurement campaigns that relate spatial variability of SWE with possible explanatory variables, mainly topographic characteristics (Anderton et al., 2004; Erickson et al., 2005; Elder et al., 1991). Pomeroy et al. (2004) analyse a multi-season data set of snow surveys for a 200 km^2 catchment. They find a log-normal distribution of SWE within single landscape classes. They discuss statistical relation between SWE and snow melt but do not find a simple dependence structure. Jost et al. (2007) were also unable to understand the small scale variability (meters), while on

the catchment scale (20 km^2), elevation, exposition and land use explained 80 to 90% of the variability after averaging out the small scale variability.

Remote sensing techniques bring the advantage that higher resolved data sets covering larger areas can be produced. However, reference measurements for calibration and verification are necessary. Direct measurement of snow height (and SWE) are recently available with new measurement techniques, not requiring the indirect way via snow covered area (SCA). Deems et al. (2006, 2008) use air borne lidar measurements of snow height. The high density spatial data were used to determine the fractal distributions of snow height. Remote sensing SWE estimates are available from radar observations (Schaffhauser et al., 2008), however the method is still under development. For example the influence of snow density on such radar based remote sensing estimates of SWE was reported (Lundberg et al., 2006).

The classical approach to estimation of SWE from remote sensing data is to use relations between SCA and SWE, since SCA can easily be determined from remote sensing data. The required assumptions and the theoretical derivation of the relationship between SWE and SCA is presented by Liston (1999). Farinotti et al. (2010) recently reported determination of SCA from photography, which allows a higher temporal resolution compared to satellite based observations. Also, observations are possible on overcast days. Assimilation of SCA into hydrological model with a Kalman filter was reported for example by Clark et al. (2006). Bayesian approaches are used to update snow depletion curves from SCA (Kolberg and Gottschalk, 2006; Kolberg et al., 2006) and to estimate storm-specific snowfall distributions (Durand et al., 2008). Approaches are available to correct SCA classification to avoid false classification of shaded and densely forested areas as snow free (Corbari et al., 2009).

Temporally resolved data with high temporal resolution is available from automatic measurement stations. Measurement methods for the estimation of SWE were compared in the Swiss alps (Egli et al., 2009) and in Scandinavia (Lundberg et al., 2010) and conclude, that the optimal technique depends on the required information. Automatic measurement stations are cost intensive in installation and maintenance. Therefore, data from such stations is generally available with a very low spatial resolution. Thus, more cost effective measurements are attractive for research and operational purposes. Lundquist and Lott (2008) presented a technique using a combination of inexpensive temperature sensors and a snow model to reconstruct spatial distribution of SWE at maximum accumulation. Also in search of more efficient measurement methods, Jonas et al. (2009) report relationships between snow height, time of year, elevation, region and snow density. This allows to estimate SWE from simple snow height measurements for the alps. In this thesis I will present a method also based on inexpensive temperature sensors to make estimations of snow height and cold content (Chapter 5). While the approach requires a higher number of sensors compared to the method by Lundquist and Lott (2008), it works without a snow melt model and the related meteorological data. Overall cost effective alternatives allow for higher spatial density of data for the same costs.

1.3.3 Snow modelling

Short overviews about snow modelling are given in Ferguson (1999); Herpertz (2001). Several approaches exist that include differing fractions of the processes described in section 1.3.1. Models with increasing complexity can be broadly distinguished into temperature index based models, models solving the energy balance, and detailed multi layer snow models for avalanche predictions. As more processes are included, also data requirements become higher.

For a full characterization of commonly used snow modelling approaches see (Dingman, 2002; Ferguson, 1999; Herpertz, 2001). An overview and recent advances with respect to the temperature index approach are presented by Hock (2003). The temperature index approach is still widely used, for example for the snow map service of Norway, which was evaluated by Dyrrdal (2009). A stepwise model refinement of a distributed temperature index model is described in (Dunn and Colohan, 1999). Also, the case study model of this thesis, WaSiM-ETH, uses temperature index approach.

If radiation and wind data are included, a complete description of the energy balance is possible (Dingman, 2002). Jost et al. (2009) use a two-layer energy- and mass-balance model to model the snow processes in a forested catchment and further refinements to the albedo decay function and the canopy transmittance.

Essery et al. (2005) suggest that calculating separate energy balances for snow patches and snow free parts gives much better results compared to calculation of a mixed energy balance for the two sub-grid landscape fractions. Extending this approach of separating energy balances, for subgrid landscape fractions, so called snowmelt runoff contribution areas are introduced by DeBeer and Pomeroy (2010). These contributing areas vary substantially over time because the cold content is reduced much faster for shallow snow packs compared to deep ones. Also, aspect enhances variability of melt onset, resulting in complex patterns. A hydrological response unit based model for blowing snow is presented by MacDonald et al. (2010). They report sublimation of blowing wind to be an important factor for reduction of SWE for their catchment in the Canadian Rocky Mountains. Snow models for avalanche predictions are too complex and data intensive for hydrological modelling (e.g. Bartelt and Lehning, 2002; Lehning et al., 2002). For testing new concepts and approaches, high quality, longterm data sets for a snow laboratory in Colorado (Williams et al., 1999) and from the cold land processes experiment (Elder et al., 2009) may be useful.

1.3.4 Measurement strategy

After this brief review of snow hydrology related literature, I will get back to the learning cycle. In order to assess whether we need to include additional processes or if variability may be presented from additional data sources introducing spatial variability, we need good measurement of the temporal development and the spatial variability. This poses the following guiding question: What are temporal and spatial structures of the snow in the catchment?

Existing methods are good in either providing spacial information at one time or temporal information at one location. Thus I will split the measurement concept accordingly. Spatial variability of snow cover is efficiently observed using remote sensing techniques. Determination of SWE from satellite based radar measurements was part of a different working package within the OPAQUE research project, to which my thesis is associated. Unfortunately, results were not available at the time of writing of this thesis. As reference measurements for the remote sensing measurements and to better understand spatial distribution of snow we measured spatial variability with a sampling design similar to (Jost et al., 2007). The design is presented in chapter 6 and allows to separate spatial variability into small scale and catchment scale fractions of variability. Measurement campaigns were carried out during two winters from 2008-2010. Measurements were taken at about 20 locations throughout the catchment in order to determine influence of topographic factors and land use on spatial variability (Chapter 6).

Repetition of spatial patterns between years is an ongoing discussion, some authors find repeating patterns (Erickson et al., 2005; Deems et al., 2008), while others did only find weak to moderate similarity between years (Wilks, 2006). Repetition of spatial snow patterns between years will have to be tested, for example using remote sensing data.

However, this is not part of this thesis.

Temporal development of the snow cover requires expensive and labor intensive measurement stations, as highlighted above. Thus I developed a more cost efficient, robust methods in order to meet the budget constraints of the research project. Temperature measurements in and above the snow cover with inexpensive temperature sensors was used to obtain temporally resolved snow height estimates (Chapter 5). Evaluation of this dataset will answer the guiding question: How much information can be obtain from measurements with inexpensive temperature sensors? One snow reference station was installed at the research station of TU Dresden and TU Freiberg including a snow pillow, snow height measurements and a surface temperature sensor (Chapter 5). Combining the two measurement strategies has the potential to bring temporal development into space or spacial distribution into time. Understanding and modelling the patterns in the snow data results in the final guiding question: What processes are required to describe the new measured data and what are the resulting updates to the model?

1.4 Research area

The Weisseritz is a fast reacting catchment with high relief energy. Thus, high flow velocities occur during flood events. A recent extreme event in 2002 caused severe flooding. In the upper catchment, the high flows swept away roads and houses while close to the confluence, the river left the channel at a redirection from the historical river bed and flooded the town of Dresden, including the railway station.

The Wilde and Rote Weisseritz join in the town Freiberg. Both catchments are about 150 km^2. Three reservoirs are used for flood protection and drinking water supply, two in the catchment of the Wilde Weisseritz (Lehnmühle 22 Mio m^3 and Klingenberg 16 Mio m^3) and one as part of the Rote Weisseritz (Malter 9 Mio m^3). An additional reservoir is planned in the catchment of the Rote Weisseritz at the Pöbelbach (1 Mio m^3). Good understanding of the relevant processes is important for reservoir control.

The topography with narrow, elongated catchments provides a special challenge for meteorological forecasts, because the location of rainfall events is a very important factor. Heavy rainfalls occurring only few kilometers from the predicted location will cause a flooding in a different catchment. Slopes are gentle with an average of 7°, 99% of the slopes are $< 20°$; calculated from a 90 m digital elevation model (SRTM, 2002).

Soils are mostly cambisols, discharge is built to a great extend by interflow. Land use is dominated by forests (≈30%) and agriculture (≈50%). Some villages and towns exist (≈15%), including (from the upper catchment downwards) Altenberg, Dippoldiswalde, Freital and Tharandt. The river discharges into the Elbe close to Dresden. In the top catchment wetlands are apparent, causing some retention of water.

Mean temperatures are 11°C and 1°C for the periods April - September and October - March, respectively. Precipitation in Zinnwald is on the order of 1100 mm/year. Some of the precipitation occurs as snow fall, forming a snow cover of up to about 1 m for 1 to 4 months. Fog is important during winter in the valleys, reducing energy input by radiation and increasing precipitation.

Besides the measurements presented in this thesis, within OPAQUE, additional catchment characteristics were observed. 10 additional rain gauges were installed to validate research with respect to radar rainfall measurements. Multi-scale soil moisture observations from the local to regional scale were made during late spring in two years (Bronstert et al., 2010). At the large scale, observations of soil moisture were made with air and satellite based radar. At the small scale, the effort included permanent, temporally and locally highly resolved soil moisture measurements at two locations (Zehe et al., 2010)

Three models were parametrized for the Weis-

seritz. LARSIM (Large Area Runoff Simulation Model; Ludwig and Bremicker, 2007) is a conceptional model that includes modules for operational application for flood prediction including model internal regionalisation and correction of meteorological data. Hydrological processes included in the model are interception, evapotranspiration, snow accumulation, snow compaction and snow melt, soil water storage as well as storage and lateral transport in streams and lakes.

WaSiM-ETH is a modular, deterministic and distributed water balance model based on the Topmodel approach (Schulla and Jasper, 2001). It was used for the Weisseritz catchment with a regularly spaced grid of 100 m resolution and an hourly time step. Interception, evapotranspiration (Penman-Monteith), and infiltration (Green and Ampt approach) as well as snow dynamics are also included as modules. The unsaturated zone is described based on the Topmodel approach with the topographic index (Beven and Kirby, 1979), which determines flow based on the saturation deficit and its spatial distribution, instead of modelling the soil water movement explicitly.

Catflow (Zehe and Fluhler, 2001; Zehe and Blöschl, 2004; Zehe et al., 2005) models hydrological processes in a quasi-3D representation, using two dimensional hillslopes, resolving the downslope and the vertical dimension with an average description across the width of a slope. Processes are described on a very detailed level. For example, soil moisture processes are described with Richards equation with an additional parametrisation for macro pores.

1.5 Knowledge management in science

Full disclosure in science refers to the praxis of documenting, archiving and sharing "all data and methodology so they are available for careful scrutiny by other scientists, thereby allowing other researchers the opportunity to verify results by attempting to reproduce them."[†] Trying to meet this principle, I took several measures to make information from my work more transparent and better available.

For data management a data-base for hydrometeorological data called GOLM-DB was developed in collaboration with David Kneis, Till Franke, Theresa Blume and Mareike Eichler (Eichler et al., 2009b,a). The efforts were started during field work in Chile in 2003, with no standards available in science. Meanwhile, Horsburgh et al. (2008) published a framework very similar to what we developed. For future work, I suggest to conform to the published framework when storing time series data.

Many lines of code were written during the development of the approaches presented in this work. In line with our call for more free tools in hydrology (Buytaert et al., 2008), the most relevant functions were released under the GNU public license as packages (Reusser and Buytaert, 2010; Reusser, 2009; Reusser and Francke, 2008; Reusser, 2008) for the data analysis environment R (Ihaka and Gentleman, 1996). In order to make changes in the software better reproducible, software engineering has developed so called version management systems which allow to retrieve every version from the development process of a software. I used cvs and subversion as version management systems for my packages.

A considerable number of model runs were necessary for the model diagnostics of WaSiM-ETH. In order to make these model runs better manageable and to allow computation on multiple computers, I wrote a small wrapper software in Java for very simple distributed computing based on a shared file system. This wrapper software is conceptually similar to the interface between simulation programs and systems analysis software as suggested by Reichert (2006).

Documentation of working steps is an impor-

[†] `http://en.wikipedia.org/wiki/Scientific_method`, accessed Mai 2010

tant aspect of scientific work. All relevant working steps were documented in little journal files, making it possible to reproduce all analysis steps. While all relevant work is documented in this way, no systematic way was used to record the exact relationship among various journal files, source codes and manuscripts. Research is progressing towards improved knowledge management (Davies et al., 2005; Pepe et al., 2009) and it may be fruitful to incorporate such methods into daily practice, with the ultimate goal of "finding information instead of searching it".

Such changes in the working attitude may also affect the ways, scientific work is published. With the goal of increased quality of publications instead of increased quantity [‡], publication may go into the direction of evolving documents and more collaborative writing. First steps in this direction are public review processes such as in HESSD. Research projects such as liquidpub address practical questions of how such collaborative and evolving documents may be featured.[§]

1.6 Overview and guiding questions

The structure of my thesis is depicted in Figure 1.5. My focus is on two steps of the learning cycle, model diagnostics and cost effective measurements. The key intention is to include knowledge about context dependent behaviour of catchments into model diagnostics with the effect that additional measurements and model improvements can be much more specific to certain model components. This is the topic of the first three chapters. The snow module is identified to be insufficient thus additional snow measurement are taken to record temporal and spatial variability of snow with a limited budget. Data and results of these measurements are topic of the remaining two chapters. Note that chapter 6 presents the condensed results from a Diploma Thesis. This chapter is an early draft of a manuscript that will eventually be submitted. Guiding questions that were formulated throughout the introduction will help to highlight the major achievement of this thesis (section 7.1):

- How to assess (poor) model performance? (Chapter 2)
- How is it possible to identify temporal patterns and context dependence in model performance? (Chapter 2)
- Can we identify relevant model components (for computationally expensive models)? (Chapter 3)
- What are the limitations of WaSiM-ETH as representation of the Weißeritz catchment? (Chapter 4)
- How much information can be obtain from measurements with inexpensive temperature sensors? (Chapter 5)
- What are temporal and spatial structures of the snow in the catchment? (Chapter 6)
- What processes are required to describe the new measured data and what are the resulting updates to the model? (Chapter 6)

[‡] http://www.dfg.de/service/presse/pressemitteilungen/2010/pressemitteilung_nr_07/index.html

[§] http://liquidpub.org/

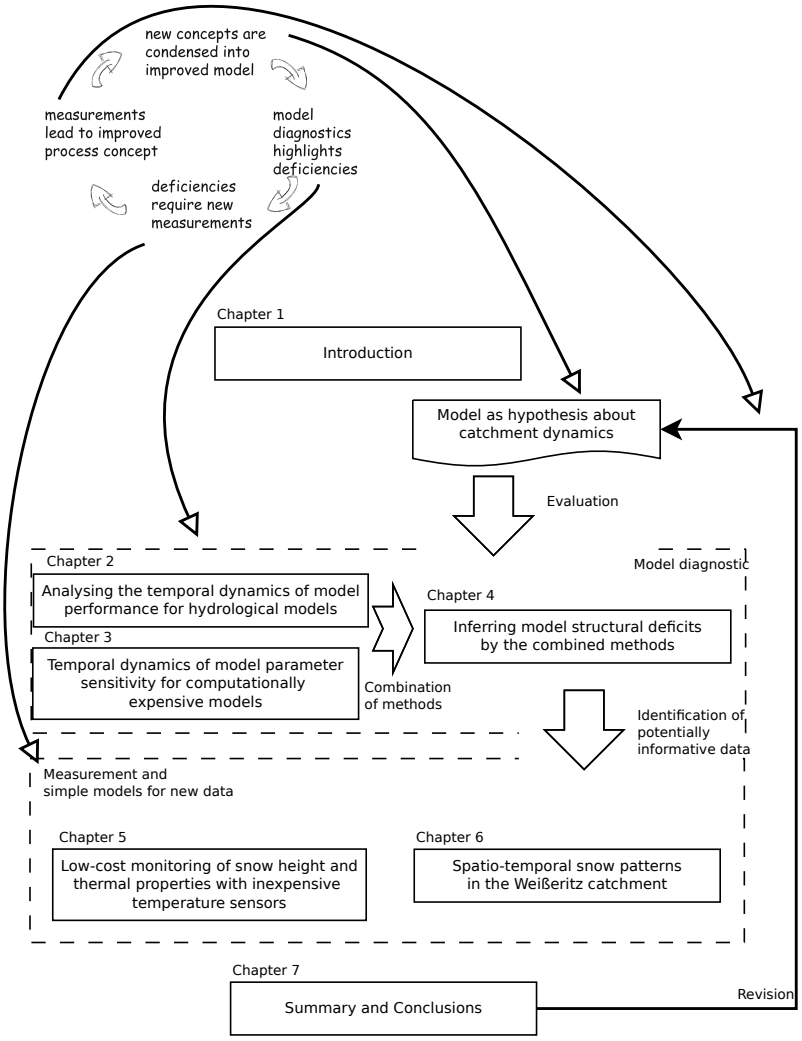


Figure 1.5: Structure of the thesis (For readability, chapter titles are short forms of the original titles)

Chapter 2

Analysing the temporal dynamics of model performance for hydrological models *

The temporal dynamics of hydrological model performance gives insights into errors that cannot be obtained from global performance measures assigning a single number to the fit of a simulated time series to an observed reference series. These errors can include errors in data, model parameters, or model structure. Dealing with a set of performance measures evaluated at a high temporal resolution implies analyzing and interpreting a high dimensional data set. This paper presents a method for such a hydrological model performance assessment with a high temporal resolution and illustrates its application for two very different rainfall-runoff modeling case studies. The first is the Wilde Weisseritz case study, a headwater catchment in the eastern Ore Mountains, simulated with the conceptual model WaSiM-ETH. The second is the Malalcahuello case study, a headwater catchment in the Chilean Andes, simulated with the physics-based model Catflow. The proposed time-resolved performance assessment starts with the computation of a large set of classically used performance measures for a moving window. The key of the developed approach is a data-reduction method based on self-organizing maps (SOMs) and cluster analysis to classify the high-dimensional performance matrix. Synthetic peak errors are used to interpret the resulting error classes. The final outcome of the proposed method is a time series of the occurrence of dominant error types. For the two case studies analyzed here, 6 such error types have been identified. They show clear temporal patterns, which can lead to the identification of model structural errors.

*D. E. Reusser, T. Blume, B. Schaefli, E. Zehe (2009), *Hydrologcial and Earth System Sciences*, 13, 999-1018

2.1 Introduction

Hydrological modelling essentially includes – implicitly or explicitly – five steps: 1) Deciding on the dominating processes and on appropriate concepts for their description. This is ideally based on data and process observations as it requires a thorough understanding of how the catchment functions. 2) Turning these concept into equations. For the more common concepts in hydrology, equations are readily available. 3) Coding and numerically solving these equations. Again, we think that it is of great advantage to use existing work if code is available (Buytaert et al., 2008). 4) Once the model structure is defined, usually a number of model parameters have to be estimated (Gupta et al., 2005). 5) Finally the model has to be tested usually based on an independent data set and we have to decide whether the model is acceptable or not. In the latter case we have to revise the initially chosen concepts and repeat steps 2–5 (see Fenicia et al., 2008, for an example of how to stepwise improve a model). However, a revision of our model concept requires a clear understanding of the model's structural deficits: What is going wrong, when does it go wrong and which part of the model is the origin?

Model evaluation is usually carried out by determining certain performance measures, thus quantitatively comparing simulation output and measured data. Various methods of model evaluation have been developed over time: Starting with visual inspection (usually used implicitly or explicitly during manual calibration) more objectivity was achieved with the calculation of performance measures, of which the most widely used in hydrology is certainly the Nash-Sutcliffe-Efficiency (Nash and Sutcliffe, 1970). Automatic calibration methods were developed based on these performance measures and lead to the realisation that a single measure is not able to catch all the features that should be reproduced by the hydrological model (Gupta et al., 1998). As a result, multiobjective calibration methods based on a range of performance measures have been and are still being developed (Gupta et al., 1998; Yapo et al., 1998; Vrugt et al., 2003).

Probably because of the development of automatic calibration procedures and their focus on the entire calibration period, the study of the *temporal dynamics* of model performance – which is implicitly used during visual inspection – did not undergo the same process of formalization.

However, we suggest that identification of temporal dynamics of performance measures can be very useful for detecting model structural errors as a first step of model improvement. This is of particular importance for operational flood forecasting because detailed knowledge about the dominant processes is necessary for credible predictions. Global performance measures are only of little use in this context, because lead times for operational forecasts are typically very short i.e. in the order of 2 to 36 h. To our knowledge, there are no studies on high resolution temporal dynamics of model performance for longer simulation periods. Pebesma et al. (2005) analyzed the temporal dynamics of the difference between observed and predicted time series for single events and used linear models to predict these differences. For longer simulation periods, it has been shown that it might be useful to split time series (for example in seasons) to obtain some minimum temporal resolution of performance measures. Choi and Beven (2007) showed with their model conditioning procedure that performance measures calculated on a seasonal scale give some additional indication about model structure deficiencies when compared to global performance measures. Similarly, Shamir et al. (2005) were able to improve identifiability of model parameters when looking at model performance on different time scales.

The rationale behind this study is that we can obtain a much clearer picture of structural model deficiencies if we know

- during which periods the model is or is not reproducing observed quantities and dynamics;

- what the nature of the error in times of bad model performance is;
- which parts / components of the model are causing this error.

A methodology to answer the first two questions is suggested here while the third topic will be the subject of a subsequent publication (see Conclusion section). The main objective of this paper is thus to present a new method to analyse the temporal dynamics of the performance of hydrological models and to be more specific about the type of error. We propose to use a combination of a) vectors of performance measures to characterize different error types, b) synthetic peak errors to support error type characterization and c) the time series of the obtained error types to analyse their occurrence with respect to observed and modelled flow dynamics.

We use multiple performance measures to capture different types of model structural deficiencies, similar to multi-objective calibration (e.g. Gupta et al., 1998; Yapo et al., 1998; Boyle et al., 2000; Vrugt et al., 2003). Dawson et al. (2007) assembled a list of 20 performance measures commonly used in hydrology. In addition, we use several performance measures introduced by Jachner et al. (2007) to test the agreement between time series in the field of ecology and which, as we will discuss, are promising for the use in the field of hydrological model calibration.

Synthetic peak errors with known characteristics will be used to better understand the model performance measures. Interpreting the values of performance measures based on modified natural reference time series has for example been proposed by Krause et al. (2005); Dawson et al. (2007). In contrast to the modified natural time series, we use an artificially generated peak as it is easier to control its properties.

As mentioned before, hydrological modelling studies do generally not analyse the temporal dynamics of model performance. However, a similar approach to the one suggested here but referring to parameter uncertainties, has been used for the dynamic identifiability analysis (Wagener et al., 2003) and the multi-period model conditioning approach (Choi and Beven, 2007), where the temporal dynamics of parameter uncertainty is analysed. The temporal dynamics of model structure uncertainties have been analysed by Clark et al. (2008), who used 79 models from a model family for their study.

The large amount of data produced in such an analysis quickly becomes overwhelming. Therefore an appropriate data reduction technique is essential to reduce the dimension of the data while at the same time loosing as little information as possible. The number of simulated time steps (N) is usually large and multiple performance measures (M) are used at each time step, therefore a set of N*M values has to be interpreted.

We propose self-organizing maps (SOM) (e.g. Kohonen, 1995; Haykin, 1999), which have already been used in several hydrological studies (see Herbst and Casper, 2008, for a short overview) and also in a comparable meteorological application where the bias of model results was determined conditional to the climatological input data (Abramowitz et al., 2008). The use of SOMs leads to a reduction of the dimension of a data set while preserving the topology of the data in a two dimensional space (i.e. similar data sets are close to each other). During this step some of the variability is lost as the number of sets N is drastically reduced (to be further explained in Sect. 2.2.3). From the SOM we will identify typical combinations of model performance measures, i.e. error types / error classes. This then leads to the assessment of the temporal dynamics of these typical combinations.

Classical methods exist to reduce M, e.g. principle component analysis, use of scatter plots (Cloke and Pappenberger, 2008), or removal of highly correlated measures (e.g. Gupta et al., 1998). In this study the analysis is performed using the full set of measures. However, only a subset of the measures is reported for readability, excluding highly correlated measures.

In the present study we propose a novel combination of key aspects of the mentioned studies as well as the use of high resolution performance measure time series and provide evidence that this is a suitable approach for model evaluation for two very different model structures.

We first present a detailed description of the methodology (Sect. 2.2) and then show its application for two case studies. These two case studies differ a) in catchment characteristics (topography, land use, soils etc.; Sect. 2.3) and b) in the hydrological model selected for simulation (process-oriented vs. physically based; Sect. 2.4). The results for the case studies are presented in Sect. 2.5 and 2.6 and discussed in Sect. 2.7. Main findings and suggested future tasks are summarized in Sect. 2.8.

2.2 Methods

The proposed methodology can be summarized as follows:

1) determination of a large set of different performance measures,

2) evaluation of the set of performance measures for a moving time window; this yields a vector of performance measures for each time step;

3) use of synthetic peak errors to interpret the values of the performance measures, i.e. to assess their error response;

4) use of SOMs and cluster analysis for data reduction and classification of error types;

5) analysis of temporal dynamics of error types with respect to measured and modelled time series.

6) removal of performance measures that have time series showing a high correlation with other time series for reporting the results;

7) analysis and characterization of error types using box plots and synthetic peak errors;

The analysis was performed with R (R Development Core Team, 2008) and the code is available as R-package (Reusser, 2009). A detailed description of the steps of the method is given below.

2.2.1 Performance measures

Dawson et al. (2007) assembled 20 performance measures used in hydrology into a test suite. This test suite includes the Nash-Sutcliffe coefficient of efficiency CE, several measures based on the absolute or squared error e.g. the mean absolute error MAE and the root mean squared error RMSE. The number of sign changes of the residuals NSC was introduced by Gupta et al. (1998). It is low if there is a bias. These and more measures are listed in Table 2.1. Detailed descriptions are available from (Dawson et al., 2007) or `https://co-public.lboro.ac.uk/cocwd/HydroTest/Details.html`. The measures have been implemented in the R package (Reusser, 2009).

Most of these measures are designed to capture the degree of exact agreement between modelled and observed values. However, we are also interested to measure the degree of qualitative agreement. Jachner et al. (2007) proposed a number of performance measures determining such a qualitative agreement (van den Boogaart et al., 2007, implemented in R;). Their measures are mainly based on MAE, MSE and RMSE defined as follows:

$$\text{MAE} = \frac{1}{n}\sum |x_{\text{obs}} - x_{\text{sim}}| \quad (2.1)$$

$$\text{MSE} = \frac{1}{n}\sum (x_{\text{obs}} - x_{\text{sim}})^2 \quad (2.2)$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum (x_{\text{obs}} - x_{\text{sim}})^2} \quad (2.3)$$

Where x_{obs} is the observed time series and x_{sim} the corresponding simulated time series. Depending on the desired qualitative comparison, they used data transformation to allow for shifts and/or

Table 2.1: List of performance measures, their abbreviations, error response group (ERG - see Sect. 2.5.2 for more details), lower (LB) and upper theoretical bound (UB) as well as the value obtained for a perfect match between model and measurement (no error).

Abr.	Full Name	ERG	LB	UB	No Error
from Dawson et al. (2007)					
MSE	mean squared error	1	-Inf	Inf	0
RMSE	root mean squared error	1	0	Inf	0
IRMSE	inertia root mean squared error	1	0	Inf	Inf[a]
R4MS4E	fourth root mean quadrupled error	1	0	Inf	0
CE	Nash-Sutcliffe efficiency	1	-Inf	1	1
PI	coefficient of persistence	1	-Inf	1	1
AME	absolute maximum error	1	0	Inf	0
PDIFF	peak difference	2	-Inf	Inf	0
MAE	mean absolute error	1	0	Inf	0
ME	mean error	3	-Inf	Inf	0
NSC	number of sign changes	9	0	LOT[b]	0
RAE	relative absolute error	1	0	Inf	0
PEP	percent error in peak	2	0	Inf	0
MARE	mean absolute relative error	1	0	Inf	0
MdAPE	median absolute percentage error	1	0	Inf	0
MRE	mean relative error	3	-Inf	Inf	0
MSRE	mean squared relative error	3	0	Inf	0
RVE	relative volume error	3	0	Inf	0
Rsqr	the square of the Pearson correlation	5	-1	1	1
IoAd	index of agreement	1	0	1	1
MSDE	mean squared derivative error	6	0	Inf	0
t_{test}	value of the paired t-test statistics	3	-Inf	Inf	0
from Jachner et al. (2007)					
CMAE	centred mean absolute error	7	0	Inf	0
CMSE	centred mean squared error	6	0	Inf	0
RCMSE	root centred mean squared error	7	0	Inf	0
RSMSE	root scaled mean squared error	5	0	Inf	0
MAPE	mean absolute percentage error	1	0	Inf	0
MALE	mean absolute log error [c]	1	0	Inf	0
MSLE	mean squared log error	1	0	Inf	0
RMSLE	root mean squared log error	1	0	Inf	0
MAGE	mean absolute geometric error	1	1	Inf	1
RMSGE	root mean squared geometric error	1	1	Inf	1
RMSOE	root mean squared ordinal error	5	0	Inf	0
MAOE	mean absolute ordinal error	5	0	Inf	0
MSOE	mean squared ordinal error	5	0	Inf	0
SMAE	scaled mean absolute error	5	0	Inf	0
SMSE	scaled mean squared error	4	0	Inf	0
SMALE	scaled mean absolute log error	1	0	Inf	0
SMSLE	scaled mean squared log error	7	0	Inf	0
SMAGE	scaled mean absolute geometric error	1	1	Inf	1
RSMSGE	root scaled mean squared geometric error	1	1	Inf	1
RSMSLE	root scaled mean squared log error	1	0	Inf	0
LCS	longest common sequence	5	0	1	1
additional measures					
t_L	lag time	8	-LOT	LOT	0
r_k	recession error	1	0	Inf	1
r_d	slope error	7	0	Inf	1
DE	direction error	8	0	LOT	0

[a] IRMSE becomes infinite for perfect match between model and observation. If the match is not perfect, small values are preferable

[b] determined by the length of the time series

[c] error of the log-transformed data

changes in scaling. To obtain measures which are insensitive to shifts, data are centred (denoted by a "C"). In order to ignore scaling, data are standardized with a linear transformation, minimizing the deviance measure ("S").

In addition, Jachner et al. (2007) provide performance measures for different scales of interest. The absolute scale is most often used and applies to the measures defined above. If the difference calculated as a ratio is of more interest (e.g. simulating twice the observed discharge, regardless of the absolute value), a relative scale ("P" from percentage), log transformed data ("L") or geometric transformed data ("G") are more appropriate (see Jachner et al., 2007, for more details). Finally they define performance measures using an ordinal scale ("O" – after transformation of the data to ranks). They also define the longest common sequence (LCS) measure: The discharge time series is reduced to a sequence of letters indicating increases ("I"), constant values ("C"), or decreases ("D"). This sequence for the observed discharge (e.g. IIIIIICCDDDDDDCCCIII) is then compared to the sequence of the simulated discharge. LCS then is defined as the longest accumulation of characters with the same order in both sequences. Thereby the method allows for deletions in one of the two series, i.e. characters can be ignored or missed (Jachner et al., 2007; van den Boogaart et al., 2007, for more details).

For this study, we complemented the above list of performance measures with the following set of four measures to obtain additional information: 1) The lag time t_L defined as the lag of the maximum in cross correlation, 2) the direction error DE, which is obtained by counting the number of times the sign of the slope differs for the observed and the modelled time series, 3) the slope error r_d and 4) the recession error r_k based on the recession constant as derived by Blume et al. (2007). r_d and r_k are defined as:

$$r_d = \frac{\frac{dx_{\mathrm{obs}}}{dt}}{\frac{dx_{\mathrm{sim}}}{dt}} \tag{2.4}$$

$$r_k = \frac{k(x_{\mathrm{obs}})}{k(x_{\mathrm{sim}})} \text{ with } k(x) = -\frac{dx}{dt}\frac{1}{x} \tag{2.5}$$

The two measures were calculated as average over the time window used to calculate the other measures (see below). Measures 2–4) work best for "smoothed" time series where noise from the measurement on short time scales has been removed.

One way to use these measures would be to translate the modelling goal into some criteria (e.g. "reproduce timing and amplitude of extreme events well") and to select the most suitable performance measures to assess them. However, we prefer a different approach. All 48 measures are calculated for a moving time window of a certain length and the vector of performance measure values for a window at a given time step t is then used as a finger print of the model performance during this time step. The finger print will be similar for time windows where the difference between model and observation has similar characteristics. Identifying and characterizing periods with comparable finger prints gives a tool to:

- objectively separate periods of differing model performance
- identify characteristics that are not easily found by visual inspection
- find recurrent patterns of differences between model and observation in longer time series.

The selection of window size depends on the process of interest and the data quality (Wagener et al., 2003). For example slow recession processes require wider windows. If data quality is suboptimal, large windows will help to reduce the influence of data errors. After some preliminary tests we selected the window size large enough to capture large events (Fig. 2.1). The selection is a compromise between looking for the local properties in the time series and having enough data to actually compute the values.

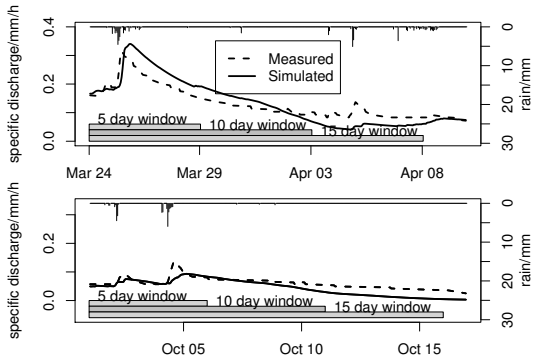


Figure 2.1: Size of the selected time window with respect to two observed events (Case study Weisseritz catchment).

The vector $\vec{p^{(t)}}$ of the M performance measures was used as finger print of the model performance for a given time step t. Of course the initial selection of the performance measures is likely to influence the result of the analysis. We regard our set of 48 measures as sufficiently large to cover the important aspects of deviations between two time series. Therefore we do not expect the results to change substantially if additional measures were added.

In order to avoid strong influence from extreme values, we transformed the values for each performance measure over all time windows to a uniform distribution in the range 0 to 1. In this transformed space, some performance measures are equivalent (e.g. MSE and RMSE). Because of this and as some performance measures behave very similarly and reporting 48 measures would make the study difficult to follow, we will report results only for a selection of the performance measures. Only one measure was used from each set of highly correlated performance measures ($|R>|0.85$ – see Sect. 2.5.1).

2.2.2 Synthetic errors

There is a need to better understand performance measures and their relationship. Two approaches exist in the literature to get familiarized with unknown measures: the first option is to calculate benchmark values for reference simple models (Schaefli and Gupta, 2007). The second option is to create artificial errors (Cloke and Pappenberger, 2008; Krause et al., 2005; Dawson et al., 2007). We used the second approach by generating synthetic errors for a single peak event as test cases (Fig. 2.2). The peak was modelled as

$$Q(t)=\begin{cases} Q_b & t<t_0 \\ Q_b * e^{(t-t_0)*k_c} & t_0<=t<t_{\max} \\ Q_b+ & \\ +\left(Q_b * e^{t_{\max}*k_c}-Q_b\right)* & t_{\max}<=t \\ *e^{(t-t_{\max})*k_r} & \end{cases} \tag{2.6}$$

Where k_r is the recession constant (negative), k_c is the constant for the rise phase and Q_b is the base flow. t, t_0 and $t_{\max}$ are the time, event starting time and the peak time, respectively. We varied the timing, baseflow, the size of the event and the recession constant to obtain the combinations shown in Fig. 2.2. Each synthetic error was generated in both possible directions of deviation (e.g. under- and overestimation) and with three different levels (small, medium and large deviation).

2.2.3 Data reduction with SOM

The dimensionality of the simulated time steps N is reduced with self-organizing maps (SOMs). A SOM (for an example see Fig. 2.3) is a method to produce a (typically) two dimensional, discretized representation of a higher-dimensional input space (Kohonen, 1995). The topological properties of the input space are preserved in the representation of the SOM. Here, the SOM helps to generate and visualize a typology of the model performance finger prints. The matrix $\mathbf{P}=(\vec{p}^{(t)})_{t=1,\ldots,N}$ of all performance measures is used as an input to the SOM. The SOM is an artificial neural network with a number $x_{\max} * y_{\max}$ of cells (or neurons) corresponding to the dimension of the map $x_{\max}, y_{\max}$.

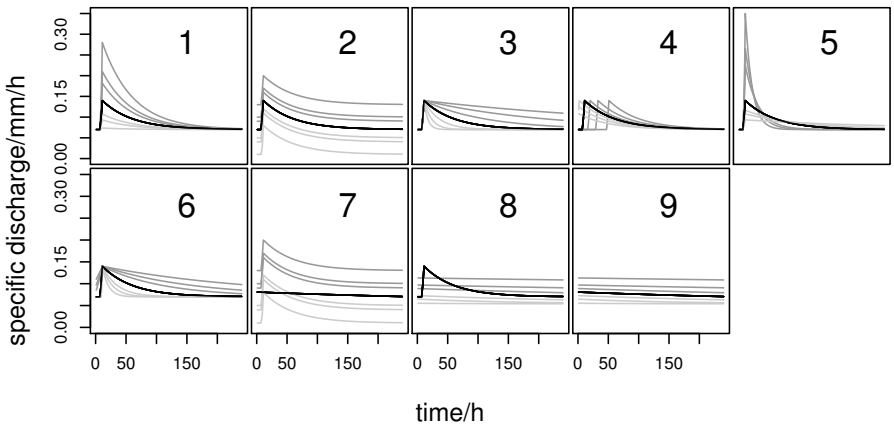


Figure 2.2: Examples of synthetic errors for a single peak event: Peak over- or underestimation **(1)**, baseflow over- or underestimation **(2)**, recession too fast or too slow **(3)**, timing: too late or too early **(4)**, maximum peak flow over- or underestimation but with correct total volume **(5)**, peak too wide (start too early, recession too slow) or too narrow **(6)**, erroneously simulated peak **(7)** or missing peak **(8)**, and over- or underestimation during a late recession phase **(9)**. The dark grey peaks will be labelled 1 to 3 with decreasing error in the remainder of this paper while light grey peaks will be labelled 4 to 6 with increasing error.

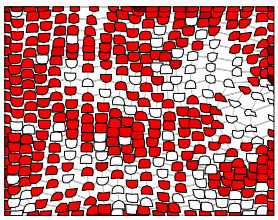


Figure 2.3: Self organizing map of the performance „finger prints" (containing 48 measures) for all N=14 827 10-day time windows (Weisseritz case study)

Each cell has a position on the map x, y and a weight vector $\vec{v}=(v_j)_{j=1,..,M}$ with the same dimension as the input vector $\vec{p}^{(t)}$. The weight vectors are initialized with random values. Then the training phase takes place with the following two steps cycling multiple times through all $\vec{p}^{(t)}$ until the weight vectors $\vec{v}$ are stable:

1) The cell most similar (best match, short BM) to the input vector $\vec{p}^{(t)}$ is determined using a Euclidean distance to the weight vector $\vec{v}$.

2) The weight for BM and its neighbours on the map are updated:

$$\vec{v}^{(i+1)}=\vec{v}^{i}+\sigma(x,y,\mathrm{BM},i)*\alpha(i)*\left(\vec{p}^{(t)}-\vec{v}^{i}\right) \tag{2.7}$$

Where x, y are the cell coordinates, $\alpha(i)$ is the learning coefficient, which monotonically decreases with iteration i and $\sigma(x, y, \mathrm{BM}, i)$ is the neighbourhood function – often a Gaussian function.

The resulting map arranges similar vectors of performance measures $\vec{p}^{(t)}$ close together while dissimilar are arranged apart. After the training phase, new input vectors can be placed on the map by finding the corresponding BM. The synthetic peak errors are placed on the map in this way in order to get a better understanding of the map.

We trained a SOM with a hexagonal and Gaussian neighbourhood with 20x20 cells with the matrix $\mathbf{P}$ as input data (Yan, 2004; Weihs et al., 2005). As mentioned before, all measures where transformed to a uniform distribution in the range [0, 1] in order to reduce effects from the differing distribution shapes and scales.

The representation of the SOM (e.g. Fig. 2.3) is based on work by Cottrell and de Bodt (1996). Each cell of the neural network is represented as a polygon. The intensity of the colouring represents the number of $\vec{p}^{(t)}$ associated with the cell (i.e. the cell weight vector $\vec{v}$ was the best match BM to the input vector $\vec{p}^{(t)}$). The shape of the polygon represents the distance (Euclidean distance) to the eight neighbouring cells. Large polygons indicate a small distance to the neighbour while if the polygon shrinks in one direction, the distance to the cell in this direction is large. Colouring of the cells can also be used to show the distribution of a specific performance measure on the map.

2.2.4 Identification of regions of the SOM

To further summarize the results, characteristic regions of the SOM with similar weight vectors $\vec{v}$ were determined using fuzzy c-means clustering (Bezdek, 1981; Dimitriadou et al., 2008). As in all clustering algorithms, the $\vec{v}$ are divided into clusters, such that they are as similar as possible within the same cluster and as different as possible between clusters. In fuzzy clustering, the $\vec{v}$ can belong to multiple clusters with all the fuzzy membership values μ_i summing up to 1. In c-means clustering the cluster memberships μ_{ki} are found by minimizing the function

$$J = \sum_{k=1}^{n} \sum_{i=1}^{c} (\mu_{ki})^m ||\vec{v}_k - \vec{w}_i||^2 \tag{2.8}$$

where the $\vec{w}_i$ are the cluster centres, $\vec{v}_k$ are the weight vectors of the SOM, and m is a parameter modifying the weight of each fuzzy membership, and $||\ ||^2$ is the Euclidean distance.

As suggested by Choi and Beven (2007), the validity index V_{XB} from Xie and Beni (1991) can be used to determine the optimal number of clusters:

$$V_{XB} = \frac{\sum_{k=1}^{n} \sum_{i=1}^{c} (\mu_{ki})^m ||\vec{v}_k - \vec{w}_i||^2}{c \left(\min_{i \neq k} ||\vec{w}_i - \vec{w}_k||^2 \right)} \tag{2.9}$$

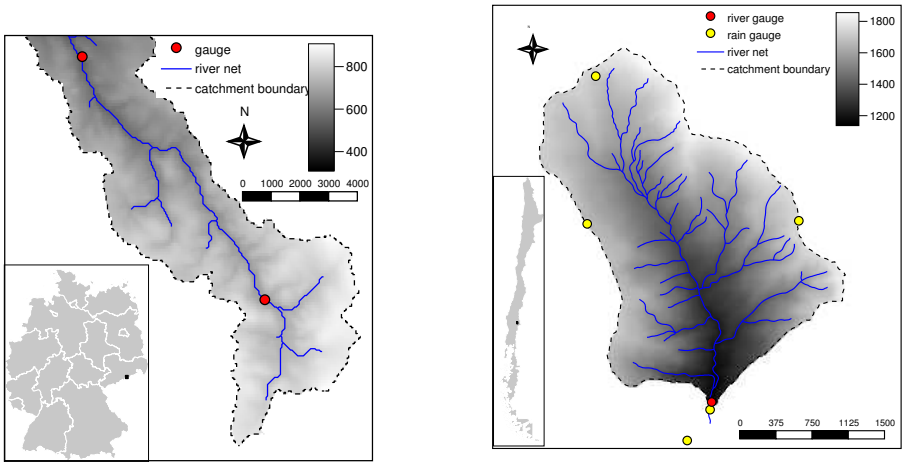
The number of clusters is thereby optimized in correspondence with the goal of the cluster analysis to have the $\vec{v}$ as similar a possible within a cluster (compactness – numerator in Eq. 2.9) and as dissimilar as possible between classes (separation – denominator in Eq. 2.9). The optimal number of clusters is the one that minimizes V_{XB}.

For the interpretation of the SOM, box plots of the performance measures for each cluster, the occurrence of the clusters in the time series and a visual inspection of the SOM are used.

2.3 Study areas

2.3.1 The Weisseritz catchment

For the first case study, the catchment of the Wilde Weisseritz, situated in the eastern Ore Mountains at the Czech-German border was used (Fig. 2.4). The lowest gauging station used in the study was Ammelsdorf (49.3 km^2). The study area has an elevation of 530 to about 900 m a.s.l. and slopes are gentle with an average of 7°, 99% are <20°; calculated from a 90 m digital elevation model (SRTM, 2002). Soils are mostly cambisols. Land use is dominated by forests (≈30%) and agriculture (≈50%). The climate is moderate with mean temperatures of 11°C and 1°C for the periods April - September



(a) Wilde Weisseritz (b) Malalcahuello

Figure 2.4: Maps of both research catchments (scales in m).

and October - March, respectively. Annual precipitation for this catchment is 1120 mm/year for the two years of the simulation period from 1 June 2000 until 1 June 2002. During winter, the catchment usually has a snow cover of up to about 1 m for 1 to 4 months with high flows during the snow melt period (Fig. 2.5 shows the pronounced peaks during spring). High flows can also be induced by convective events during summer. WASY (2006) conclude from their analysis based on topography, soil types and land use that subsurface stormflow is likely to be the dominant process. Meteorological data for 11 surrounding climate stations was obtained from the German Weather Service (DWD, 2007). Discharge data, as well as data about land use and soil was obtained from the state office for environment and geology (LfUG, 2007).

2.3.2 The Malalcahuello catchment

As a second case study the Malalcahuello catchment (Chile) was used. This research area is located in the Reserva Forestal Malalcahuello, on the southern slope of Volcán Lonquimay. The catchment covers an area of 6.26 km^2. Elevations range from 1120 m to 1856 m a.s.l., with average slopes of 51%. 80% of the catchment is covered with native forest. There is no anthropogenic intervention.

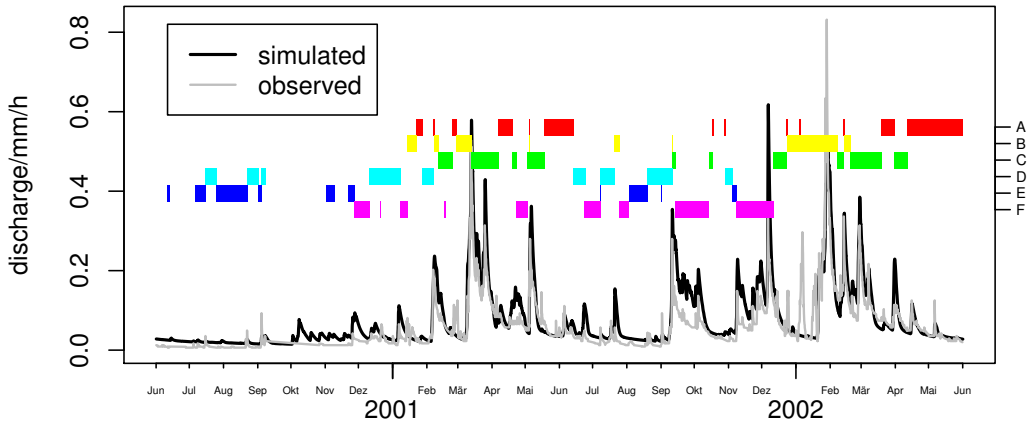
The soils are young, little developed and strongly layered volcanic ash soils (Andosols, in Chile known as Trumaos) (Iroumé, 2003; Blume et al., 2008b). High permeabilities (saturated and unsaturated), high porosities and low bulk densities are typical for volcanic ash soils. Soil hydraulic conductivities for the soils in the Malalcahuello catchment range from $1.22{*}10^{-5}$ to $5.53{*}10^{-3}$ m/s for the top 45 cm. Porosities for all horizons sampled range from 56.8% to 82.1%. Layer thickness is also highly heterogeneous, and can range from 2–4 cm to several meters. For a more detailed description of the Malalcahuello catchment see Blume et al. (2008b).

The climate of this area is humid-temperate with altitudinal effects. There is snow at higher elevations during winter and little precipitation during the summer months January and February. Annual rainfall amounts range from 2000 to over 3000 mm, depending on elevation. An overview of catchment topography and basic instrumentation is given in Fig. 2.4.

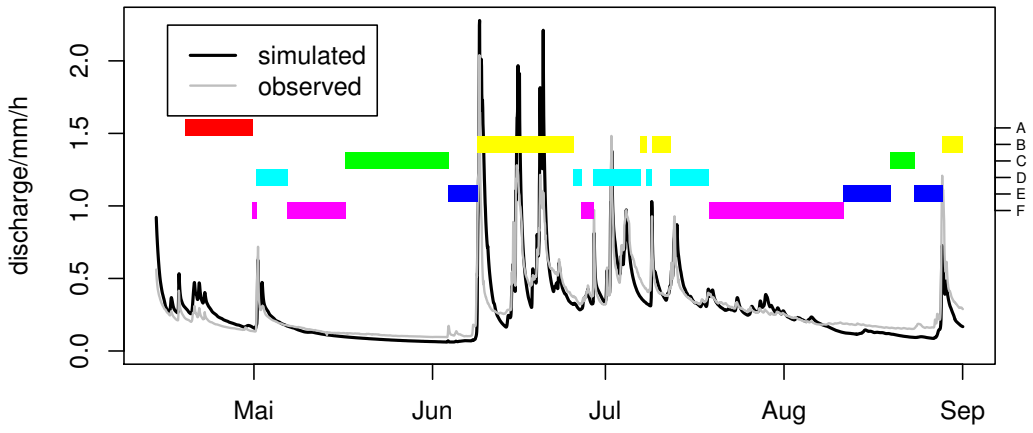
2.4 Hydrological models

2.4.1 WaSiM-ETH

As subsurface storm flow is deemed to be a dominant process in the Weisseritz catchment, the Topmodel approach (Beven and Kirby, 1979) appears suitable to conceptualise runoff generation. We therefore selected WaSiM-ETH, which is a modular, deterministic and distributed water balance model based on the Topmodel approach (Schulla and Jasper, 2001). It was used for the Weisseritz catchment with a regularly spaced grid of 100 m resolution and an hourly time step. Interception, evapotranspiration (Penman-Monteith), and infiltration (Green and Ampt approach) as well as snow dynamics are also included as modules. The unsaturated zone is described based on the Topmodel approach with the topographic index (Beven and Kirby, 1979), which determines flow based on the saturation deficit and its spatial distribution, instead of modelling the soil water movement explicitly. For the exact formulations of WaSiM-ETH see Schulla and Jasper (2001). We used an extension by Niehoff et al. (2002), which includes macropore flow, siltation and water retention in the landscape. Direct flow and interflow are calculated as linear storage per grid cell while baseflow is calculated as linear storage for the entire subcatchment. The snow cover dynamics are simulated with a temperature index approach (Rango and Martinec, 1995). The routing of streamflow is computed with the kinematic wave approach (Niehoff et al., 2002).



(a) Wilde Weisseritz



(b) Malalcahuello (2004)

Figure 2.5: Simulated and observed discharge series. The colour bars indicate the error class during this time period.

2.4.2 Catflow

The hillslope module of the physically based model Catflow (Zehe and Fluhler, 2001; Zehe and Blöschl, 2004; Zehe et al., 2005) was used to model runoff generation in the Malalcahuello catchment. It relies on detailed process representation such as soil water dynamics with the Richards equation, evapotranspiration with the Penman-Monteith equation and surface runoff with the convection diffusion approximation to the 1D Saint Venant equation. The processes saturation and infiltration excess runoff, reinfiltration of surface runoff, lateral subsurface flow and return flow can be simulated. Macropores were included with a simplified effective approach (Zehe et al., 2001). The simulation time step is dynamically adjusted to achieve a fast convergence of the Picard iteration. The hillslope is discretized as a 2-D vertical grid along the main slope line. This grid is defined by curvilinear coordinates (Zehe et al., 2001). As the hillslope is defined along its main slope line, each element extends over the whole width of the hillslope, making the representation quasi-3-D. Catflow has proved to be successful for a number of applications (Graeff et al., 2009; Lee et al., 2007; Lindenmaier et al., 2005; Zehe et al., 2001, 2005, 2006).

For this investigation the hillslope module was used to simulate a single hillslope. As the outflow at the lower end of the slope is compared with stream hydrographs measured at the main stream gauging station, this carries the inherent assumption that the structure and physical characteristics of this single slope are representative of all slopes in the catchment. While this is a strong assumption it is not completely unrealistic for the Malalcahuello catchment.

For soil parametrization values of saturated hydraulic conductivities, porosities, pF curves and fitted Van Genuchten parameters were used. Details on set-up and parametrization can be found in (Blume, 2008). 2004 data from a climate station just outside the catchment was used as climatic input data with a temporal resolution of 30 min. Rainfall time series stem from a rain gauge close to the catchment outlet.

2.5 Weisseritz case study – results

2.5.1 Performance measures

The performance measures introduced in Sect. 2.2.1 were calculated for the entire simulation period with a moving 10 day window (hourly time steps, 240 data points for each window, N=14 827). We repeated this case study also with window sizes of 5 days and 15 days in order to test the sensitivity of the method with respect to the selected window length (Sect. 2.5.5). We will report only 19 performance measures (see Sect. 2.2.1 and Table 2.2). The summary of the measures shows that the ranges of the measures vary considerably (Table 2.3).

2.5.2 Synthetic errors

The synthetic peak errors are used to improve our understanding of the performance measures. In Fig. 2.6, nine plots show the response of some representative measures (y-axis) to the synthetic peak errors, each of which is shown with a different symbol. On the x-axis, no error would be in the centre and the severity of the error increases to each side. Note that synthetic errors are generated to match the peaks of the case study (size, width, base flow). Therefore, Fig. 2.6 is valid for the Weisseritz case study and looks slightly different for the other case study. However, the following summary of the results also applies to the Malalcahuello case study. Some performance measures are very specific to a certain type of error. 23 out of 48 measures react to all peak errors, which is similar to the Nash-Sutcliffe efficiency CE in Fig. 2.6. We call this error response group (ERG) 1 (Table 2.1). This grouping is obtained by visual inspection of Fig. 2.6 and similar plots for all performance measures. The ERGs give a qualitative as-

Table 2.2: Performance measures to remove based on high correlation for the Weisseritz study. The table does not list all measures.

Measure to keep		Correlated measure ($\|R\|$>0.85) to be removed
RMSE	root mean squared error	AME, MAE, CMAE, R4MS4E, MSE
CE	Nash-Sutcliffe efficiency	RAE
PI	coefficient of persistence	IRMSE
MARE	mean absolute relative error	MdAPE, MRE, MSRE, RVE, MSLE, MAGE, MALE, MAPE, RMSGE RMSLE
MSDE	mean squared derivative error	CMSE, RCMSE, RSMSE, SMAE, SMSE
MAOE	mean absolute ordinal error	MSOE, RMSOE
RSMSGE	root scaled mean squared geometric error	RSMSLE, SMAGE, SMALE, SMSLE

sessment of the measures used in this study. Measures from ERG 2 (e.g. PDIFF in Fig. 2.6) are insensitive to the error in recession (error 3), lag (error 4) and width (error 6). These three error types do not change the maximum of the peak. Measures from ERG 3 (e.g. ME in Fig. 2.6) show no or only little sensitivity to the lag time error (error 4) and the error in peak size with correct total volume (error 5). SMSE (the only measure from ERG 4) is insensitive to errors related to shifts, the false peak, and peak size (errors 1, 2, 7, 9). Measures from ERG 5 (e.g. Rsqr in Fig. 2.6) are insensitive to errors related to shifts and peak size (errors 1, 2, 9). Measures from ERG 6 (e.g. MSDE in Fig. 2.6) are insensitive to errors related to shifts and shifts during the late recession phase (errors 2, 9). Measures from ERG 7 (e.g. SMALE in Fig. 2.6) are not sensitive for the shift only (error 2). Measures from ERG 8 (e.g. t_L in Fig. 2.6) are only sensitive to the lag time and the missing / false peak (errors 4, 7, 8). NSC (the only measure from ERG 9) has a value of 0 for most synthetic peak errors. Values above zero occur only if the sign of the error changes along the time series (errors 4, 5, 7, 8). The plots for all measures for both case studies are available from the first authors homepage.

2.5.3 Data reduction with SOM

Based on the transformed $\vec{p}^{(t)}$ of the model performance, a SOM was created. The representation according to Cottrell and de Bodt (1996) is shown in Fig. 2.3. Remember that the shape of the polygons indicates the distance between the cells and that the intensity of the colour is proportional to the number of $\vec{p}^{(t)}$ represented by a cell. No $\vec{p}^{(t)}$ are associated with white cells.

The 19 representations of the SOM in Fig. 2.7 help to identify a typology of the model performance finger prints. It is noteworthy that not all performance measures are shown (see Sect. 2.5.1). The value associated with each cell is colour coded using white for no error and black for the highest deviation from the optimal value. For performance measures with a central optimal value, no error is – again – shown in white while errors are displayed in red in one direction and blue in the other direction. A careful inspection of the SOMs (Fig. 2.7) allows identification of patterns that are related to certain errors. For example, positive lag times t_L are found in the top right corner of the SOM. In the center on the right hand side the model strongly overestimates observed peaks as indicated by negative values for t_{test} and ME, PEP, and PDIFF. However, a clear interpretation is difficult. Hence, a

Table 2.3: Summary of performance measures for the Weisseritz simulation.

Measure	Min	1st.Q	Median	Mean	3rd Q.	Max
PDIFF	-0.355	-0.059	-0.014	-0.015	0.014	0.364
ME	-0.1052	-0.0287	-0.0119	-0.0172	-0.0020	0.0614
RMSE	0.000	0.012	0.020	0.032	0.050	0.125
NSC	0.0	0.0	1.0	1.9	4.0	11.0
PEP	-343	-86	-27	-37	20	88
MARE	6.1e-02	2.9e-01	5.0e-01	7.4e-01	1.1e+00	2.6e+00
Rsqr	1.9e-08	3.1e-01	6.1e-01	5.5e-01	8.2e-01	9.8e-01
CE	-Inf	-18.27	-2.53	-Inf	-0.29	0.91
IoAd	0.00	0.27	0.48	0.48	0.71	0.98
PI	-Inf	-1008.8	-269.3	-Inf	-83.4	-5.3
MSDE	1.2e-09	8.2e-07	3.1e-06	1.1e-05	9.4e-06	1.6e-04
t_{test}	-3240.8	-44.6	-20.3	-39.7	-5.2	54.2
t_L	-20.0	0.0	1.0	2.2	5.0	20.0
r_d	-31.02	0.00	0.00	0.27	0.62	12.41
DE	0	10	24	29	41	134
r_k	0.00	0.48	1.36	1.89	2.62	14.16
MAOE	0.000	0.066	0.123	0.150	0.217	0.502
LCS	4.2e-03	5.4e-01	6.8e-01	6.8e-01	8.3e-01	1.0e+00
RSMSGE	1.0	1.2	1.2	1.3	1.4	2.5

further condensation of the SOMs is necessary to identify how different criteria cluster into different error classes and how we can interpret these error classes with respect to model failure.

2.5.4 Identification of regions of the SOM

In order to identify error classes on the SOM, fuzzy c-means clustering was applied to the weight vectors $\vec{v}$ of the SOM. The validity index V_{XB} for the identification of the optimal cluster number is shown in Fig. 2.8. Based on the V_{XB}, we chose the solution with 6 clusters for further analysis. Note that the 2 and 5 cluster solutions have similar values for V_{XB}. The 2 cluster solution combines clusters A-C and D-F from the 6 cluster solution while the 5 cluster solutions combines clusters B and D from the 6 cluster solution. Therefore, the 6 cluster solution also represents the 2 and 3 cluster solu-

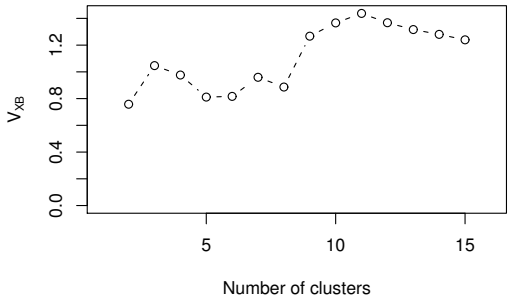


Figure 2.8: Validity index for the identification of the optimal cluster number for c-means clustering (Weisseritz case study).

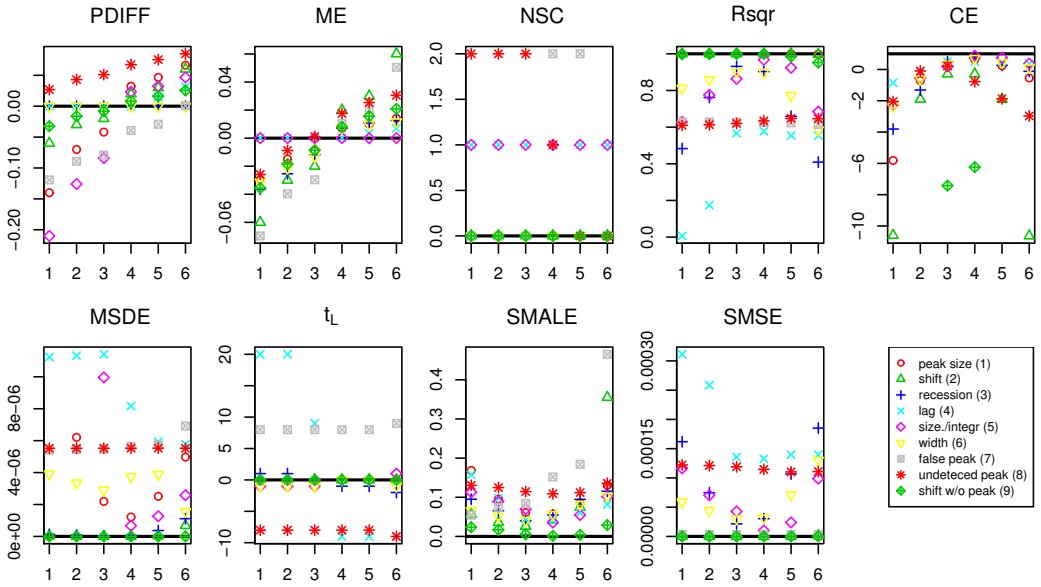


Figure 2.6: Performance measures for synthetic peak errors. Along the x-axes, the degree of error varies, with index 1 to 3 indicating a peak that is much (some, little) too large (shift to too high discharges, too slow recession, too late, too wide) and 4 to 6 indicating too small peaks. The black line indicates the position of "perfect fit".

tions. We also checked if the clustering algorithm could be applied to the $\vec{p}^{(t)}$ directly. For the two case studies presented here, we obtained equivalent results without SOMs. However, several test cases used during the development of the methodology suggested that the raw data is highly likely to not enable an identification of error clusters. In addition, the planned combination of the present method with a parameter sensitivity analysis (see also Conclusion section) will require an appropriate data reduction technique. We, thus, present here the full methodology including SOMs for data reduction.

The 6 clusters are represented with colour coding in the SOM in Fig. 2.9. Uncoloured cells do not have any associated $\vec{p}^{(t)}$ vectors As expected, the clusters form connected regions on the SOM, since similar performance "finger prints" are placed close together on the SOM.

The temporal occurrence of the error classes is

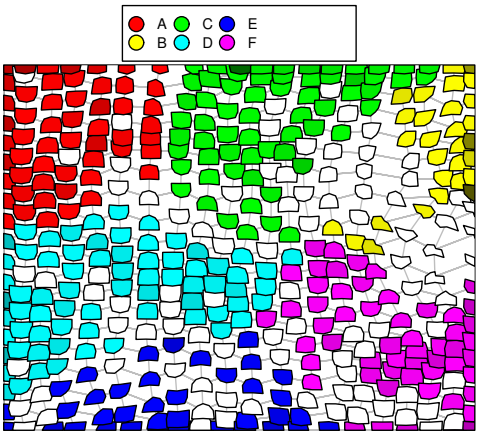


Figure 2.9: Self organizing map with color coded error cluster assignment (see Sect. 2.5.4)

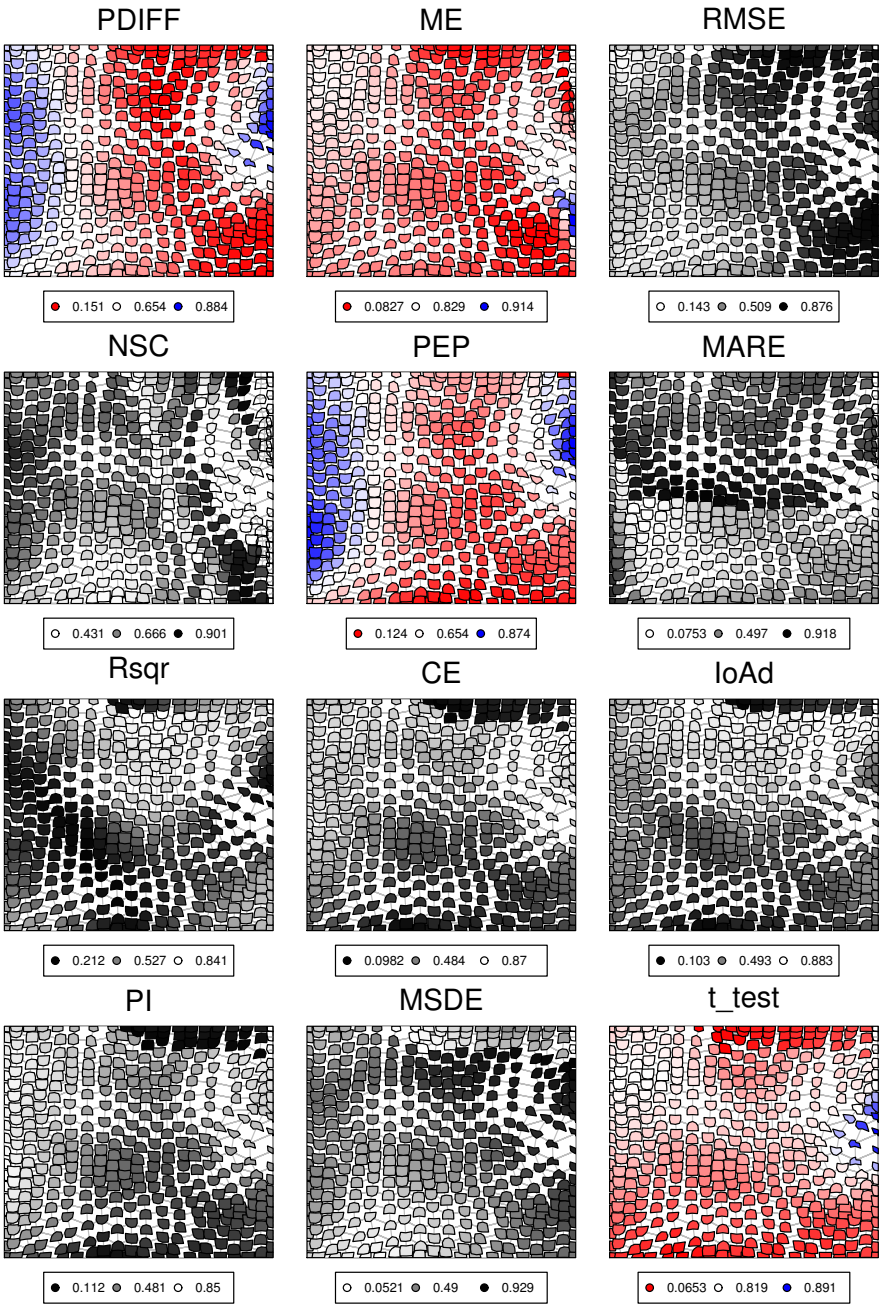


Figure 2.7: Self organizing maps. The performance measure value of each cell of the SOM is color coded. White cells indicate no error, increasing saturation of grey (for single sided performance measures), and blue and red (for double sided performance measures) indicate increasing deviation from optimal performance (see Sect. 2.5.3 for more details).

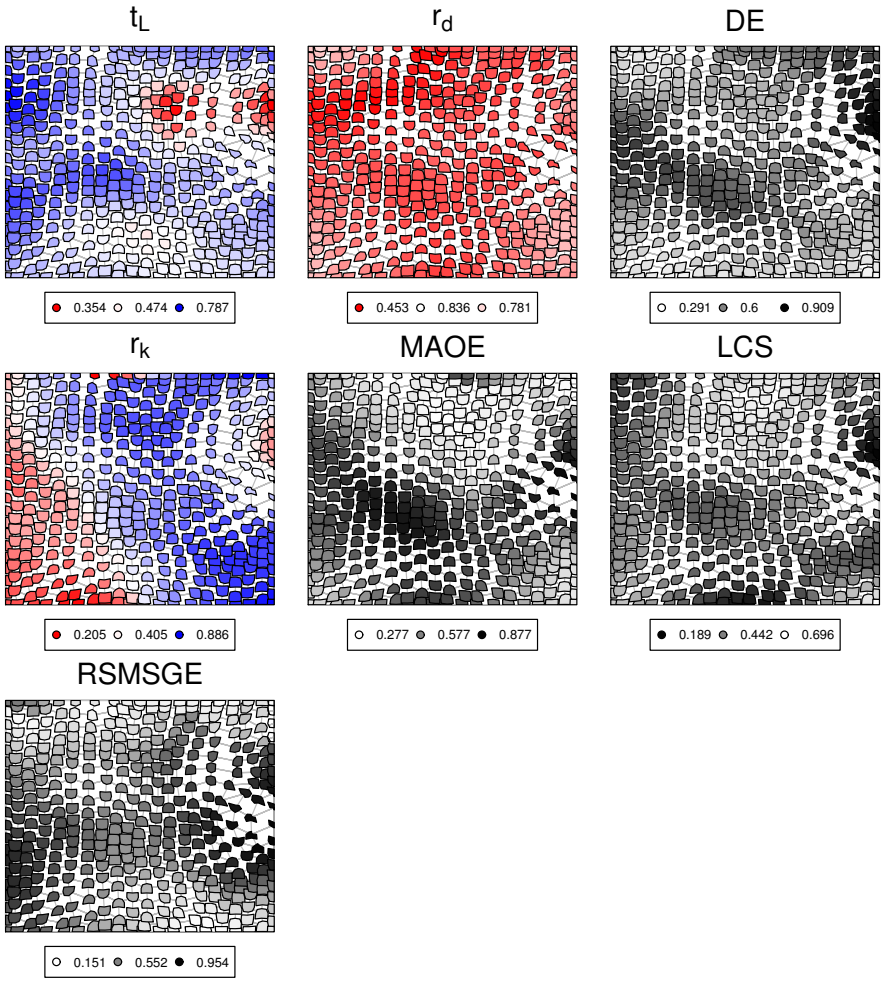


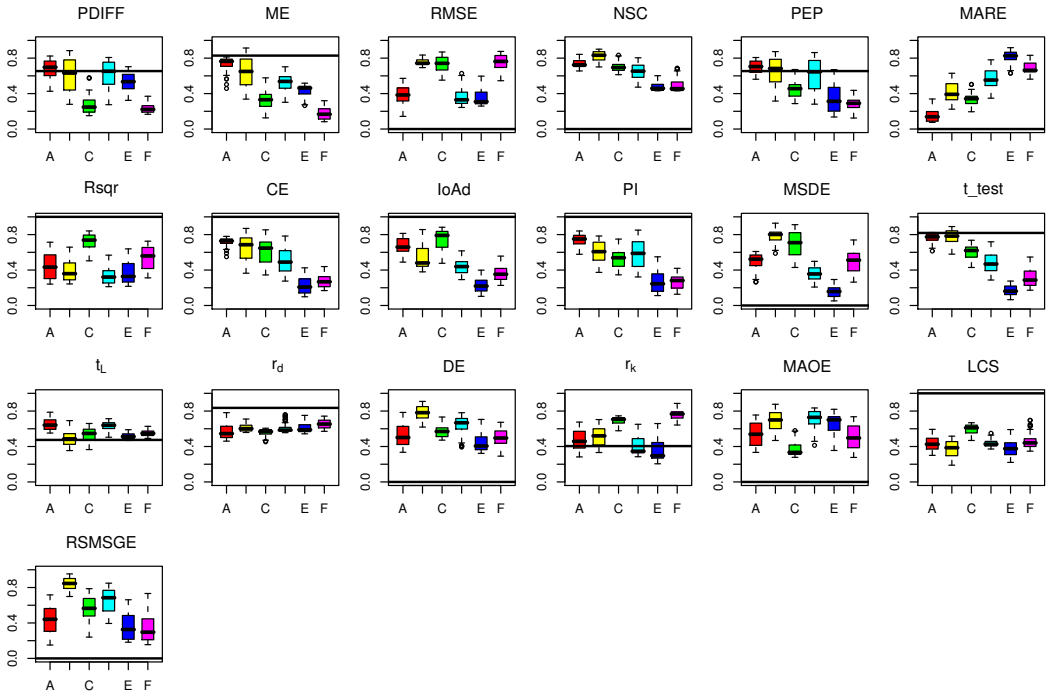
Figure 2.7: continued

shown in Fig. 2.5 as colour bars in the discharge time series. The colour coding is equivalent to Fig. 2.9. The plot shows clear patterns in the occurrence of the error classes, which are identified by visual inspection and described hereafter. Note that the cluster descriptions in parentheses will be further explained in the subsequent paragraphs. Cluster A (best fit, includes most synthetic peak errors) occurs mainly during late spring / early summer. Cluster B (underestimation, false peaks, differences for smaller values but good agreement for peaks) and C (dynamics well reproduced but overestimation) occur during snow melt events. Cluster D (bad reproduction of dynamics but small RMSE and maximum error) occurs mainly during late summer, fall and early winter. Cluster E (very bad agreement in terms of dynamics and volume, strong underestimation of peaks due to shift) occurs only a few times, mainly during the initial simulation period. Finally, cluster F (overestimation due to shift and false peaks, recession periods do not agree well, relative dynamics represented well) occurs during times where the model overestimates the observed data, mainly during summer and fall.

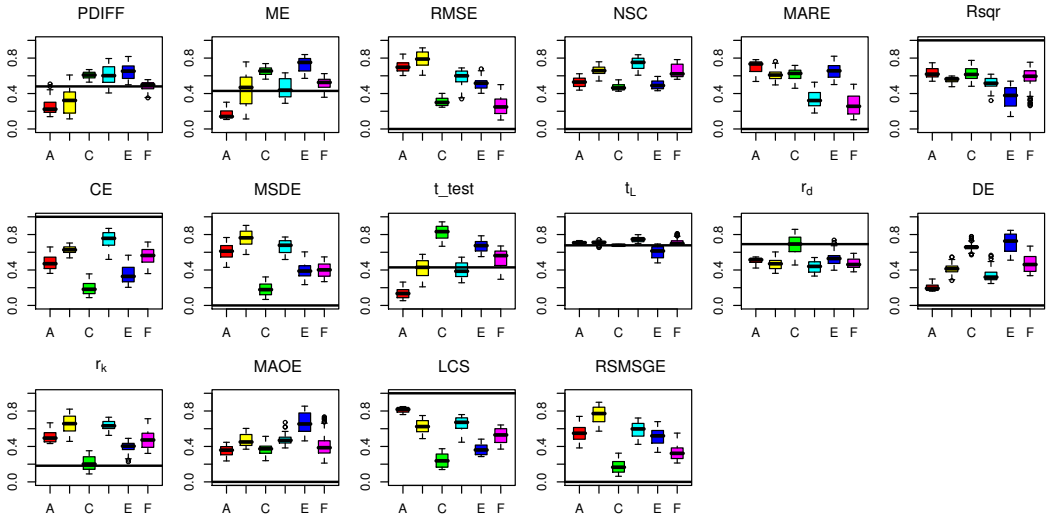
In order to associate the synthetic peak errors (Sect. 2.5.2) with the error clusters, the synthetic peak errors were placed on the SOM by finding the best matching cell (BM). Table 2.4 shows, to which clusters the synthetic peak errors are associated. Levels 1 to 3 correspond to overestimated values by the model compared to the observed data (the darker grey peaks in Fig. 2.2) while levels 4 to 6 correspond to underestimated values (to the lighter grey peaks). Cluster A includes most of the synthetic peak errors and especially the synthetic peak errors with small deviations. Cluster B includes the strong underestimation with a false peak. Cluster C includes strong overestimation due to the peak size error and errors due to undetected peaks. None of the errors were placed within Cluster D. Cluster E includes the strong underestimation of the peak due to shift. Cluster F corresponds to peaks with strong overestimation due to a shift and a shift during the late recession phase and due to false peaks. Note that cluster F is clearly related to overestimation, and Clusters B and E are clearly related to underestimation. Clusters A and C correspond to either over- or underestimation and no information is available about Cluster D from the synthetic peak errors.

Looking at the behavior of the performance measures within each cluster will provide us with more information. We therefore analyze box plots of the preformance measure values for each cluster. The box plots (Fig. 2.10) were created from the normalized weight vectors $\vec{v}$ of the cells in the SOM. The value for a perfect match between observation and model is shown as black line in the box plot. The normalized weight vectors $\vec{v}$ do not span the entire range from 0 to 1 because each cell in the SOM only represents the centre of the associated $\vec{p}^{(t)}$. The box plots are read the following way: For example, looking at PDIFF, the black line indicating a perfect match between observation and model falls within the interquartile range for clusters A, B and D. Therefore, peaks are generally matched well for these clusters. However, as the interquartile range is large for cluster B, this cluster also includes cases with strong differences between peaks. Cluster E is found slightly below the black line, which indicates that peaks are generally slightly overestimated in this cluster. Clusters C and F are found far below the black line, which shows that peaks are strongly overestimated for these clusters.

The findings from the box plots are summarized in Table 2.5. If the cluster median value was closest or the most distant from the perfect match value (no error), this cluster was entered into the table as “best” or “worst”, respectively. “Worst” was replaced by “high” and “low” if the deviation occurred to both sides of the optimal value. If the median of the second highest / lowest cluster was within the inner quartiles and on the same side of the value for no error, it was also highlighted in the table. For the example from above, PDIFF is rated best for clusters B, D and E, and low for clusters C



(a) Wilde Weisseritz



(b) Malalcahuello

Figure 2.10: Matrix of box plots comparing the normalized error measure values $\vec{v}$ (see Sect. 2.2.3). The black line indicates the "perfect fit" for each of the performance measures.

Table 2.5: Characterization of performance measures clusters derived from visual inspection of the box plots in Fig. 2.10a and 2.10b.

Cluster	Description	
Weisseritz Case Study		
A	best:	ME, RMSE, MARE, CE, IoAd, PI, t_{test}, DE, r_k, RSMSGE
	worst:	t_L, r_d, LCS
B	best:	PDIFF, t_{test}, t_L, r_k
	worst:	RMSE, NSC, Rsqr, MSDE, r_d, DE, MAOE, LCS, RSMSGE
C	best:	PEP, Rsqr, IoAd, MAOE, LCS
	worst:	RMSE, r_d
	low:	PDIFF
D	best:	PDIFF, RMSE, PEP
	worst:	Rsqr, t_L, r_d, MAOE, LCS
E	best:	PDIFF, RMSE, NSC, MSDE, t_L, DE, RSMSGE
	worst:	MARE, Rsqr, CE, IoAd, PI, t_{test}, r_d, MAOE, LCS
	low:	PEP
F	best:	NSC, r_d, DE, RSMSGE
	worst:	ME, RMSE, CE, PI, LCS
	low:	PDIFF, PEP
	high:	r_k
Malalcahuello Case Study		
A	best:	Rsqr, DE, MAOE, LCS
	worst:	MARE
	low:	PDIFF, ME, t_{test}
B	best:	ME, t_{test}
	worst:	RMSE, MSDE, r_d, r_k, RSMSGE
C	best:	RMSE, NSC, Rsqr, MSDE, t_L, r_d, r_k, MAOE, RSMSGE
	worst:	CE, DE, LCS
	high:	PDIFF, ME, t_{test}
D	best:	ME, MARE, CE
	worst:	NSC, r_d, r_k
	high:	PDIFF, t_L
E	best:	NSC
	worst:	MARE, Rsqr, DE, MAOE
	low:	t_L
	high:	PDIFF, ME
F	best:	PDIFF, ME, RMSE, MARE, Rsqr, MAOE
	worst:	r_d

Table 2.4: Cluster allocation of synthetic peak errors. For details on peak characteristics see Figs. 2.2 and 2.6. Levels 1–3 generally overestimate flow while levels 4–6 underestimate it.

Weisseritz Case Study		
Cluster	Error	Levels
A	peak size (1)	2 3 4 5 6
	shift (2)	2 3 4 5
	recession (3)	2 3 4 5 6
	lag (4)	1 2 3 4 5 6
	size./integr (5)	2 3 4 5 6
	width (6)	1 2 3 4 5 6
	undeteced peak (8)	2 3 4 5 6
	shift w/o peak (9)	2 3 4 5 6
B	false peak (7)	6
C	peak size (1)	1
	recession (3)	1
	size./integr (5)	1
	false peak (7)	4 5
	undeteced peak (8)	1
E	shift (2)	6
F	shift (2)	1
	false peak (7)	1 2 3
	shift w/o peak (9)	1
Malalcahuello Case Study		
Cluster	Error	Level
A	peak size (1)	1 2
	shift (2)	1 2 3
	recession (3)	3
	width (6)	1 2
	false peak (7)	1 2 3
	shift w/o peak (9)	1 2 3
B	shift (2)	5 6
	recession (3)	1 2 5 6
	lag (4)	6
	size./integr (5)	1
	width (6)	6
	false peak (7)	4 5
	undeteced peak (8)	1 2 3 4 5 6
C	shift w/o peak (9)	5 6
D	peak size (1)	5 6
	shift (2)	4
	recession (3)	4
	lag (4)	1 2 3 4 5
	size./integr (5)	2 3 5 6
	width (6)	3 4 5
E	false peak (7)	6
F	peak size (1)	3 4
	size./integr (5)	4
	shift w/o peak (9)	4

and F.

From the box plots (Fig. 2.10) and Table 2.5 we find that cluster A shows the best fit according to 9 performance measures. In this cluster there is thus a good agreement in (high flow) dynamics (CE, PI) and amounts (ME, RMSE, MARE, t_{test}) of simulated and observed stream flows. Peaks are late (t_L above target values) and the derivative is sometimes overestimated. LCS is the worst for cluster A. Since LCS is quite far from the optimal value for all clusters, this fact is negligible.

Cluster B has a good match between the observed and modelled time series in terms of high flows (PDIFF, CE, PI, t_{test}). Dynamics are not represented very well by the model (Rsqr, DE, MSDE), and data do not agree well after rescaling and ordering (MAOE, RSMSGE). Overall, this indicates differences for smaller values but good agreement for large values. For Cluster C, dynamics are matched reasonable (best values for PEP,Rsqr, IoAD, LCS, MAOE) but levels do not agree well (PDIFF). Also RMSE is high. For Cluster D on the other hand, the agreement is reasonable in terms of level (PDIFF, PEP, RMSE) but dynamics are not reproduced well (Rsqr, t_L, MAOE, LCS). Cluster E shows bad agreement between model and observation in terms of dynamics (Rsqr, CE, IoAd, PI, r_d, LCS) and level (t_{test}). The observed best values for PDIFF,RMSE, MSDE, t_L, DE and RSMSGE are initially somewhat surprising but can be explained by the fact that this cluster is related to low flow periods with little dynamics. In Cluster F, the level is not well represented as indicated by bad values for ME, RMSE, CE, PI, PDIFF and, PEP. Also, recession periods do not match well (r_k). Good values for r_d, DE and RSMSGE indicate that the relative dynamics are matched relatively well for cluster F.

2.5.5 Sensitivity for the size of the moving window and the size of the SOM

The entire case study was repeated two more times with a moving window of 5 days and 15 days, in

order to test the sensitivity of the method for this choice. In short, the alternative window sizes resulted also in 6 clusters. The identified clusters had very similar error types and the temporal occurrence of the clusters was comparable to the 10 days window, the solution we retained for the present paper. In general, with smaller window sizes, the temporal occurrence of the error clusters becomes more fragmented.

The entire case study was also repeated with SOM sizes of 10x10, 15x15, 25x25, 30x30, and 10x20. In this case, solutions were found for 5 or 6 clusters. The solutions with 5 clusters (30x30) combined two of the clusters presented above to a single cluster. Again, descriptions of the error types and temporal occurrence of the clusters were similar. The validity index and the interquartile ranges on the box plots (comparable to Fig. 2.10) were generally smaller for SOMs with a smaller number of cells because more variability was reduced during the generation of the SOM.

Detailed results (plots and tables) are available on the corresponding authors homepage at `http://www.uni-potsdam.de/u/Geooekologie/institut/wasserhaushalt/hessd_homep`.

2.6 Malalcahuello case study – results

2.6.1 Performance measures and synthetic errors

For the Malalcahuello case study a time window of 120 h (5 days; hourly time step, 120 points) was chosen as streamflow here is faster in response and dynamics than in the Weisseritz catchment. After excluding correlated measures, a set of 16 performance measures (N=3241) remained. All of these measures were also used in the Weisseritz case study. The 9 synthetic errors proposed in Sect. 2.2.2 were adapted for the time window as well as the range in flows.

2.6.2 SOM and fuzzy clustering

As in the Weisseritz case study, data reduction was achieved by producing a self-organizing map. 6 error clusters were identified. Looking at the distribution of the error clusters over the time series (Fig. 2.5) we find a distinct pattern of errors, which mainly occur in larger blocks.

Cluster A (good correlation but overestimation) was attributed to a longer period in April. Again, the descriptions in parenthesis will be further explained below. Cluster B (strong differences in peak width – including recession errors, false and undetected peaks – large errors also for rescaled data, bad performance in terms of derivatives) is allocated to a series of peaks in June. Times attributed to cluster C (small RMSE but dynamics not reproduced well, underestimation of recession phase) are the late recessions in May and August. These periods have very little dynamics and the model does indeed show a general underestimation of flow. Cluster D (dynamics well reproduced, low mean errors, time lags) occurs in shorter time blocks in May and late June / beginning of July. Cluster E (worst performance, underestimation with false peaks) is attributed to the late recessions in June and August. Some of the discrepancies in dynamics, especially in August, are the result of snow melt. As Catflow does not contain a snow model, these dynamics cannot be reproduced in the simulation. The early recession phases in May and July / August are attributed to cluster F (good reproduction of long term behaviour / balance, bad scores for the ratio of the recession constant).

Locating the synthetic peak errors (corresponding to Fig. 2.6) on the SOM (see Table 2.4) leads to the following characterization: Cluster A contains most of the overestimating synthetic errors. Cluster B includes the slight underestimation due to a false peak (error 7) and the extreme peaks related to wrong recessions (error 3). In addition, the most extreme too early lag time error (error 4) and the most extreme overestimating errors due to peak

size with correct integral and undetected peaks are found in this cluster. Most of these synthetic errors are related to a strong difference in peak width. Cluster C contains the most extreme error shifting the modelled below the measured time series in absence of a peak (error 9). Cluster D includes a number of intermediate / underestimating errors and all but one error related to lag times. Cluster E includes the underestimating error due to a false peak (baseline shifted far below the reference). Cluster F contains the intermediate errors related to peak size with and without correct total volume and shift during the late recession phase.

The box plots for each performance measures and clusters are shown in Fig. 2.10. A summary of the specific characteristics of each cluster is given in Table 2.5. Cluster A shows the best performance for those measures looking at the correlation of the time series (Rsqr, DE, LCS, MAOE) but has the characteristic values for overestimating the time series in general (ME and t_{test} below aim). Peaks are also overestimated (PDIFF below aim). Cluster B strongly overestimates the peaks (RMSE, PDIFF low) and fits the worst after rescaling (RSMSGE). Also, derivative based measures are worst for this cluster (r_k, r_d MSDE). Good values for t_{test} and ME and intermediate values for CE and Rsqr indicate that the dynamics are still reproduced quite well. Cluster C shows good performance for derivative based measures and a small RMSE but dynamics (CE, LCS) and peaks (PDIFF, ME and t_{test}) are badly reproduced. For Cluster D, dynamics (CE) and overall volume (ME, t_{test}) agree well. However, derivative based measures (r_d, r_k) show bad values. A high NSC indicates that the modelled time series changes often between lying above and below the measured time series. Cluster D thus describes times where the model has only slight over and underestimation in peaks, quite good correlation and low mean errors. Cluster E can easily be identified as having the worst performance measures (scores worst on 7 of the performance measures and best only for the NSC). Peaks as well as the overall time series are underestimated (PDIFF and ME above target value). The correlation between modelled and measured time series is low as it has the worst scores on Rsqr, MARE, MAOE, and DE. Finally, cluster F might be regarded as the best performing cluster. However, it corresponds to recession periods with little dynamics, therefore CE values are only intermediate. Scores are good for mean and mean relative errors (ME, MARE) and RMSE. However, the derivatives r_d do not match well.

2.7 Discussion

In both case studies we found 6 classes or clusters of model performance (Fig. 2.10). A temporal pattern of the occurrence could be identified in both cases, indicating that the model has different deviations during different phases. For the Weisseritz simulation we found the following weaknesses:

- times of "best" performance (cluster A) still show a great range of variability (most synthetic peak errors attributed to this period)
- completely missing peaks during snow season (cluster B). More detailed analysis showed that these were events occurring at times with reported temperatures well below freezing - which must be clearly radiation induced melt events. This process is missing in the model.
- major snow melt events are generally overestimated
- periods during summer / fall, where observed peaks are completely missing
- strong underestimation of low flow during late summer, together with
- strong overestimation of recession periods occurring during autumn, which indicates that soil and interflow storage is not well parametrized.

From this analysis, we suggest to test the following model improvements. The snow melt component may be better suited for this catchment after including radiation induced snow melt. We will check the data again very carefully for the peaks that are completely missing during summer periods. If the data is valid, we are likely to miss an important process in the model. We will also try to further improve the parametrization of the soil and interflow storage. However, as model runs take about 20 minutes, classical calibration methods with more than 1000 required runs are time consuming. Strong storage parameter interactions in WaSiM-ETH with the Topmodel soil storage additionally complicate calibration attempts.

For the Malacahuello case study the main findings are:

- During the first month, the model overestimates the observed discharge, indicating too high initial filling of the soil storage.
- In the recession period in August, the model completely fails to reproduce stream flow dynamics
- The three major events in June form a distinct group as they are strongly overestimated by the model. Both the missed dynamics in August as well as this strong overestimation are likely to be the result of the lacking representation of snow dynamics in the model.
- flow was found to be underestimated during the longer recession periods.

The first step for model improvement will be to include a snow module. The long-term storage behaviour could probably be improved by coupling the model with a ground water model. Moreover, the evaluation exercise shows that the observed discharge data needs to be preprocessed in order to remove variability / noise on the very short time scales.

While some of the identified errors are already apparent in a first visual inspection of the model output, others are less obvious and might be overlooked – especially for longer simulation periods.

2.8 Conclusions

This paper presents a new method to analyse the temporal dynamics of the performance of hydrological models and to characterize the types of errors. This new method is consistent with the diagnostic evaluation approach presented by Gupta et al. (2008). They suggest to use "signature indices that measure theoretically relevant system process behaviors" and argue that a single criterion is not sufficient for diagnosis of current environmental models. Instead, multiple diagnostic signatures should be derived from theory and used to compare modelled and observed behavior. This corresponds to the main idea of the performance finger prints presented in this paper.

The developed methodology combining time-resolved performance analysis and data reduction techniques is applied successfully in two case studies. These two case studies differ strongly in both, model type and runoff generation processes and thus the method seems to be applicable for a wide range of research areas and modelling approaches.

In the two case studies, a set of uncorrelated performance measures calculated for a moving 5 or 10 day window is used to characterize the temporal dynamics of the model performance (model performance finger print). As the results show, the combination of multiple measures provides a better characterization of the performance compared to any single measure, which agrees with the basic idea of multi-objective calibration.

Self organizing maps (SOM) are used to reduce the amount of data and in a subsequent step, different clusters of performance finger prints are identified. These clusters are in fact not readily identifiable in the raw data (before data reduction).

To test the sensitivity of the performance measures as well as to characterize the error clusters, the presented model diagnostics methodology in-

cludes synthetic peak errors. They show that some performance measures are very specific for a certain type of errors while others react to all types of error. Some of these errors are visible in visual inspection of the simulated and the observed reference time series. However, as illustrated for the two case studies, analyzing the temporal patterns of the identified error types gives valuable additional insights into model structural deficiencies.

In summary, the proposed methodology has the following main benefits:

- Identification and separation of time periods with different model performance characteristics are achieved in an objective way.
- Long simulation periods, for which analysis of single events becomes almost impossible can be processed. Recurrent patterns become apparent.
- Subtle but important differences between observation and model can be detected.

Especially the patterns of error repetition are likely to contain valuable information if they can be connected to parameter sensitivities. The next step will thus be to combine the analysis of the temporal dynamics of model performance with the analysis of the temporal dynamics of parameter sensitivity in order to enhance our understanding of the model. The model performance will tell us, during which periods the model is failing while the parameter sensitivity will show, which model component is the most important during these periods. Overall the methodology presented here proves to be viable and valuable for the analysis of the temporal dynamics of model performance.

Acknowledgements This study has been funded as part of OPAQUE (operational discharge and flooding predictions in head catchments), a project within the BMBF-Förderaktivität "Risikomanagement extremer Hochwasserereignisse" (RIMAX). We would like to thank Jenny Eckart for her support with the data preprocessing for WaSiM-ETH. A major part of the analysis was carried out with the free statistical software R and contributed packages, we would like to thank its community.

Chapter 3

Temporal dynamics of model parameter sensitivity for computationally expensive models with FAST (Fourier Amplitude Sensitivity Test) *

The quest for improved hydrological models is one of the big challenges in hydrology. When discrepancies are observed between simulated and measured discharge, it is essential to identify which algorithms may be responsible for poor model behaviour. Particularly in complex hydrological models, different process representations may dominate at different moments and interact with each other, thus highly complicating this task. This paper investigates the analysis of the temporal dynamics of parameter sensitivity as a way to disentangle the simulation of a hydrological model and identify dominant parameterisations. In a first part, three existing methods, (the Fourier amplitude sensitivity test, extended Fourier amplitude sensitivity test and Sobol's method) are compared by applying them to a Topmodel implementation in a small mountainous catchment in the tropics. For the major part of the simulation period, the three methods give comparable results, while the Fourier amplitude sensitivity test is much more computationally efficient. In a second part, this method is applied to the complex hydrological model WaSiM-ETH implemented in the Weisseritz catchment, Germany. A qualitative model validation was performed based on the identification of relevant model components. The validation revealed that the saturation deficit parameterisation of WaSiM-ETH is highly susceptible to parameter interaction and lack of identifiability. We conclude that temporal dynamics of model parameter sensitivity can be a powerful tool for hydrological model analysis, especially to identify parameter interaction as well as the dominant hydrological response modes. Finally, an open source implementation of the Fourier amplitude sensitivity test is provided.

*Dominik Reusser, Wouter Buytaert, Erwin Zehe (in review), *WRR*

3.1 Introduction

Rainfall runoff models have become important tools to represent and test our knowledge about the processes in a hydrological catchment. One of the most important aims in model building is to keep the model structure as parsimonious as possible, to aid calibration and uncertainty analysis, and to avoid parameter interaction and lack of identifiability.

For this purpose, it is necessary to identify dominant hydrological processes and to parameterise them adequately in the model as functional components. This is often not straightforward. Depending on the hydrological context (e.g., rainfall driven; energy driven; occurrence of snowmelt) different processes will be active in the hydrological system at different moments in time. Ideally, in a parsimonious model with low parameter interaction, this should be reflected in the model structure, with different model components dominating simulated dynamics over time. Hence, we expect simulation results to be most sensitive to variations of exactly those parameters that belong to the corresponding model component. For instance, we expect a good model to be sensitive to variation of snow melt parameters during snow melt periods, but rather insensitive in snow free periods.

This paper explores the application of the Fourier Amplitude Sensitivity Test (FAST) as a powerful and computationally efficient method to analyse and visualise the temporal dynamics of model parameter sensitivity for computationally expensive hydrological models.

3.1.1 Sensitivity analysis for temporal dynamics

Sensitivity analysis (SA) assesses the impact of model parameters on the model outcome, and is therefore a convenient tool to assess model behaviour and particularly the importance of certain parameterisations within the model.

Classically, SA is of most interest in the context of model calibration. The goal is then to determine the most important parameters for the calibration process as well as the unimportant parameters that may be fixed at a predefined value. Therefore, in hydrology, sensitivity is most often calculated for some objective function, for example the root mean square error RMSE or the Nash-Sutcliffe coefficient of efficiency. In contrast, we do not approach the question of sensitivity from a calibration point of view. By analysing temporal dynamics of parameter sensitivity (TEDPAS) of model output variables, such as discharge, groundwater level or snow water equivalents, we can quantify which model components dominate the simulation response. This information can then be used as an indicator for dominant processes in the catchment, as well as the functioning of the model. TEDPAS as an analytic tool for identification of dominant model components has been reported before by Sieber and Uhlenbrook (2005) and Cloke et al. (2008). The same SA methods can be applied for both approaches, with the main difference that for TEDPAS, SA is performed for each time step individually.

Using TEDPAS as an analytic tool is related to the dynamic identifiability analysis introduced by Wagener et al. (2003). However, they serve a different purpose. Identifiability analysis aims at identifying parameters that can be confined by given observations. It is a necessary, but not sufficient condition that parameters must be sensitive to be identifiable. Non-identifiable but sensitive parameters occur for instance, when parameters are strongly correlated as for the Nash cascade where a decrease of one parameter can be compensated by increasing the other (Bárdossy, 2007).

3.1.2 Sensitivity analysis methods

A wide range of SA methods exist. Many methods characterise local gradients at a given point in parameter space by assessing the response of the model output to a small variation of single parameters (the so-called one at a time method).

This is a sensible approach if SA is used in the context of model calibration. The main disadvantage of this method (and other local SA methods) is that information is available for this very specific location in parameter space only, which is usually not representative for the physically possible parameter space. To overcome this problem global SA methods have been proposed, where multiple locations in the physically possible parameter space are evaluated. Global methods may be used without prior calibration of the model, which may reduce the total computing time required considerably as calibration often requires a large number of model runs. Global SA with regression based methods rests on the estimation of linear models between parameters and model output. The method provides good estimates of parameter sensitivity for nearly linear models, but fail if the model output shows non-linear (especially non-monotonic) dependence on model parameters, which is very common for hydrological models. Regional sensitivity analysis (RSA) (Hornberger and Spear, 1981) and derived methods approach the question by comparing an initial distribution of model parameters to the distribution after conditioning of the model output to observations. This approach is more suited to find identifiable parameters compared to sensitive parameters (see above). Finally, a number of methods are based on ANOVA-like analysis of the dependence of the model output variance to simultaneously modified parameters (partial variance based methods):

$$V = \sum_i V_i + \sum_{i<j} V_{ij} + \cdots + V_{1,2,3,\cdots,n} \quad (3.1)$$

V is the total variance, V_i is the variance caused by parameter θ_i (first order variance), V_{ij} is the covariance caused by θ_i (second order variance) and θ_j and higher order terms show the variance contribution from multiple parameters. Sensitivities in terms of partial variance are then calculated by dividing by the total variance V. Therefore, all such defined sensitivities add up to 1. Variance based methods result in reliable estimates of sensitivities also for strongly non-linear models, as was often demonstrated using examples where the analytical solution can be computed (e.g. Saltelli and Bolado, 1998). The main drawback of these methods is that they are not easy to implement and the required number of model runs is very high (usually >1000) for most approaches. The most important variants of this method are Sobol's method 2001 and the (extended) Fourier Amplitude Sensitivity Test ((e)FAST – Schaibly and Shuler, 1973; Cukier et al., 1973, 1975; Fang et al., 2003; Saltelli and Bolado, 1998).

Some of the recent studies applying SA to rainfall runoff, flood inundation, and water quality models are listed in Table 3.1. 5 out of the 14 studies use variance based methods. In 5 studies, on the order of 10'000 model runs were computed to calculate sensitivities, which is impossible for computationally expensive models. To our surprise, we were unable to find an application of FAST or eFAST to rainfall-runoff modelling.

The selection of the appropriate method for analyzing parameter sensitivity depends strongly on the goal of the sensitivity analysis (Saltelli et al., 2006, and Figure 3.1): 1) If the correct values of a parameter can be fixed from additional, independent data before calibration, then which parameter causes the greatest reduction in variance (called factor prioritisation setting in Saltelli et al., 2006)? This use is illustrated in Figure 3.1A). On the left hand side, the distribution of parameter values (normalized to the range between 0 and 1) is shown. The dots illustrate the physically possible parameter space while the grey box for parameter 2 (P2) indicates the parameter range, to which P2 is fixed from independent data without calibration. In reality, there is no true parameter set, because the parameters are not observable (Beven, 2002; Zehe et al., 2007) or can not be identified (Klaus and Zehe, 2010), however we assume perfectly determinable parameters for the illustration of the factor prioritisation setting.

Study	Model	Parameters	Method	Runs	FOPV[a]	Evaluated model output
Deflandre et al. (2006)	QUESTOR[b]	5 to 12 parameters:chemical reaction constants and oxigen exchange constants	eFAST	500	x	Nash-Sutcliffe coefficient of efficiency
van Werkhoven et al. (2009)	SAC-SMA	14 p. for upper (3 p.) and lower (5 p.) zone, partition. (3 p.) and percol. (3 p.)	Sobol's method	7.5 *10^6	x	4 objective functions
Tang et al. (2007a)	SNOW-17 – SAC-SMA	5 p. for SNOW (precip. correction, melting factors (2 p.), wind and SCA parameter) and 13 p. for SAC-SMA (see above)	Sobol, RSA, ANOVA, PEST	$\leq 10'000$	x	RMSE of discharge, RMSE of Box-Cox transformed discharge
Cloke et al. (2008)	ESTEL-2D	9 p.: moisture, hydraulic conduct., Brooks-Corey and van Gen. p., storage p., upslope pressure, river stage, rainfall	MMGSA [c]	< 1280	x	fuzzy membership (overall, $f(t)$); sum squared errors
Pappenberger et al. (2008)	HEC-RAS	6 p.: input quality (1 p.), 3 p. for roughness, p. for numerical solution, downstream initial slope	multiple [d]	not reported	x	mean absolute error; Nash-Sutcliffe measure
Pappenberger et al. (2006)	HEC-RAS[e]	3 surface roughnesses, 3 model input p. and 1 p. for numerical solution	SARS-RT [f], correlation, RSA	3000		inundation performance measure
Demaria et al. (2007)	VIC (variable infiltation capacity)	10 p.: base flow, layer thickness, hydraulic cond., infiltration, Brooks-Corey p.	RSA+ (Freer96), scatter plots	60'000		RMSE, RMSE box-cox, ARE (absolute relatvie error) of discharge
McIntyre et al. (2003)	WaterRAT	26 p. related to chemical processes	RSA	10'000		water quality concs.
Sieber and Uhlenbrook (2005)	TAC^D [g]	20p.: precip. (2 p.), snow (7 p.), soil (13 p.), runoff (22 p.) and routing (4 p.)	RSA, regr. with L. Hypercube sampling	400		q(t), Nash Sutcliffe of q and log(q)
Christiaens (2002)	MIKE SHE	soil hydraulic paramters	regression with L. Hypercube sampling	25		multiple [h]
Van Griensven et al. (2006)	SWAT	41 parameters related to snow, soil, groundwater, geomorphology, evaporation, channel flow, runoff, erosion and crop	Morris modified: multiple OAT with L. Hypercube sampling	not reported		Total sums and Sum squared errors of flow, sediment and nutrients
Foglia et al. (2009)	TOPKAPI	35 (thickness of soil, hydraulic conductivity, water content, Manning roughness)	OAT local SA	71		discharge (composite sensitivity)
Benke et al. (2008)	2C-model	5 p.: 2 store shape p., evaporation p., and 2 maximum discharge p.	step-wise fixing of p., local SA	30'000		annual discharge
Cullmann et al. (2006)	WaSiM-ETH	6 soil module related parameters	OAT local SA	13		peak discharge

Table 3.1: Recent sensitivtiy analysis studies in surface hydrology and water quality modelling

[a] first order partial variance
[b] QUality Evaluation and Simulation TOol for River systems
[c] multi-methods global sensitivity analysis (Sobol, K-L Entropy, Morris)
[d] Sobol, Kullback–Leibler entropy, Morris, RSA and regression method
[e] and 2 simple model structures for testing
[f] Sensitivity analysis based on regional splits and regression trees
[g] tracer aided catchment model, distributed
[h] cumulative and peak discharge, average base flow, average groundwater elevation, average soil moisture

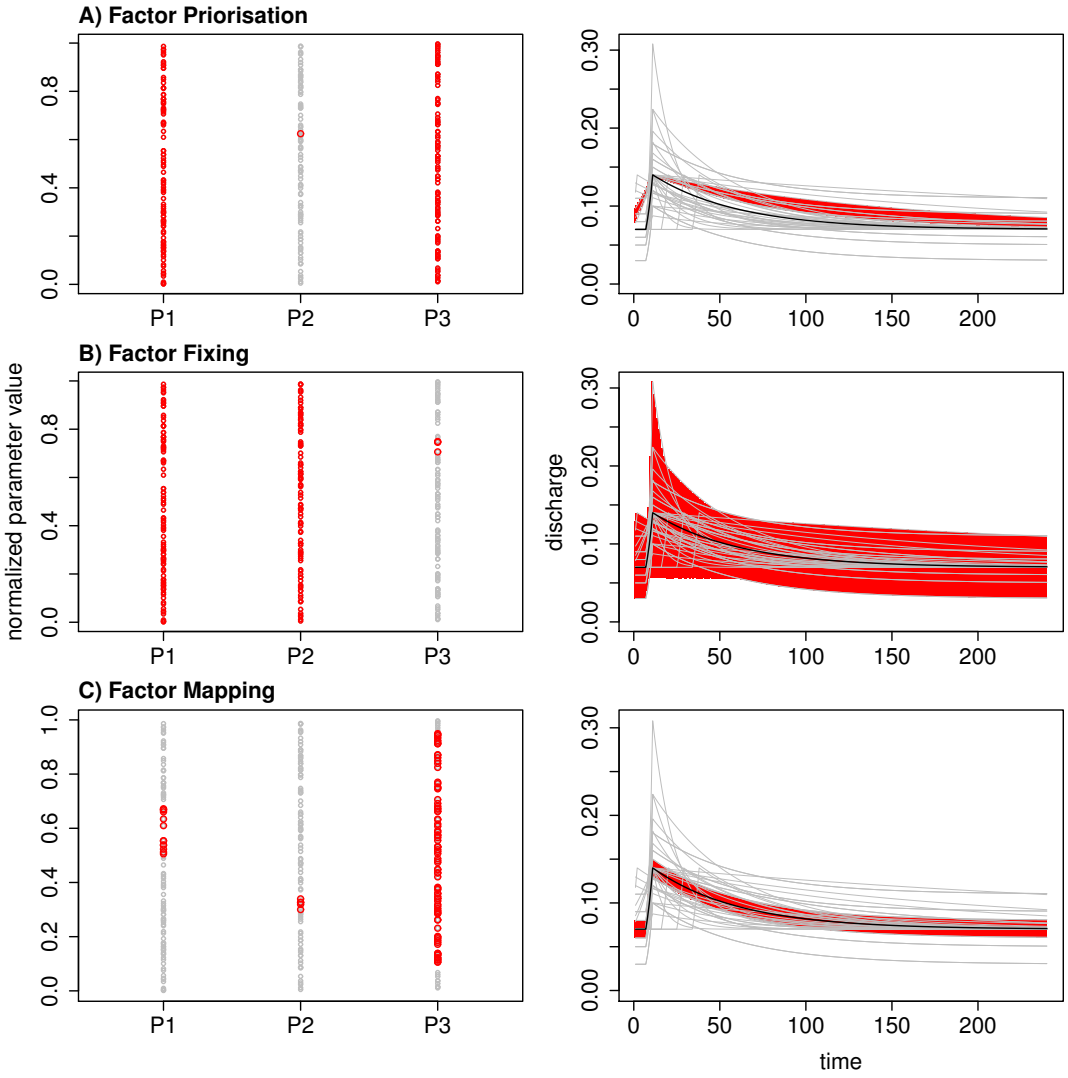


Figure 3.1: Illustration of the different purposes of sensitivity analysis (SA) (Saltelli et al., 2006): A) factor prioritisation investigates the most influential parameter, B) factor fixing investigates the least influential parameter, and C) factor mapping is related to calibration and GLUE like procedures. Normalized parameter ranges are shown on the left hand side together with possible selections for parameter ranges. The right hand side shows measured discharge (black), possible simulation runs (light grey), and remaining variability (red) to visualize SA settings.

The right hand side shows the distribution of the model output for the sampled parameter space in light grey. The red shaded area shows the greatly reduced variance after fixing P2. The black line shows the observations. Factor prioritisation is used to identify the relevant model components for a certain time step or to identify periods with high information content for the calibration of these parameters.

2) A second use of SA is the identification of parameters that can be fixed at any value in their possible range without significantly reducing the output variance. The remaining parameters explain the variance (factors fixing setting). This is the appropriate question if we want to exclude irrelevant parameters from the model calibration and set their values to an arbitrary one. Figure 3.1B) shows that P3 may be set to any value in the full range, without reducing the variance in the model output significantly. The variance in the model output covers the full range without depending on whether P3 is selected from one of the ranges around 0.1 or 0.3 or 0.7.

3) A third use of SA concerns the selection of parameters values to use for accurate prediction with adequate uncertainty bounds (factors mapping setting). This is a typical use for the determination of parameter identifiability and GLUE-like procedures, when behavioral parameter sets are to be identified. Figure 3.1C) illustrates that confining the model output to the observation will strongly reduce the range of the sensitive parameter P2, somewhat reduce the range of P1 and hardly reduce the range of the insensitive P3.

As "best practice" to determine the dominant model components Saltelli et al. (2006) suggested to use measures based on first order partial variance (FOPV). As stated above, partial variance based methods belong to the global sensitivity analysis (SA) methods, which determine the parameter sensitivity for an entire region in parameter space with distributed sampling techniques in parameter space.

3.1.3 Advantages of FAST

There are several methods to compute FOPV sensitivities. FAST was originally developed for the analysis of chemical reaction systems, providing a computationally efficient way to compute FOPV (Schaibly and Shuler, 1973; Cukier et al., 1973, 1975) As for all partial variance based methods, FAST is able to reliably estimate sensitivities of parameters also for non-linear models and is therefore well suited for hydrological models. Multiple sensitivity measures are reported to give contradicting results for the same application (Tang et al., 2007b; Cloke et al., 2008). (This is not very surprising, if for example local and global sensitivities, regional SA and variance based methods are compared). In contrast, results for FOPV appear to show better comparability. For example, Saltelli and Bolado (1998) demonstrated the equivalence of sensitivities computed using Sobol's method with FAST. Saltelli and Bolado (1998) concluded that FAST is computationally much more efficient, requiring for example 150 runs to determine reliably the sensitivity of 6 parameters. This may help to overcome problems with high computational expenses when calculating TEDPAS such as those of Sieber and Uhlenbrook (2005). However, FAST has some limitations which make the method unsuitable for certain types of problems. Results from FAST are not accurate for discrete parameter values (Saltelli et al., 2000; Frey and Patil, 2002). Also parameter interactions can not be detected by the FAST method. The focus of this study is on factor prioritisation and therefore, FOPV may be appropriate.

As argued, FOPV SA is an appropriate method to analyse TEDPAS in order to quantify which model components dominate the simulation as an indicator for dominant processes in the catchment. However, for complex hydrological models we need a highly efficient method to estimate sensitivities. Therefore this paper aims at (1) implementing the computationally efficient original FAST method; (2) investigating whether FAST is

applicable to rainfall-runoff models; (3) comparing FAST to existing implementations of (e)FAST and Sobol's method using a lumped (computationally inexpensive) hydrological model; (4) applying FAST to a computationally expensive hydrological model (WaSiM-ETH); and (5) showing how the resulting sensitivities are a useful diagnostic tool by performing a qualitative model validation and we will identify interactions among parameters in WaSiM-ETH which result in problems during model calibration.

3.2 Methods and Study Area

3.2.1 Fourier Amplitude Sensitivity Test

As with other global SA methods, for FAST, parameters are varied according to the physically possible parameter space, which usually is determined with pretests including expert knowledge, model documentation or model runs. FAST allows to perform a global SA with a small number of runs. All SA methods have in common that the parameters are modified between the model runs $j = 1, 2, 3, \cdots, N$ (which we denote as the model run dimension). FAST is based on the fact that the model output in this model run dimension can be expanded into a Fourier series. The coefficients of the Fourier series can then be used to estimate the mean expected model outcome as well as the variance. If individual parameters are varied with specific frequencies, the corresponding Fourier coefficients allow estimation of the partial variance or model parameter sensitivity. In other words, the underlying idea is to label the parameters by using different frequencies to modify the parameters between the model runs $j = 1, 2, 3, \cdots, N$. This constitutes *the first step* of the method (subsection 3.2.1.1): generating the Fourier parameter set for the sensitivity test. This is exemplified in Figure 3.2 with the labeled arrows indicating the analysis steps. Parameter a and b in the example are "labeled" with frequencies $\omega = 3$ and 7 respectively in the model run dimension ($j = 1, 2, 3, \cdots, N$) and varied according to a uniform distribution in the range $-0.5 \cdots 0.5$. Note that the order of the parameter sets along the model run dimension j needs to be maintained for the evaluation method to work.

The model is then evaluated for each of the parameter sets (*the second step*). Figure 3.2 shows the evaluation for the very simple models depending on a time dependent, weighted average of parameter a and b. To retrieve the information from the frequency labels, i.e. to analyse the sensitivity of the model output for the different parameters, the model output is Fourier transformed in the model run dimension $j = 1, 2, 3, \cdots, N$ (*the third step* (subsection 3.2.1.2). Note that in our simple example, we are repeating the FAST analysis separately for the model output at a fixed time. The fraction of the variance in the model run dimension that can be explained by a certain parameter is proportional to the power in the Fourier spectrum for the corresponding frequency and its multiples (see Cukier et al., 1978, for further details). In the example (Figure 3.2) we observe Fourier coefficients above 0.02 only for frequencies 3 and 9 for $t = 0$, frequencies 3, 7, 9, and 21 for $t = 25$ and frequencies 7 and 21 for $t = 50$. Thus, the variance at $t = 0$ can be fully explained by parameter a (frequency 3 and its multiples). Corresponding statements are possible for $t = 25$ and $t = 50$.

The three steps are described in more details in the subsequent sections. The method is available both as part of SimLab and as software package (Reusser, 2008) for the open source data analysis language R (R Development Core Team, 2008).

3.2.1.1 Generation of the parameter set and model evaluation

As stated above, when generating the Fourier parameter set, we want to modify each parameter with a different frequency among the model runs in order to assign a "label". Generation of the parameter set can be subdivided into selection of appropriate frequencies ω_i, the generation of a value set

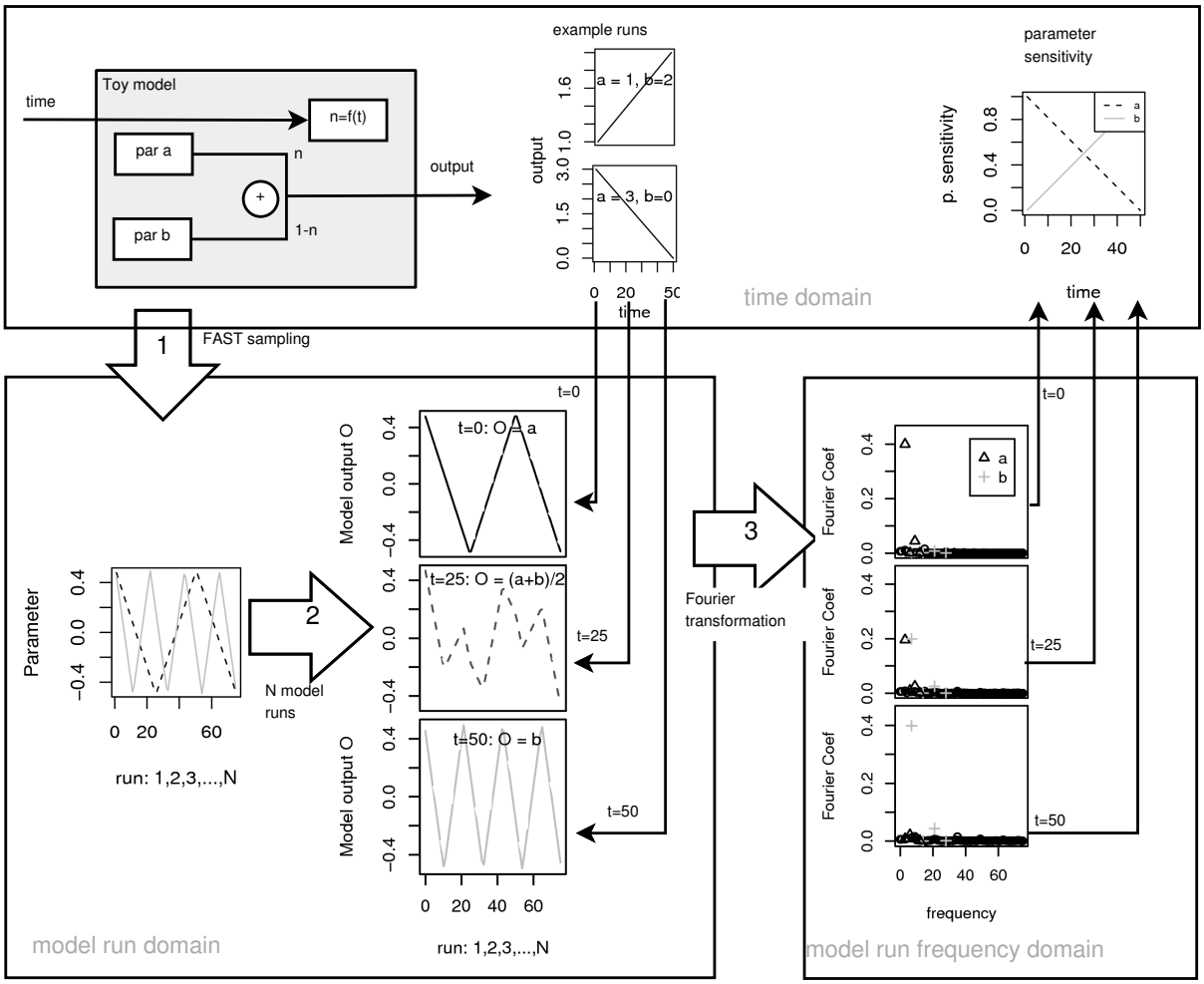


Figure 3.2: Illustration of the Fourier amplitude sensitivity test (FAST) with a simple toy model: $o = n * a + (1 - n) * b$ with $n = 1 - t/50$ for $t = 0 \cdots 50$. In the first step (outlined arrow), parameters are sampled according to a predefined sampling scheme for multiple model runs (model run domain). The second step includes running the model for each time step. Due to the special sampling design, parameter sensitivities can be calculated with Fourier transformation in the third step (model run frequency domain). Applying FAST for each time step allows calculation of TEDPAS. See text for more details.

S with uniform distribution between -0.5 and 0.5 and a transformation of S into the actual Fourier parameters θ.

Cukier et al. (1975) present a table with suggested frequencies ω_i that are mutually independent (for further details see Cukier et al., 1975; McRae et al., 1982) and have mutually independent multiples up to order four. The higher the order, the smaller the error of the numerical approximation of the FAST method (for further details see Cukier et al., 1975). With higher number of parameters, the number of required model runs (also presented in Cukier et al., 1975) increases in order to assure independence of frequencies. Therefore, the selected frequencies $\omega(i)$ and the required sampling size N ($j = 1, ..., N$) depend on the number of parameters n ($i = 1, ..., n$). The selected frequencies together with the model run index are then converted into a supporting variable $S(j, i)$ (equation 3.2), which varies at the appropriate frequency in the range $-0.5 \cdots 0.5$. Calculation of $S(j, i)$ was initially proposed as an exponential function (Cukier et al., 1973), which has the disadvantage of resulting in a distribution that over emphasizes low and high values (Saltelli et al., 1999). Saltelli et al. (1999) proposed to use the function shown in equation 3.2 which results in a uniform distribution of $S(j, i)$. The final transformation of $S(j, i)$ to the actual Fourier parameters $\theta(j, i)$ has undergone some development since the publication of the original method. The transformation based on the cumulative density function $F(\theta)$ of the parameter as shown in equation 3.3 was presented by Fang et al. (2003). Compared to the original method (Cukier et al., 1978) this method has an advantage if non-uniformly distributed parameters are used. We used uniform distributions for all parameters with ranges as shown in table 3.3. Note that in this case, the transformation procedure of (Fang et al., 2003) does not provide an improvement of the method compared to the original method of Cukier et al. (1978).

$$S(j,i) = arcsin(sin(\omega_i * \pi/N * (2j - N - 1)/2))/\pi \tag{3.2}$$

$j = 1....N \; i = 1..n$

$$\theta(j,i) = F_i^{-1}(S(j,i) + 0.5) \tag{3.3}$$

F_i^{-1} being the inverse of the cumulative density function for parameter i

3.2.1.2 Analysis of parameter sensitivity

For each model time step t, a model output series $M = y(j, t)$ was transformed with fast Fourier transformation resulting in a power spectrum. The variance σ_i^2 that could be explained by a certain parameter i was calculated from the sum of the power in the spectrum for the frequencies $\omega(i), 2\omega(i), 3\omega(i), 4\omega(i)$ (see Cukier et al., 1975, for further details on why to use higher order frequencies to the order of 4). Whereas the total variance σ^2 was calculated as the sum of the power spectrum over all frequencies. The sensitivity of model output y on parameter i is then calculated as the partial variance, which is the ratio σ_i^2 / σ^2.

3.2.2 eFAST

We used the implementation of eFAST from the software package SimLab (Saltelli et al., 1999; Saltelli and Bolado, 1998). In eFAST, total order sensitivity can also be determined. This sacrifices the efficiency of FAST to obtain simultaneously the FOPV with a limited number of runs. While the basic idea to "label" parameters with a certain frequency in the model run dimension remains, some modifications to the algorithm need to be introduces in order to asses total order sensitivity. The reader is referred to Saltelli et al. (1999); Saltelli and Bolado (1998) for further details.

3.2.3 Sobol's method

For Sobol's method, a special sampling scheme is applied as well. For a given sampling size N and n parameters, a sub sample size N_s is calculated as $N_s = N/2n + 2$. Parameters θ_i are then sampled randomly for two sub sample sets M_1 and M_2, each consisting of N_s independent parameter sets. Variances are then estimated by evaluating the model for parameter sets, where one parameter in M_1 is replaced by the corresponding parameter of M_2, thereby assessing the effect of changing this single parameter. For further details, see Sobol (2001); Saltelli (2002)

3.2.4 Study regions

3.2.4.1 Huagrahuma catchment, Ecuador

The Huagrahuma catchment is located in the south Ecuadorian Andes, as part of the Paute river basin (Fig. 3.3). The geology consists of Cretaceous and early Tertiary lavas and andesitic volcanoclastic deposits, shaped and compacted by glacier activity during the last ice age (Hungerbühler et al., 2002). The hydraulic conductivity of the bedrock is low, particularly compared to the hydraulic conductivity of the thin layer of volcanic ashes that constitute the soil layer (Buytaert et al., 2005). On average, the soil layer is about 80 cm thick, with some bedrock outcroppings at convex locations and hilltops (Buytaert et al., 2006b). No deep aquifers are present, and water flow is restricted to overland flow and subsurface flow in the soil layer above the bedrock. The vegetation of the Huagrahuma site consists of noetropical alpine grasses and shrubs and some low statured cloud forest. The climate regime is bimodal, with a average annual precipitation of around 1300 mm y^{-1} but a very low seasonality. Precipitation is characterised by frequent low intensity events (drizzle), resulting in around 75% of wet days throughout the year.

3.2.4.2 Weisseritz catchment

The catchment of the Wilde Weisseritz upstream of the gauging station Ammelsdorf (49.3 km^2) served as a second case study. The catchment is situated in the eastern Ore Mountains at the Czech-German border (Fig. 3.3) and has an elevation of 530 m to about 900 m a.s.l. Slopes are gentle with an average of 7°, 99% are $<20°$; calculated from a 90 m digital elevation model (SRTM, 2002). Soils are mostly shallow cambisols of 1 to 2 m thickness. Land use is dominated by forests (≈30%) and agriculture (≈50%). The climate is moderate with mean temperatures of 11°C and 1°C for the periods April - September and October - March, respectively. Annual precipitation for this catchment is 1120 mm/year for the two years of the simulation period from 1 June 2000 until 1 June 2002. During winter, the catchment usually has a snow cover of up to about 1 m for 1 to 4 months with high flows during the snow melt period (Fig. 3.6 shows the pronounced peaks during spring). High flows can also be induced by convective events during summer. WASY (2006) conclude from their analysis based on topography, soil types and land use that subsurface stormflow is likely to be the dominant process. Meteorological data including precipitation, temperature, wind speed, humidity, and global radiation for 11 surrounding climate stations was obtained from the German Weather Service (DWD, 2007). Discharge data, as well as data about land use and soil was obtained from the state office for environment and geology (LfUG, 2007).

3.2.5 Hydrological models

3.2.5.1 Topmodel

The hydrological model TOPMODEL is used in this study (Beven and Kirby, 1979). TOPMODEL is a frequently used model, based on simple physical approximations, and is well documented in the literature (for an overview see Beven et al., 1995; Beven, 1997, 2001). It has been applied to a wide range of catchments, including regionalisation studies (e.g., Ibbitt et al., 2000; Bastola

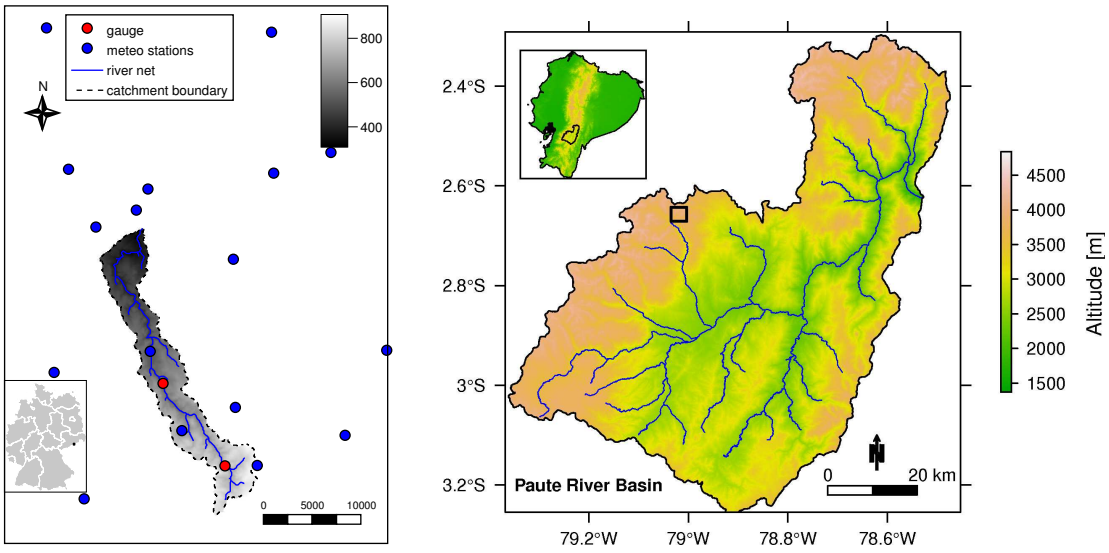


Figure 3.3: Maps of the Wilde Weisseritz catchment (left, scales in m), and the Huagrahuma catchment within the Paute basin, Ecuador

et al., 2008). The choice of TOPMODEL as a good model structure for the hydrology of the páramo ecosystem is based on extensive field experience (Buytaert et al., 2006a; Buytaert and Beven, 2009). The steep topography induces large spatial differences in soil moisture and tendency for the generation of overland flow, which are captured by the topographic index. Additionally, the absence of a dry season, and the marked drop of soil hydraulic conductivity in non-saturated conditions result in continuously wet soils (>60 vol% Buytaert et al., 2005). Field research has shown that also in dry periods a saturated soil layer exists above the bedrock, even on steep slopes (Buytaert et al., 2005). This suggests that the variation in the contributing area is minimal, and that the entire catchment contributes to base flow most of the time.

Finally, the high porosity and low bulk density (typically below 0.6g/cm^3 Buytaert et al., 2006b) give rise to easily compressible soils. Bulk density tends to rise and hydraulic conductivity tends to fall with depth (Buytaert et al., 2006b), giving support to the use of a nonlinear transmissivity profile. The TOPMODEL assumption of an exponential function of the storage deficit appears to give a good representation of the recession curves in these catchments.

The model has 7 parameters and 2 initialisation values (table 3.2). $qs0$ and $Sr0$ initialise respectively the initial subsurface flow per unit area and the initial root zone storage deficit. The surface hydraulic conductivity ($k0$) and the capillary drive (CD) are only used in the infiltration excess routine. The maximum root zone storage deficit ($Srmax$) is part of the root zone equations, the unsaturated zone time delay (td) controls the flow from unsaturated to saturated zone, while the areal average of the transmissivity ($lnTe$, log transformed), and the rate of decline of transmissivity with increasing storage deficit (m) are related to the saturated subsurface flow. vr is the river flow velocity.

name	symbol	min	max	FAST frequency
Initial subsurface flow	qs0	1e-05	6e-05	19
Soil transmissivity (log transformed)	lnTe	-7e-01	-4e-01	59
Shape of the transmissivity curve	m	1e-02	4e-02	91
Initial root zone storage deficit	Sr0	1e-03	4e-02	113
Maximum root zone storage deficit	Srmax	1e-01	1e+00	133
Unsaturated zone time delay	td	-3e+00	1e+00	143
Channel flow velocity	vr	8e+02	2e+03	149
Surface hydraulic conductivity	k0	1e-03	1e-02	157
Capillary drive	CD	1e-01	1e+00	161

Table 3.2: Parameter ranges for Topmodel

3.2.5.2 WaSiM-ETH

WaSiM-ETH is a modular, distributed model (Schulla and Jasper, 2001) and was used for the Weisseritz catchment with a regularly spaced grid of 100 m resolution. The model provides methods for the interpolation of meteorological input data. For each cell, a surface runoff storage and an interflow storage are parametrized with the corresponding linear recession constants and a maximum storage size for the interflow storage (see Table 3.3). The precipitation intensity limit defines a threshold, above which macro pore flow is active and rainfall enters the lower soil storage directly. Interception (leaf area index dependend simple bucket), evapotranspiration (Penman-Monteith) and snow (temperature-index-approach) are also included as modules. Four parameters of the snow module were investigated more closely. Snow accumulation is determined by the snow/rain temperature limit. The temperature melt index defines the amount of snow melted for each degree and hour the temperature is above the snow melt limiting temperature (third parameter). Finally, the fraction of snow melt which builds surface runoff is the fourth parameter. The unsaturated zone is described for each sub basin based on the Topmodel approach (Beven and Kirby, 1979). The Topmodel regionalization parameter m determines how strong the gradient in the saturation deficit is due to differences in the topographic index. m also enters the equations for the vertical flow q_v (Eq 3.5) and the baseflow Q_B (Eq 3.4). Vertical flow and baseflow are both calibrated with the scaling factors T_{korr} and K_{korr}. Channel flow is routed with a simple storage to account for diffusion.

$$Q_B = T_{korr} * e^{-\gamma} * e^{-S_m/m} \tag{3.4}$$

$$q_v = K_{korr} * k_f * e^{-S_i/m} \tag{3.5}$$

γ is the mean value of the topograpic index, a constant for a given basin, k_f the saturated hydaulic conductivity, S_m and S_i the mean and local saturation deficit for a subbasin, model state variables.

WaSiM-ETH was set up and run 487 times, the number of required runs (see section 3.2.1.1) for sensitivity analysis with 11 varying parameters. The set of resulting discharge time series $y(j,t)$, one for each of the N parameter sets was then further analysed to calculate sensitivities.

3.3 Results

3.3.1 Comparison of Sensitivity analysis methods with Topmodel

Figure 3.4 shows sensitivities calculated with the following SA methods: a) Sobol in SimLab 3.2.6

Parameter name	Process	Symbol	Range	FAST frequency
temperature limit for snow melt	T_{m0}	snow melt	$-2 \cdots 2$	41
difference between snow/rain temperature limit and temperature limit for snow melt (the first is always higher)	snow accumulation	$T_{\mathrm{R/S}}$	$0 \cdots 2$	67
temperature melt index	snow melt	C_0	$0.7 \cdots 2$	105
fraction on snow melt which is surface runoff	snow melt	c_{melt}	$0.2 \cdots 0.5$	145
Topmodel regionalization parameter	baseflow	m	$0.005 \cdots 0.04$	177
scaling factor for transmissivities	baseflow	T_{korr}	$0.005 \cdots 0.4$	199
scaling factor for vertical flow	baseflow	K_{korr}	$800 \cdots 8000$	219
recession constant for surface runoff single linear storage	surface runoff	k_D	$1 \cdots 120$	229
maximum content of the interflow storage	interflow	SH_{max}	$1 \cdots 150$	235
recession constant for interflow runoff single linear storage	interflow	k_H	$50 \cdots 300$	243
precipitation intensity limit	fast infiltration	P_{limit}	$0.2 \cdots 20$	247

Table 3.3: Parameters of the model WaSiM-ETH used for the SA

(n=5632), b) eFAST in SimLab 2.2 (n=5000) and c) FAST (SimLab 3.2.6, n=1289) d) FAST (R-package (Reusser, 2008), n=487). The number of model runs for eFAST and Sobol was selected as a balance between the reduction of numerical artefacts (e.g. first order sensitivities outside the possible range from 0 to 1, random fluctuations of TEDPAS) and computation time, while the minimum requirements as suggested by the implementation were used for FAST. The 487 runs for the method in the R-package are reported by Cukier et al. (1975, 1978) (with the corresponding frequencies listed in table 3.2), while the minimum requirement of 1289 simulation runs for SimLab 3.2.6 is undocumented.

For each method (a-d) two graphs show the sensitivity of the modelled discharge for different parameters, grouped according to sensitivity. The first graph shows parameters $qs0$, $Sr0$, m, and vr. For all 3 methods, the TEDPAS is equivalent for these four variables: initial conditions ($qs0$ and $Sr0$) are dominant until mid of the simulation, thereafter m and vr of highest importance. Looking at the 4 parameters $qs0$, $Sr0$, m, and vr, 90% of the time, the same parameter dominates in all of the methods (rank 1 in sensitivities). The second graph shows the remaining 5 variables. With Sobol's method, the modelled discharge depends on td after the initial period, while with eFAST $lnTe$ has some influence on modelled discharge for June and July. With FAST, the output shows only minor sensitivity for the two parameters.

As explained in the method section, the sensitivity is reported as partial variance that can be explained by this parameter at this time step. For example, a value of around 0.7 for parameter m during June indicates that 70% of the observed variation between the model runs $j = 1, \cdots, N$ can be explained by this parameter. The sum over all parameter sensitivities never exceeds 1.0 but may be lower because of the numeric approximation (Cukier et al., 1975) or when parameter interactions are of importance (non-additive models – Saltelli et al., 2006). The fourth graph (e) shows the 25 best (selected according to RMSE) modelled discharge time series in black and the mea-

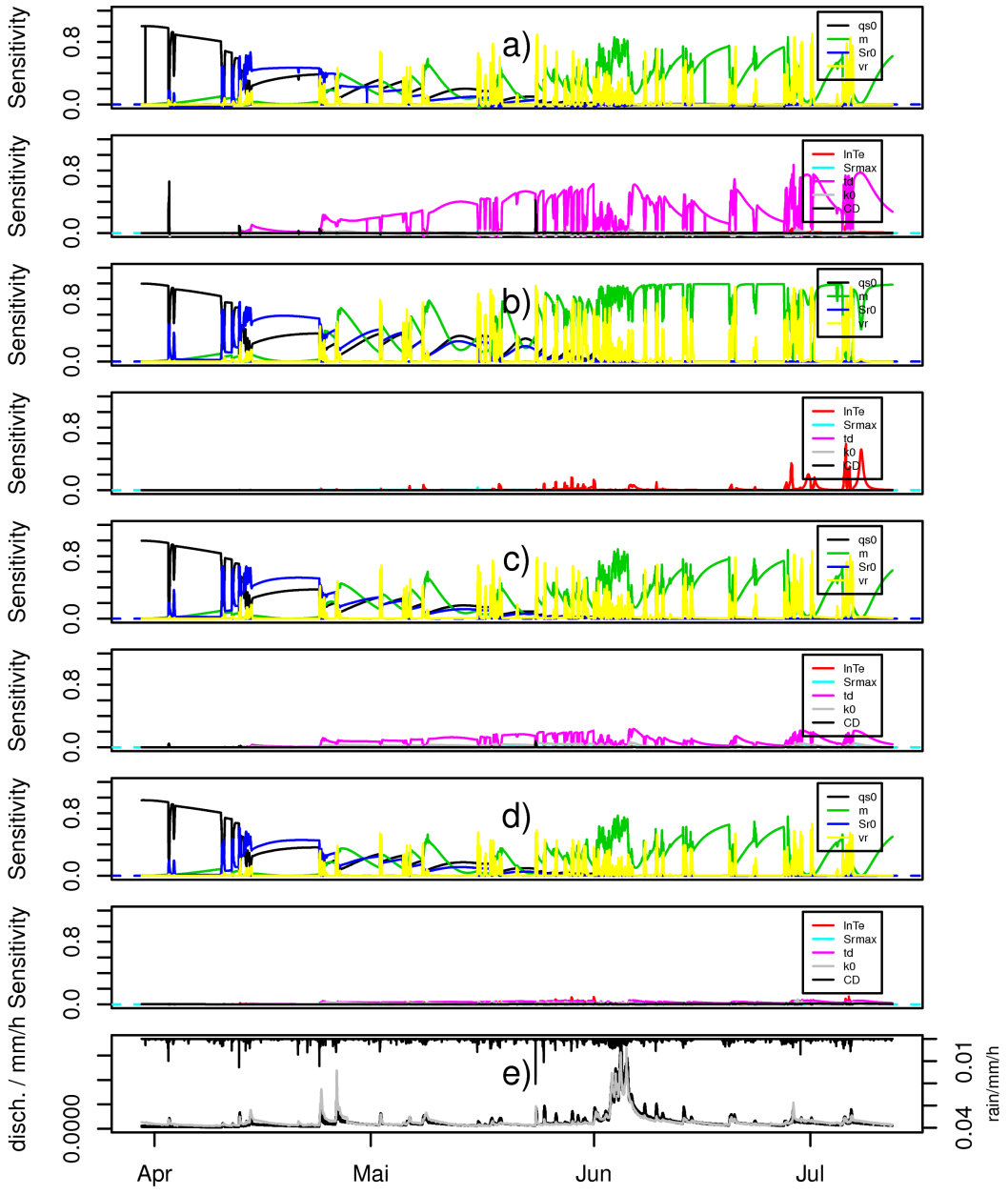


Figure 3.4: Parameter sensitivities of Topmodel, with a) Sobol's method (SimLab 3.2.6, n=5632), b) eFAST (SimLab 2.2, n=5000), c) FAST (SimLab 3.2.6, n=1289) and d) FAST (R-package, n=487). Panel e) shows discharge for the 25 best simulation runs and the observation

sured time series in gray.

3.3.2 FAST WaSiM-ETH

Sensitivities for the computationally expensive WaSiM-ETH model were calculated with the FAST-method only (FAST frequencies are listed in table 3.3). For TEDPAS at the event time scale we will present two examples, which are shown in Fig 3.5. The two examples consist of 4 four graphs (a-d) each. The three top graphs (a-c) show the sensitivity of the modelled discharge for different parameters, grouped according to the different model components. The first graph (a) shows the snow melt related parameters. The three saturation deficit related parameters m, T_{korr} and K_{korr} are shown in the second graph (b). The third graph (c) shows the remaining parameters k_D, k_H, $SH_{\max}$, and P_{limit}. The fourth graph (d) shows the 25 best (selected according to RMSE) modelled discharge time series in black and the measured time series in grey.

The first example is in February 2001. Simulated discharge is strongly dependent on the snow melt temperature limit T_{m0} during the entire winter (see also Fig 3.6). At the beginning of the event, the modelled discharge shows some sensitivity for the shift of the snow/rain temperature limit $T_{\mathrm{R/S}}$ and the temperature melt index C_0. The discharge also shows some sensitivity towards the direct flow recession constant k_D. When approaching the end of February, sensitivity of the discharge decreases for k_D and C_0 and increases for the interflow recession constant k_H.

The second example is in July 2001. At the beginning of the event, the modelled discharge shows increased sensitivity for the direct flow recession constant k_D and the precipitation intensity limit P_{lim}. In the following period, the sensitivity of the discharge mainly depends on the interflow recession constant k_H and shows slight sensitivity towards the interflow reservoir size $SH_{\max}$. At the end of the event simulated discharge is sensitive for the three saturation deficit related parameters.

Figure 3.6 shows TEDPAS of the modelled discharge on the annual time scale. We observe that snow related parameters are important during winter and spring as expected. Also, saturation deficit related parameters are unimportant for discharge during snow melt periods. Note that the plots only show first order effects. Influence of interacting parameters are not visible from these plots. Therefore, in order to exclude the influence of a parameter, higher order terms (or total order sensitivity) need to be calculated. TEDPAS of the three saturation deficit related parameters is highly correlated. This suggests a strong interaction of the these parameters, as will be further discussed in section 3.4.

3.4 Discussion

3.4.1 Comparing Sensitivity Methods for Topmodel

3.4.1.1 Performance

All 3 SA methods result in very similar TEDPAS for four important variables. Differences exist for two parameters: the model appears to be more sensitive for $lnTE$ in July with eFAST, sensitivity is high for td with Sobol's method and with FAST, the model shows only minor sensitivity for the two parameters. The reasons for this difference are unclear, but may be related to the different sampling schemes or the differing methods for calculating the sensitivities.

3.4.1.2 Computational requirements

The time required for the analyses differ substantially. Using FAST (Reusser, 2008) in R, the analysis was finished within less than one hour, while approximately 8 hours were necessary to produce the results with the eFAST algorithm implemented in SimLab 2.2. Results for Sobol's method (SimLab 2.2) also required about 8 hours, however these results were discarded because of

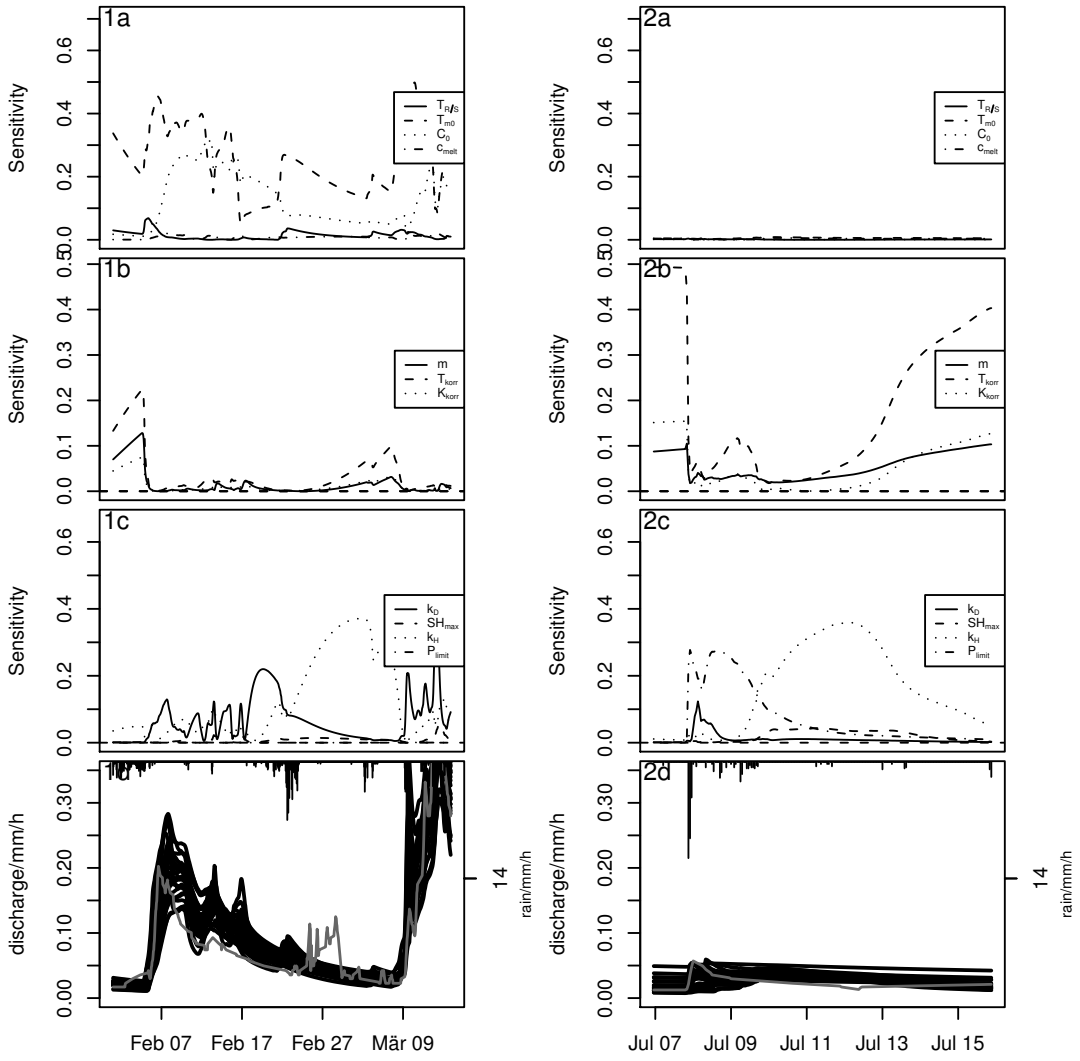


Figure 3.5: TEDPAS for periods in 2001. Parts (a-c) show the parameter sensitivity of the modelled discharge (a: snow model related parameters; b: saturation deficit related parameters; c: remaining parameters k_D, k_H, $SH_{\max}$, and P_{limit}). The sensitivity is reported as partial variance that can be explained by the corresponding parameter. The fourth graph (d) shows the 25 best modelled discharge time series in black and the measured time series in grey.

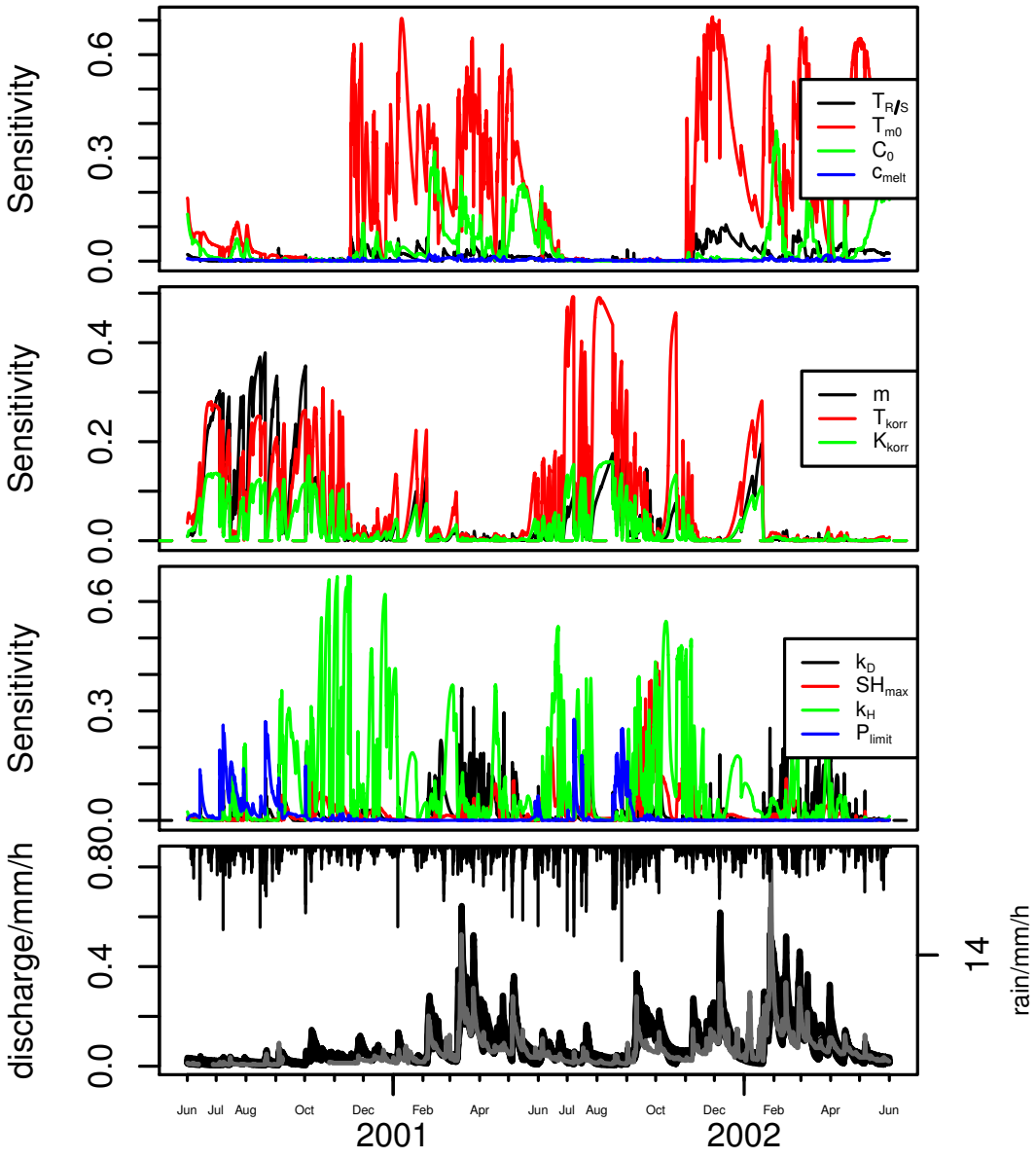


Figure 3.6: As Figure 3.5 for the entire simulation period.

implausibilities: Total order sensitivities were alway smaller than FOPV and FOPV and Total order sensitivities added up to more than one for all variables. This is by definition impossible for first order sensitivities. However, due to the closed nature of the SimLab computer application, it is not possible to investigate this issue in more detail. With SimLab 3.2.6, the analysis with Sobol's method and the original FAST method required about 50 and 12 hours, respectively.

In summary, consistent results for FAST, eFAST and Sobol's method were obtained for Topmodel, which agrees with previous reports about the equivalence of results for the two methods (Saltelli and Bolado, 1998). Based on this comparison, we may expect results with our implementation of the FAST algorithm (Reusser, 2008) to be similar in quality to existing implementations, however with about 1/10th of the model runs required compared to other FOPV SA methods.

3.4.2 TEDPAS of Topmodel

The results of the temporal sensitivity analysis for TOPMODEL on the Huagrahuma catchment are very much as expected. Although the effect of the initialisation values decreases over time, several thousands of time steps are required for the effect to die out. This highlights the importance of a warm-up period when using the model in prediction mode. Model parameter m, which defines the shape of the transmissivity curve, is known to be a sensitive parameter (e.g., Buytaert and Beven, 2009), with an effect over the entire recession curve. The channel velocity parameter vr has a major effect on the time to peak flow, and to a lesser extent on the shape of the steep part of the recession curve. The observation that the sensitivy of vr shows a peaky behaviour, with high sensitivy related to precipitation events is therefore physically plausible.

Other parameters, particularly $Srmax$ and td are known to be relatively insensitive in the ecosystem. Evapotranspiration is nearly independent of soil moisture in the continuously wet grasslands. The organic soils also tend to have a very high soil moisture, accelerating the flow from the unsaturated to the saturated zone (Buytaert et al., 2006a; Buytaert and Beven, 2009). Therefore, the relatively high sensitivity of td in the Sobol method is not very clear, but may be related to interaction between td and m. It is indeed possible that an artificially high time delay between the unsaturated and saturated zone compensates for too quick saturated flow, which is controlled by both $lnTe$ and m. The peaks in the sensitivity of CD are related to the very rare occurrence of infiltration excess overland flow in the study catchment. Hence, this model routine is nearly always inactive, apart from a few occurrences of intense precipitation events.

Finally, it is interesting to note that the sensitivity of the parameters does not change during the major high-flow event in early June. This suggests that the similar model structures are operational during this period as during the dry periods and that the parsimonious model structure performs well for a relatively wide range of hydrological conditions.

3.4.3 TEDPAS of WaSiM-ETH

Checking the yearly patterns in TEDPAS is a first assessment to verify the model structure. A first pattern on the annual scale is – as expected – the model showing high sensitivities for snow related parameters during winter and spring. However, simulated discharge shows some sensitivity for snow related parameters approximately until end of June, although in reality snow cover was completely depleted by the end of May. We, thus, suggest to revise the parameter range of $C0$ used for sensitivity analysis for this catchment for subsequent analyses.

A second pattern on the annual scale is the first order insensitivity of discharge for saturation deficit related parameters during snow melt periods. This is plausible since discharge is (by definition) dominated by melt water during these peri-

ods.

Plausibility checks are also possible on the event time scale by checking the sequence of parameter sensitivity and comparing it to expectations based on model design as suggested by Sieber and Uhlenbrook (2005). The two examples presented in Section 3.3.2 are compliant with our expectations. The expectations for the snow melt event are: at the beginning of the event, discharge depends on whether precipitation occurs as snow or rain (snow/rain temperature limit) and the amount of snow melting (temperature melt index). Because a part of the melt water forms overland flow, we expect the discharge to be sensitive for the direct flow recession constant.

For the summer event we expect the following chronology of relevant parameters: first direct flow recession constant k_D and the precipitation intensity limit $P_{\lim}$ followed by interflow related parameters and finally saturation deficit related parameter which determine base flow.

Two additional benefits: First, TEDPAS is also a valuable tool for calibration of model parameters. The fraction of melt water contributing to overland flow c_{melt} will hardly ever be well identifiable, because the sensitivity of simulated discharge for this parameter is always smaller than 2% despite the large range for c_{melt} of $20 \cdots 50\%$. Note that this result needs to be confirmed with calculation of higher order sensitivities in order to definintely exclude any influcence of c_{melt}.

As stated before, sensitivity is a necessary but not sufficient conditon for identifiablity: parameters may show high sensitivity but be poorly identifiable if compensatory effects between two parameters make them interdependent. Such compensatory effects of parameters may be detected (second benefit), indicated by a highly correlated sensitivity of the model output for multiple parameters. Correlated model parameters can be a major source for poor identifiability in hydrological modelling (Bárdossy, 2007). In the case study, we observe correlated parameters for the saturation deficit related parameters, which will complicate proper identification of these parameters during calibration. Brun et al. (2001) demonstrate how to derive a set of identifiable parameters from such a set of correlated parameters.

3.5 Conclusions

We demonstrated that SA can provide valuable information for improved model understanding, which goes beyond the most often selected approach to date, where the most influential parameters for calibration are determined.

For our case study we found that TEDPAS is consistent with expectations on both the annual and event time scale. In addition, SA may enhance calibration because time periods of high parameter sensitivity are the relevant periods for calibration. Based on the understanding of the importance of parameters, a priori assumptions may be revised and field experiments may be guided. Finally, the method allows to detect compensatory effects of parameters, which we found for the saturation deficit related parameters of WaSiM-ETH.

To make full use of SA-methods, we need highly efficent methods. We applied such a highly efficient SA method called FAST to two rainfall-runoff models. FAST allows to determine global sensitivity for parameters with only a limited number of model runs. The current case study required around 150 and 490 runs for resp. 6 and 11 parameters. Our analysis of parameter sensitivities for WaSiM-ETH would not have been possible without this very efficient sampling scheme. In addition, based on the efficient calculation of sensitivities from model output variables, entire time series of parameter sensitivity can be calculated (e.g. for discharge).

That extended SA presented here is only a first step in obtaining better model understanding. The ultimate goal is to gain insight into model behavior by answering the three research questions: (1) during which periods is the model reproducing observed quantities and dynamics; (2) what is the na-

ture of the error in times of poor model performance; and (3) which components of the model are causing this error. The first two research questions may be answered using the TIGER method (Chapter 2), while the third question is answered using SA methods as presented here. This approach closely relates to the framework for diagnostic model evaluation proposed by Gupta et al. (2008).

3.6 Acknowledgments

The study has been funded as part of OPAQUE (operational discharge and flooding predictions in head catchments), a project within the BMBF-Förderaktivität "Risikomanagement extremer Hochwasserereignisse" (RIMAX). WB was funded by a Marie Curie Intra-European Fellowship at the University of Lancaster during the implementation of Topmodel for the Huagrahuma Catchment. We would like to thank Jenny Eckart for her support with the data preprocessing for WaSiM-ETH. A major part of the analysis was carried out with the open source statistical software R and contributed packages, we would like to thank its community.

Chapter 4

Inferring model structural deficits by analyzing temporal dynamics of model performance and parameter sensitivity *

In this paper we investigate the use of hydrological models as learning tools to help improve our understanding of the hydrological functioning of a catchment. With the model as a hypothetical conceptualisation of how dominant hydrological processes contribute to catchment scale response, we investigate three questions: 1) during which periods does the model (not) reproduce observed quantities and dynamics. 2) what is the nature of the error during times of bad model performance, and 3) which model components are responsible for this error.

To investigate these questions we combine a method for detecting repeating patterns of typical differences between model and observations (TIGER) with a method for identifying the active model components during each simulation time step based on parameter sensitivity (TEDPAS). The approach generates a time series of occurrence of dominant error types and time series of parameter sensitivities. A synoptic discussion of these time series highlights deficiencies in the assumptions about the functioning of the catchment.

The approach is demonstrated for the Weisseritz headwater catchment in the eastern Ore Mountains. Our results indicate that the WaSiM-ETH complex grid based model is not a sufficient working hypothesis for the functioning of the Weisseritz catchment, and point towards future steps that can help improve our understanding of the catchment.

*Dominik Reusser, Erwin Zehe (in press), *WRR*

4.1 Introduction

The major goal of hydrological research is to learn from the past to understand and predict the future. In the standard approach to this learning process, conceptual models are used to predict future behaviors of the investigated catchment, for instance in the long-term context of climate or land use change impacts (Niehoff et al., 2002; Niehoff and Bronstert, 2001; Kleinn et al., 2005, 2003).

Models can, however, also help to shed light on our incomplete understanding of how hydrological processes translate into catchment response in different landscapes. By assuming that the model is the best conceptualisation of our understanding of the relevant processes and their transformation into catchment response, we may learn from periods in which the model is found to perform poorly. Such an approach, that uses models as learning tools, can help us to improve our models in a much more targeted way, and to better identify and possibly reduce predictive uncertainty.

The core idea of our model assessment approach is to evaluate the interdependence between patterns of poor model performance and patterns of dominant model components. The model consists of functional components that represent different hydrological processes and their interactions, and the conditions that lead to the different processes of the hydrological cycle represent the hydrological context. For example, the hydrology of the catchment may be a) dominated by either mass input or energy input. b) thresholds may alter the functioning of the catchment; for example snow influenced periods occur when temperatures drop below snow melt temperature, or hydrophobicity dominated reactions occur when catchment wetness drops below a certain threshold (Blume et al., 2009, 2008b,a). The catchment may spend most of its time c) either close to or far from equilibria. Depending on the relevant hydrological context, different components of the model will dominate the simulation of catchment dynamics. While some model components may represent the catchment response well, others are likely to be deficient. Thus we can expect to see repeating patterns of poor model performance, these being related to the relevance/dominance of model components during those periods.

Existing approaches to identify errors in the model structure and the resulting predictive uncertainties are generally based on various methods of uncertainty analysis. Often used in hydrology are GLUE based approaches (Beven and Binley, 1992) and more formal Bayesian approaches to uncertainty estimation (Thiemann et al., 2001; Gupta, 2003). For example, the multi-period model conditioning approach (Choi and Beven, 2007) analyses the temporal dynamics of parameter uncertainty. In dynamic identifiability analysis (Wagener et al., 2003) non-stationarites in the optimal range for a certain parameter are detected. An alternative approach was recently presented by Reichert and Mieleitner (2009) where stochastic, time-dependent parameters are used to identify model components with the potential to reduce model uncertainty. The temporal dynamics of model structure uncertainties have been analysed by Clark et al. (2008), who used 79 models from a model family for their study. Bayesian total error analysis provides the possibility to simultaneously assess the uncertainty from various sources (Kuczera et al., 2006). None of the existing approaches explicitly evaluates the interdependence between patterns of poor model performance and patterns of dominant model components. In addition these approaches require a large number of realizations, which may not be feasible for models with a more complex process representations, as computing time becomes quickly the bottle neck. This calls for other methods that require less model runs, which we will present here.

We are able to reduce the number of model runs with our approach because: a) it is not necessary to calibrate the model in advance, b) a highly efficient method is used to sample the parameter space, and c) all model runs are evaluated (to determine parameter sensitivity) while other Monte Carlo based

methods often discard the 90% worst runs as a first step.

Our key idea is that the model structural deficiencies can be better identified and understood when first analysing patterns of poor model performance and patterns of dominant model components independently during model assessment and subsequently combining the information. In doing so it is important to recognize that poor model performance might be caused either by data errors or by deficiencies of certain model components. Accumulated state variable errors caused by data errors and/or past model structural errors are probably the most challenging cause of poor model performance. Error contaminated input data can force sensitive parameters towards wrong values to compensate for poor performance during the calibration phase. Using a Bayesian framework to address input uncertainty (Kavetski et al., 2006b,a) or simply excluding poor input data from the calibration process can help to minimize this problem. In the case that certain model components are deficient, one can expect that simulated discharge will exhibit higher sensitivity to parameters belonging to the model component with inadequate representation of the system. If so, the process conceptualization associated with that model component should be revised. This approach provides a much more targeted process for improving the model with respect to dominant processes and for reducing specific errors.

The main innovations of this work are this strategy for targeted improvement of the model (by disentangling the temporal dynamics of model performance and parameter sensitivity), which has the ability to provide model diagnostic analysis with a limited number of model runs.

The core idea of our approach can be condensed into three interlinked research questions:

1) during which periods is or is not the model reproducing observed quantities and dynamics?

2) what is the nature of the error in times of poor model performance?

3) which components of the model are causing this error?

A methodology to address the first two questions was presented by chapter 2. Their TIGER (TIme series of Grouped ERrors) method uses a combination of a) a large selection of performance measures to characterize different error types, b) synthetic peak errors to support error type characterization and c) analysis of the time series occurrence of error types with respect to observed and modeled flow dynamics. To investigate the third research question, we combine TIGER with a method for analyzing the temporal dynamics of parameter sensitivities (TEDPAS) introduced in a closely related study (Chapter 3). We provide an overview of the two methods (TIGER and TEDPAS) and a brief description of the model and study catchment in section 4.2. Results for the case study are presented (Section 4.3) and discussed (Section 4.4). The study closes with conclusions and suggestions for future work in Section 4.5.

4.2 Methods and Study Area

4.2.1 Weisseritz catchment

The catchment of the Wilde Weisseritz upstream of gauging station Ammelsdorf (49.3 km^2) served as the study area. The catchment is situated in the eastern Ore Mountains at the Czech-German border (Fig. 4.1) and has an elevation range of 530 to about 900 m a.s.l. Slopes are gentle with an average of 7°, 99% are $<20°$; calculated from a 90 m digital elevation model (SRTM, 2002). Soils are mostly cambisols. Land use is dominated by forests (≈30%) and agriculture (≈50%). The climate is moderate with mean temperatures of 11°C and 1°C for the periods April - September and October - March, respectively. Annual precipitation for this catchment is 1120 mm/year for the two years of the simulation period from 1 June 2000 until 1 June 2002. During winter, the catchment usually has a snow cover of up to about 1 m for

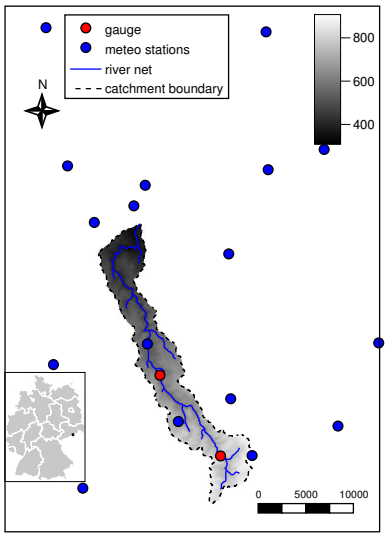


Figure 4.1: Wilde Weisseritz catchment (scales in m).

1 to 4 months with high flows during the snow melt period (Fig. 4.5 (2d) and (4d) shows the pronounced peaks during spring). High flows can also be induced by convective events during summer. WASY (2006) conclude from their analysis based on topography, soil types and land use that subsurface stormflow is likely to be the dominant process. Meteorological data including precipitation, temperature, wind speed, humidity, and global radiation for 11 surrounding climate stations was obtained from the German Weather Service (DWD, 2007). Discharge data, as well as data about land use and soil were obtained from the state office for environment and geology (LfUG, 2007).

4.2.2 Hydrological model WaSiM-ETH

WaSiM-ETH is a modular, distributed model (Schulla and Jasper, 2001) and was used for the Weisseritz catchment with a regularly spaced grid of 100 m resolution. The model provides methods for the interpolation of meteorological input data. For each cell, a surface runoff storage and an interflow storage are parametrized with the corresponding linear recession constants and a maximum storage size for the interflow storage (see Table 4.1). The precipitation intensity limit defines a threshold, above which macro pore flow is active and rainfall enters the lower soil storage directly. Interception (leaf area index dependent simple bucket), evapotranspiration (Penman-Monteith) and snow (temperature-index-approach) are also included as modules. Four parameters of the snow module were investigated more closely. Snow accumulation is determined by the snow/rain temperature limit. The temperature melt index defines the amount of snow melted for each degree and hour the temperature is above the snow melt limiting temperature (third parameter). Finally, the fraction of snow melt which builds surface runoff is the fourth parameter. The unsaturated zone is described for each sub basin based on the Topmodel approach Beven and Kirby (1979). The Topmodel regionalisation parameter m determines how strong the gradient in the saturation deficit is due to differences in the topographic index. m also enters the equations for the vertical flow q_v (Eq 4.2) and the baseflow Q_B (Eq 4.1). Vertical flow and baseflow are both calibrated with the scaling factors T_{korr} and K_{korr}. Channel flow is routed with a simple storage to account for diffusion.

$$Q_B = T_{korr} * e^{-\gamma} * e^{-S_m/m} \tag{4.1}$$

$$q_v = K_{korr} * k_f * e^{-S_i/m} \tag{4.2}$$

γ is the mean value of the topographic index, a

constant for a given basin, k_f is the saturated hydraulic conductivity, S_m and S_i are the mean and local saturation deficit for a sub basin, two model state variables.

4.2.3 Parameter sensitivity (TEDPAS)

An analysis of the temporal dynamics of parameter sensitivity (TEDPAS) of the modelled discharge provides insight into the relevant model components. To calculate the temporal dynamics of parameter sensitivity, a sensitivity analysis is performed repeatedly for each time step (Chapter 3) by: 1) Generating the appropriate sets of model parameters ϕ. 2) Evaluating the model for each parameter set. 3) Processing the model output of interest for the set of all model runs to calculate the parameter sensitivity. Note that by explicitly splitting analysis of parameter sensitivity and model performance, one avoids a potential source of errors. If parameter sensitivity of some performance criterion F were considered, then the sensitivity $S = dF/d\phi$ would depend on two components – the sensitivity $dQ_{\text{sim}}/d\phi$ and the size of the model error ($Q_{\text{obs}} - Q_{\text{sim}}$) represented in F. In this case, the sensitivity would depend on both the process sensitivity and the size of the model error. This problem does not occur with our method as the sensitivity is calculated as $dQ_{\text{sim}}/d\phi$.

We used the Fourier amplitude sensitivity test method (FAST Schaibly and Shuler, 1973; Cukier et al., 1973, 1975; Fang et al., 2003) because of its computational efficiency. Sensitivity analysis for eleven parameters (Table 4.1) required 487 simulation runs. Parameters were sampled in a representative way from the parameter space according to the FAST sampling scheme. The algorithm is freely available as a software package (Reusser, 2008) coded using the open source data analysis language R (R Development Core Team, 2008). In chapter 3, we report that FAST produces the same results as three other methods but with at least eight times less computational burden.

4.2.4 Model performance (TIGER)

The TIGER approach investigates time series of grouped errors to detect repeating patterns of similar poor model performance (Chapter 2). To explain the method we present a simple "toy example" using the time series shown in Figure 4.2. Notice that the simulations (shown as 2 runs with different parameters - black lines) deviate from the "observations" (grey line) in such a way that peaks 1 and 3 are overestimated and peaks 2 and 5 appear too late, while peak 4 is matched exactly.

The essence of the method is to compute a "fingerprint" of error type for a moving time window of length 250 time steps, based on an analysis of several performance measures. For this simple example, the finger print is based on three performance measures – the root mean square error (RMSE), Nash-Sutcliffe coefficient of efficiency (NSCE) and lag time (t_L) (see Figure 4.2a, lower panel). Next, a clustering of fingerprints, based on self-organizing maps (SOM; Reusser et al., 2009; Kohonen, 1995; Haykin, 1999; Kalteh et al., 2008) and fuzzy clustering (Reusser et al., 2009; Bezdek, 1981; Dimitriadou et al., 2008) is performed to identify similar types of model behaviour along the modelled time period. Figure 4.2 shows the cluster membership beneath the discharge time series as bars with varying shading; the saturation of the color bar is proportional to the cluster membership, with full saturation indicating a membership of 1 and white indicating a membership of 0. As can be seen, periods of overestimation (peaks 1 and 3) appear in cluster A, periods of good agreement in cluster B, and periods with a time lag are assigned to cluster C.

To get a better understanding of the nature of each cluster we next examine error types for synthetic hydrographs representing a single flood event (see Reusser et al., 2009, for the mathematical function used to generate synthetic peak errors). By constructing the synthetic hydrographs to be of the length of the time window we can compare the data with the magnitude and duration

Parameter name	Process	Symbol	Range
temperature limit for snow melt	T_{m0}	snow melt	$-2 \cdots 2$
difference between snow/rain temperature limit and temperature limit for snow melt (the first is always higher)	snow accumulation	$T_{\mathrm{R/S}}$	$0 \cdots 2$
temperature melt index	snow melt	C_0	$0.7 \cdots 2$
fraction on snow melt which is surface runoff	snow melt	c_{melt}	$0.2 \cdots 0.5$
Topmodel regionalisation parameter	baseflow	m	$0.005 \cdots 0.04$
scaling factor for transmissivities	baseflow	T_{korr}	$0.005 \cdots 0.4$
scaling factor for vertical flow	baseflow	K_{korr}	$800 \cdots 8000$
recession constant for surface runoff single linear storage	surface runoff	k_D	$1 \cdots 120$
maximum content of the interflow storage	interflow	SH_{max}	$1 \cdots 150$
recession constant for interflow runoff single linear storage	interflow	k_H	$50 \cdots 300$
precipitation intensity limit	fast infiltration	P_{limit}	$0.2 \cdots 20$

Table 4.1: Parameters of the model WaSiM-ETH used for the sensitivity analysis

of a "typical" event. Based on this artificial reference hydrograph we construct the two error types - including under- and overestimation and positive and negative lag times with three levels of deviation (Figure 4.2b). Only clusters A and C appear in this plot, because the presented synthetic peak errors correspond only to these two performance measures clusters . This example illustrates how we interpret clusters A and C as representative of periods of overestimation and time lags, respectively.

Whereas the toy example presented above uses only 3 performance measures, the approach used in our full case study uses 44 additional performance measures (in addition to NSCE, RMSE and t_L) to provide a complete fingerprint of model performance (Chapter 2). In addition, we selected the 25 best model runs from the set of 487 model runs (section 4.2.3) for the analysis (best 5% of model runs) based on the Nash-Sutcliffe coefficient of efficiency NSCE (Nash and Sutcliffe, 1970). Further, we supplement the synthetic hydrograph error types and provide a second aid to the understanding of the nature of each cluster, based on the range of the performance measures for each cluster is visualised with box plots. The error types for synthetic hydrographs are extended to a set of a) peak errors, b) timing errors, c) volume errors or d) recession errors in the simulation as shown in Figure 4.3.

As preparation for understanding the results of the Wilde Weisseritz case study, we introduce here the error types generated using the synthetic hydrographs (Figure 4.3). Cluster A (red) includes the synthetic peak errors closest to the reference peak, and therefore corresponds to periods with the best accordance between models and observation. Cluster B (yellow) includes peaks where the recession period is generally too fast and peaks in the reference do not occur in the synthetic errors. Cluster C (green) includes peaks with overestimated discharge, mainly due to recession periods that are too slow. Cluster D (blue) includes strong underestimation where the discharge time series is shifted below the reference. Cluster E finally includes false peaks and overestimations due to an upwards shift. A summary of cluster characterization (including observed characteristics from per-

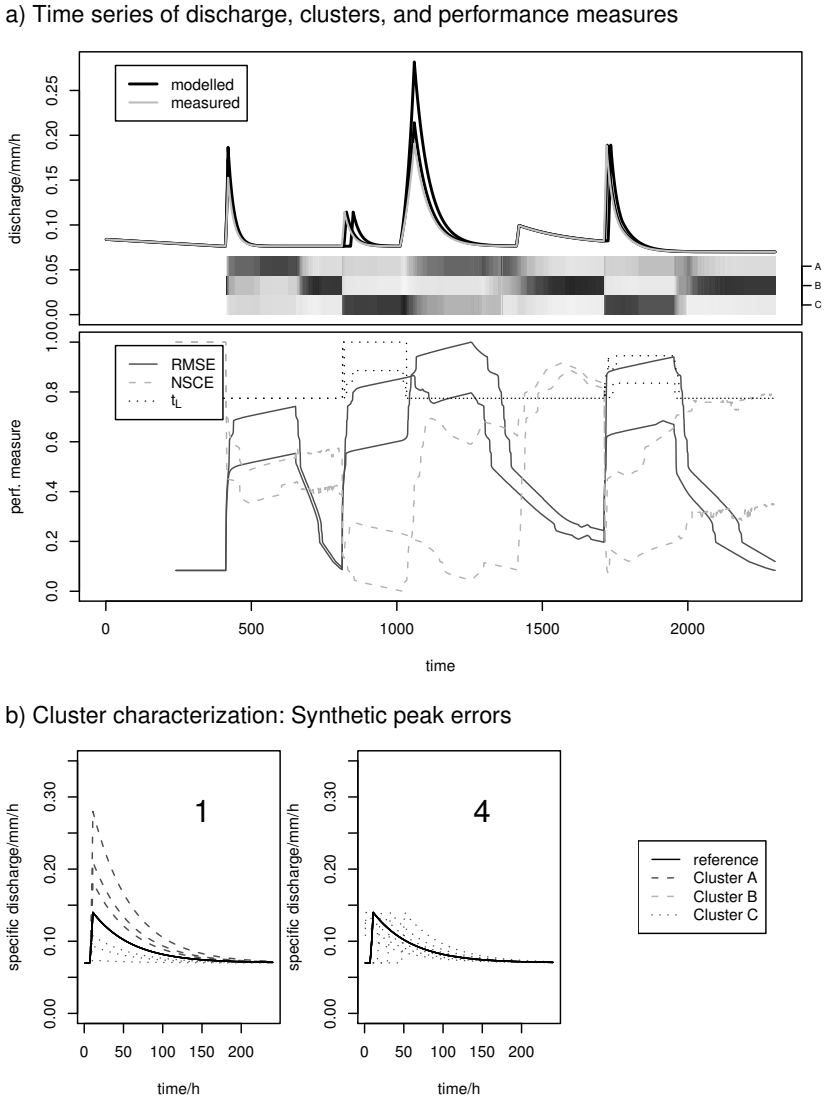


Figure 4.2: Synthetic discharge time series for demonstration of the TIGER method. Part a) shows time series data: observed and 2 simulated discharge time series and cluster memberships (A,B,C) in the first panel. The second panel shows time series for the performance measures root mean square error (RMSE), Nash-Sutcliffe coefficient of efficiency (NSCE) and lag time (t_L). Part b) characterization of clusters with synthetic peak errors (see also Figure 4.3 – only clusters A and C appear in the plot).

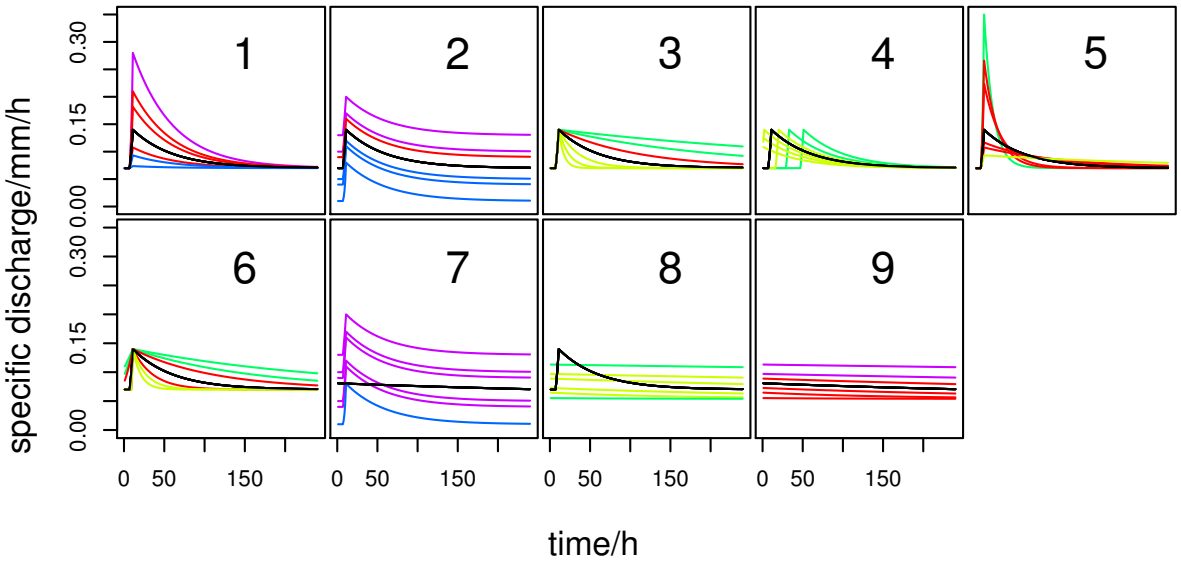


Figure 4.3: Examples of synthetic errors for a single peak event: Peak over- or underestimation **(1)**, baseflow over- or underestimation **(2)**, recession too fast or too slow **(3)**, timing: too late or too early **(4)**, maximum peak flow over- or underestimation but with correct total volume **(5)**, peak too wide (start too early, recession too slow) or too narrow **(6)**, erroneously simulated peak **(7)** or missing peak **(8)**, and over- or underestimation during a late recession phase **(9)**. The colors indicate the cluster with which the synthetic errors are associated (Cluster A: red, B: yellow, C: green, D: blue, and E:purple - see section 4.3.2 for a description of the clusters)

Cluster	Nr.	Year	Period	Parameters	Model component
Cluster A: best fit, includes synthetic peak errors with small level, low flow periods not represented very well	1	2000	Jul.18 - Jul.19		
	2	2000	Aug.22 - Sep.04	m	sat. deficit
	3	2001	Jan.23 - Feb.03	T_{m0}	snow
	4	2001	Jul.08 - Jul.18	T_{korr}, k_H	sat. deficit, interflow
	5	2001	Aug.05 - Aug.21	T_{korr}	sat. deficit
	6	2001	Oct.28 - Oct.29	k_H	interflow
	7	2002	Mar.23 - Mar.30	k_H	interflow
	8	2002	Apr.28 - May.07	T_{m0}	snow
	9	2002	May.15 - May.31	T_{m0}	snow
Cluster B: underestimation (mainly due too fast recession), missing peaks, peaks too early, differences for smaller values but good agreement for peaks	10	2000	Dec.22 - Dec.31	k_H	interflow
	11	2001	Jan.15 - Jan.23	T_{m0}	snow
	12	2001	Feb.06 - Feb.09	T_{m0}	snow
	13	2001	Feb.26 - Mar.09	k_H	interflow
	14	2001	Jun.07 - Jun.12	k_H	interflow
	15	2001	Jul.20 - Jul.21	T_{korr}, k_H	sat. deficit, interflow
	16	2001	Sep.04 - Sep.11	k_H	interflow
	17	2001	Dec.24 - Jan.29	T_{m0}	snow
	18	2002	Apr.18 - Apr.20	T_{m0}	snow
Cluster C: dynamics well reproduced but overestimation (mainly due to too slow recession), peaks too late	19	2001	Feb.09 - Feb.15	T_{m0}	snow
	20	2001	Mar.11 - Apr.08	T_{m0}	snow
	21	2001	May.05 - May.19	T_{m0}	snow
	22	2001	Dec.06 - Dec.23	T_{m0}	snow
	23	2002	Jan.29 - Mar.19	T_{m0}	snow
	24	2002	Mar.31 - Apr.11	T_{m0}	snow
Cluster D: bad reproduction of dynamics, underestimation mainly due to downwards shift of time series	25	2000	Jun.11 - Jun.13		
	26	2000	Jul.06 - Jul.08	m	sat. deficit
	27	2000	Jul.25 - Aug.21	m	sat. deficit
	28	2000	Sep.02 - Sep.02	m	sat. deficit
	29	2000	Nov.02 - Nov.09	k_H	interflow
	30	2000	Nov.21 - Dec.22	T_{m0}	snow
	31	2001	Jan.03 - Jan.08	T_{m0}	snow
	32	2001	Apr.20 - Apr.29	T_{m0}	snow
	33	2001	Jun.15 - Jun.16	k_H	interflow
	34	2001	Nov.01 - Nov.08	k_H	interflow
Cluster E: overestimation due to upwards shift and false peaks, recession periods do not agree well, good agreement after rescaling	35	2000	Dec.07 - Dec.10	k_H	interflow
	36	2001	Jan.10 - Jan.14	T_{m0}	snow
	37	2001	Feb.18 - Feb.19	k_D	direct flow
	38	2001	Apr.15 - Apr.16	k_H	interflow
	39	2001	May.01 - May.03	T_{m0}	snow
	40	2001	Jun.24 - Jul.05	T_{korr}	sat. deficit
	41	2001	Jul.22 - Aug.02	T_{korr}	sat. deficit
	42	2001	Sep.14 - Oct.16	k_H	interflow
	43	2001	Nov.09 - Dec.06	T_{m0}	snow
	44	2002	Feb.26 - Feb.26	T_{m0}	snow

Table 4.2: Time periods with a high cluster membership (>0.7) and the corresponding dominating parameters (sensitivity > 0.2 for at least 40% of the period).

formance measures) is provided in Table 4.2.

4.3 Results

4.3.1 Analysis of parameter sensitivity (TEDPAS)

Two examples of TEDPAS at the event time scale are shown in Fig 4.4. The two examples (left and right columns) each consist of four graphs (a-d). The three top graphs (a-c) show the sensitivity of the modelled discharge for different parameters, grouped by different model components. The first graph (a) shows the snow melt related parameters. The three saturation deficit related parameters m, T_{korr} and K_{korr} are shown in the second graph (b). The third graph (c) shows the remaining parameters k_D, k_H, $SH_{\max}$, and P_{limit}. Sensitivity is reported in terms of the partial variance explained by a parameter at this time step. For example a value of around 0.3 for parameter k_H during July 2000 indicates that 30% of the observed variation between the model runs can be explained by this parameter. The sum over all parameter sensitivities never exceeds 1.0 but may be lower because of the numerical approximation (Cukier et al., 1975) or if parameter interactions are of importance (non-additive models – Saltelli et al., 2006). The fourth graph (d) shows the 25 modelled discharge time series in black and the measured time series in grey.

The first example (left column) is in February 2001. Simulated discharge depends strongly on the snow melt temperature limit T_{m0} (Fig 4.4-1a) during the entire winter (see also Fig 4.5). At the beginning of the event, the modelled discharge shows some sensitivity to shift of the snow/rain temperature limit $T_{\mathrm{R/S}}$ and the temperature melt index C_0 (Fig 4.4-1a). Discharge also shows some sensitivity towards the direct flow recession constant k_D (Fig 4.4-1c). Towards the end of February, sensitivity of the discharge decreases for k_D and C_0 and increases for the interflow recession constant k_H.

The second example (right column) is in July 2001. At the beginning of the event, modelled discharge shows increased sensitivity to the direct flow recession constant k_D and the precipitation intensity limit P_{lim} (Fig 4.4-2c). Subsequently, the sensitivity of discharge mainly depends on the interflow recession constant k_H and shows slight sensitivity towards the interflow reservoir size $SH_{\max}$. At the end of the event simulated discharge is sensitive to the three saturation deficit related parameters (Fig 4.4-2b).

4.3.2 Analysis of model performance (TIGER)

The temporal dynamics of model performance was calculated for the best 25 runs (section 4.2.4). The corresponding 25 sets of model parameters are listed in Table 4.3. Nash-Sutcliffe efficiencies (NSCE) between $0.47 \cdots 0.63$, with an average of 0.54 were observed for the 25 model runs. A closer look reveals that acceptable model performance is observed mainly during late spring / summer (see below, Cluster A) and various differences between models and the observation occur during the remaining periods.

The 25 modelled discharge time series and the measured time series are shown in Figure 4.5 (bottom subplot) in black and grey, respectively. The top three subplots show parameter sensitivities, while the cluster membership for each time step is shown in the fourth subplot using color-coded bars. Cluster A (best accordance) occurs mainly during late spring / early summer. Cluster B (recession too fast) and cluster C (discharge overestimated) occur during snow melt events. Cluster B is also present during summer. Cluster D (strong underestimation) occurs only a few times, mainly during the initial simulation period. Finally, cluster E (false peaks or upwards shift) occurs during times where the model overestimates the observed data throughout the entire period. For more details on the results of the TIGER analysis see (Chapter 2).

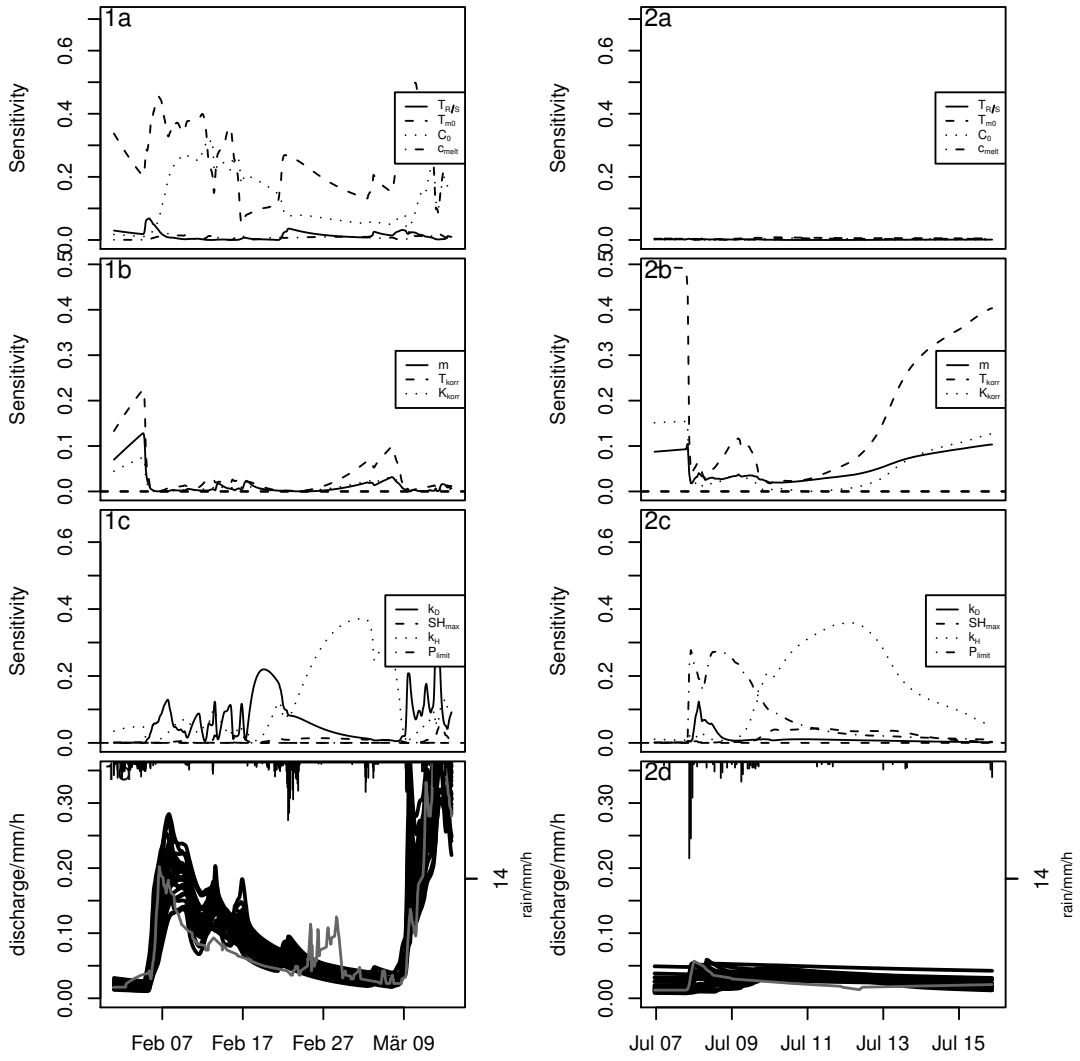


Figure 4.4: Temporal dynamics of parameter sensitivity for periods in 2001. Panels (a-c) show the parameter sensitivity of the modelled discharge (a: snow model related parameters; b: saturation deficit related parameters; c: remaining parameters k_D, k_H, $SH_{\max}$, and P_{limit}). The sensitivity is reported as partial variance that can be explained by the corresponding parameter. The fourth graph (d) shows the 25 modelled discharge time series in black and the measured time series in grey. The two columns refer to different periods.

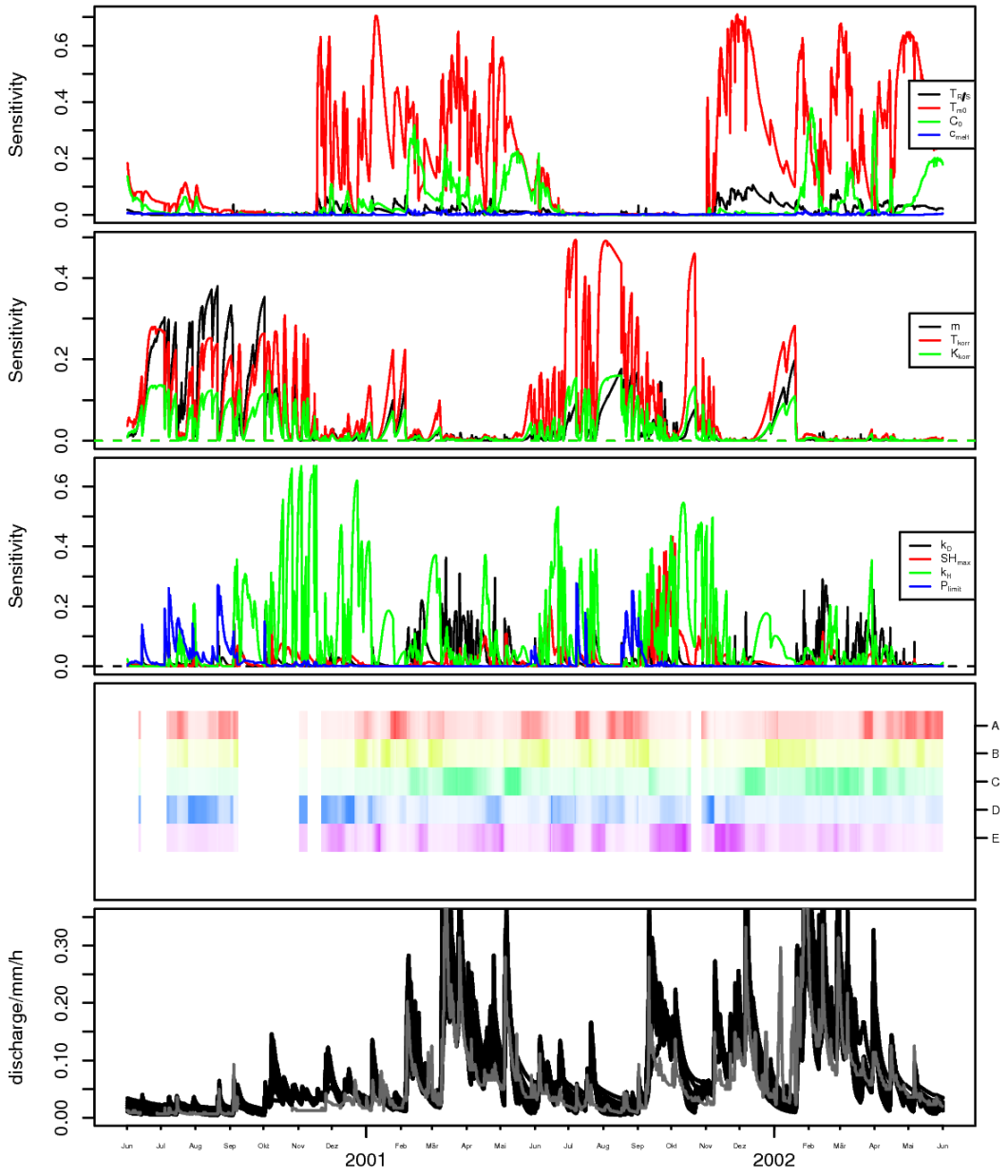


Figure 4.5: As Figure 4.4 for the entire simulation period. The figure also shows the performance cluster membership μ_t as color coded bars (Cluster A: red, B: yellow, C: green, D: blue, and E:purple) where full color saturation corresponds to a membership of 1 and white to a membership of 0.

	m	T_{korr}	K_{korr}	k_D	$SH_{\max}$	k_H	$T_{\mathrm{R/S}}$	T_{m0}	C_0	c_{melt}	P_{limit}
	0.018	0.16	7800	80	55	260	2	-1.5	1.8	0.43	14
	0.03	0.072	2000	64	140	170	1.6	-2	1.9	0.43	14
	0.015	0.076	3800	13	83	210	0.44	-0.56	1.9	0.43	12
	0.01	0.21	4400	49	110	200	0.93	-2	1.6	0.28	1.5
	0.022	0.38	3400	94	110	290	0.53	-1.5	1.7	0.28	0.93
	0.04	0.33	6800	35	57	220	0.49	-1.9	1.1	0.41	16
	0.019	0.0083	5300	93	130	180	1.1	-1.7	1.5	0.39	7.4
	0.023	0.092	4900	77	59	240	0.77	-1.3	1.8	0.47	4.8
	0.034	0.012	7100	72	87	220	0.84	-0.84	1.5	0.47	7
	0.037	0.051	7300	36	32	120	0.053	-1.3	1.9	0.32	17
	0.031	0.27	1100	110	110	210	0.35	-1.8	2	0.32	18
	0.019	0.31	6700	11	100	300	0.75	-1.8	1.9	0.32	18
	0.021	0.046	1900	26	140	270	1.8	-1.7	1.6	0.36	3.3
	0.029	0.23	4500	40	89	240	0.41	-1.7	1	0.35	12
	0.022	0.064	7400	9.6	130	200	0.21	-1.4	1.1	0.35	12
	0.018	0.33	3800	24	120	230	1.6	-1.5	1.6	0.36	3.5
	0.03	0.25	6000	74	100	280	2	-2	1.5	0.37	3
	0.032	0.2	2600	39	64	250	0.95	-1.5	1.9	0.33	19
	0.02	0.027	5200	59	150	260	0.55	-2	2	0.32	18
	0.0082	0.24	4600	84	73	160	0.14	-1.5	2	0.32	17
	0.028	0.38	7000	29	100	200	0.56	-1.5	1.8	0.47	5
	0.02	0.3	1200	45	92	230	0.87	-2	1.4	0.39	7.1
	0.031	0.053	5100	31	120	96	0.021	-1.7	1.4	0.26	11
	0.017	0.14	2300	46	68	270	0.33	-1.3	1.7	0.28	0.65
	0.029	0.09	7500	98	150	240	0.73	-1.7	1.6	0.28	1.2
min	0.0082	0.0083	1100	9.6	32	96	0.021	-2	1	0.26	0.65
max	0.04	0.38	7800	110	150	300	2	-0.56	2	0.47	19
prior min	0.005	0.005	800	1	1	50	0	-2	0.7	0.2	0.2
prior max	0.04	0.4	8000	120	150	300	2	2	2	0.5	20

Table 4.3: Parameter values for the best TEDPAS runs. Min and max define the range covered by the best TEDPAS runs, while prior min and prior max correspond to the initial range from table 4.1

4.3.3 Combined analysis of model performance and parameter sensitivity

Table 4.2 lists the periods during which cluster membership > 0.7 occurs along with the relevant parameters. The latter are defined as those for which the modelled discharge has a partial sensitivity > 0.2 for at least 40% of the period. In other words, parameters are considered relevant if they explain at least 20% of the discharge variance for at least 40% of the period. While these thresholds (20% and 40%) are subjectively chosen, they provide an objective basis for analysis and make it easier to reproduce results than by direct visual inspection of Figure 4.5.

Cluster A is not further discussed here because these correspond to periods having reasonable accordance between simulation and observation. From Table 4.2 we see that if cluster B (recession too fast) occurs, the relevant parameter during summer and fall is normally k_H while the relevant parameter during snow melt periods is T_{m0}. Error cluster C (discharge overestimated) always coincides with T_{m0} as relevant parameter. Periods that correspond to cluster D (strong underestimation) have m as the relevant parameter during the initial phase of the simulation, while k_H and T_{m0} become relevant later. For cluster E (false peaks or upwards shift) 4 different parameters are important at different times.

This general pattern can be better understood by also looking at Figures 4.6-4.8, which show details from Figure 4.5:

1) **Example A** (Figure 4.6): during both winter periods, the simulated discharges do not respond in concert with observed events (cluster B (recession too fast) – period nr. 11, 13, 17). Discharge is seen to be sensitive to the temperature limit for snow T_{m0}. However, the temperature melt index c_{melt} has no influence on simulated discharge, indicating that no snow melt is occurring during these periods. Interestingly, discharge is sensitive to the Topmodel parameters m, T_{korr}, and K_{korr}, which would typically not be the case during snow melt periods. (a possible explanation will be presented in the discussion).

2) **Example B** (Figure 4.7): for cluster C (discharge overestimated) we see that the patterns of snow melt period dynamics are reproduced fairly well but there is an upwards shift that results in overestimation. During all these periods The most important parameter is T_{m0} - the snow melt temperature, but discharge is also sensitive to the snow melt index $C0$ and the direct flow recession constant k_D.

3) **Example C** (Figure 4.8): For periods in cluster E (false peaks or upwards shift), the recession in the model is too rapid compared to the observed discharge. In terms of sensitivity, we observe 4 different main patterns where discharge depends on a) K_h and T_{m0} (period nr. 35, 38, 43, 44), b) T_{m0} only (period nr. 36, 39), c) k_D, $C0$, and T_{m0} (Period 37), d) k_h, T_{korr}, and $SH_{\max}$ (period 40, 41, 42).

4.4 Discussion

4.4.1 Parameter sensitivity (TEDPAS)

The two examples presented in Section 4.3.1 are compliant with our expectations regarding parameter sensitivity, that at the beginning of the snowmelt event, discharge depends on whether precipitation occurs as snow or rain (snow/rain temperature limit) and on the amount of snow melting (temperature melt index). Because a part of the melt water forms overland flow, we also expect the discharge to be sensitive to the direct flow recession constant. Similarly, during the summer event we expect the following chronology of relevant parameters: first direct flow recession constant k_D and the precipitation intensity limit P_{lim} followed by interflow related parameters and finally the saturation deficit related parameter which determine base flow.

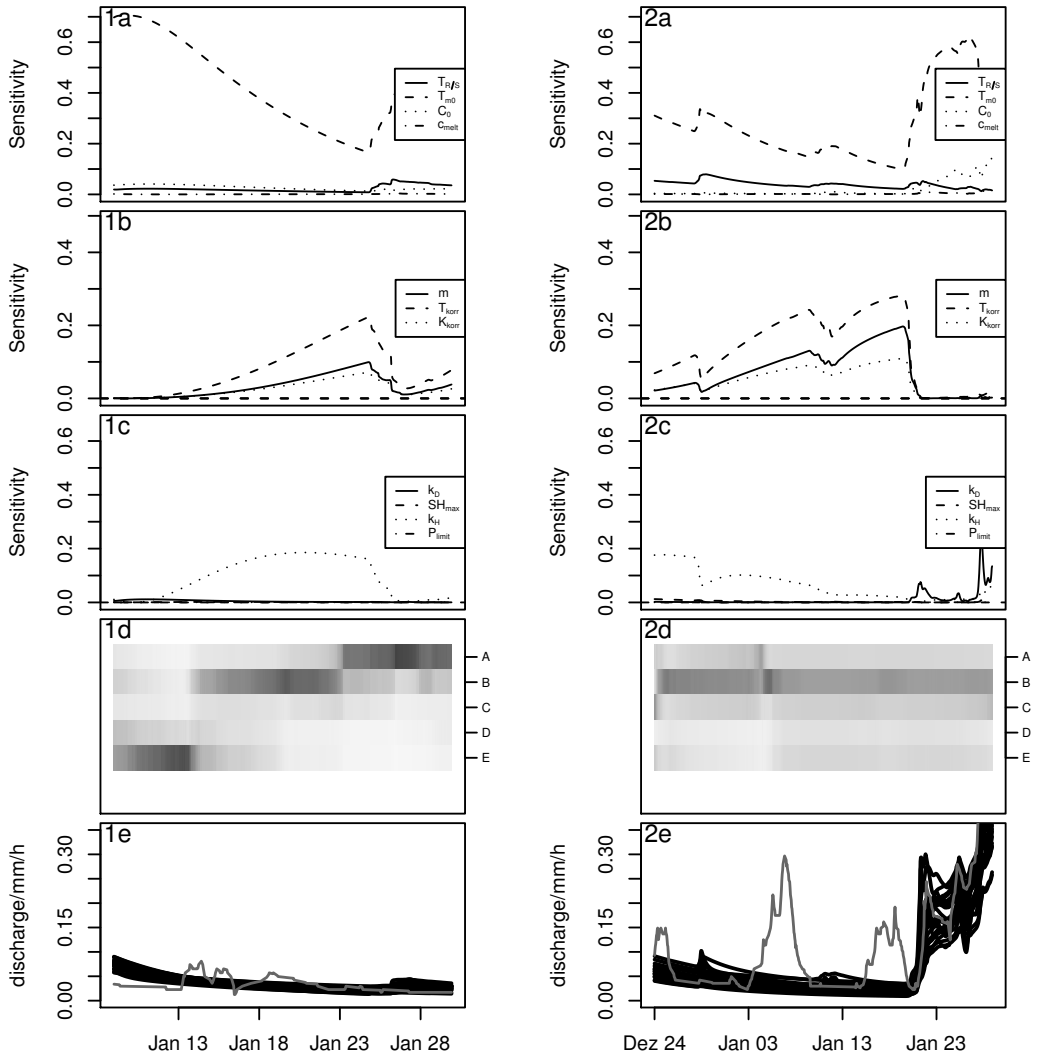


Figure 4.6: Details from Figure 4.5 for January 2001 (left plot) and January 2002 (right plot).

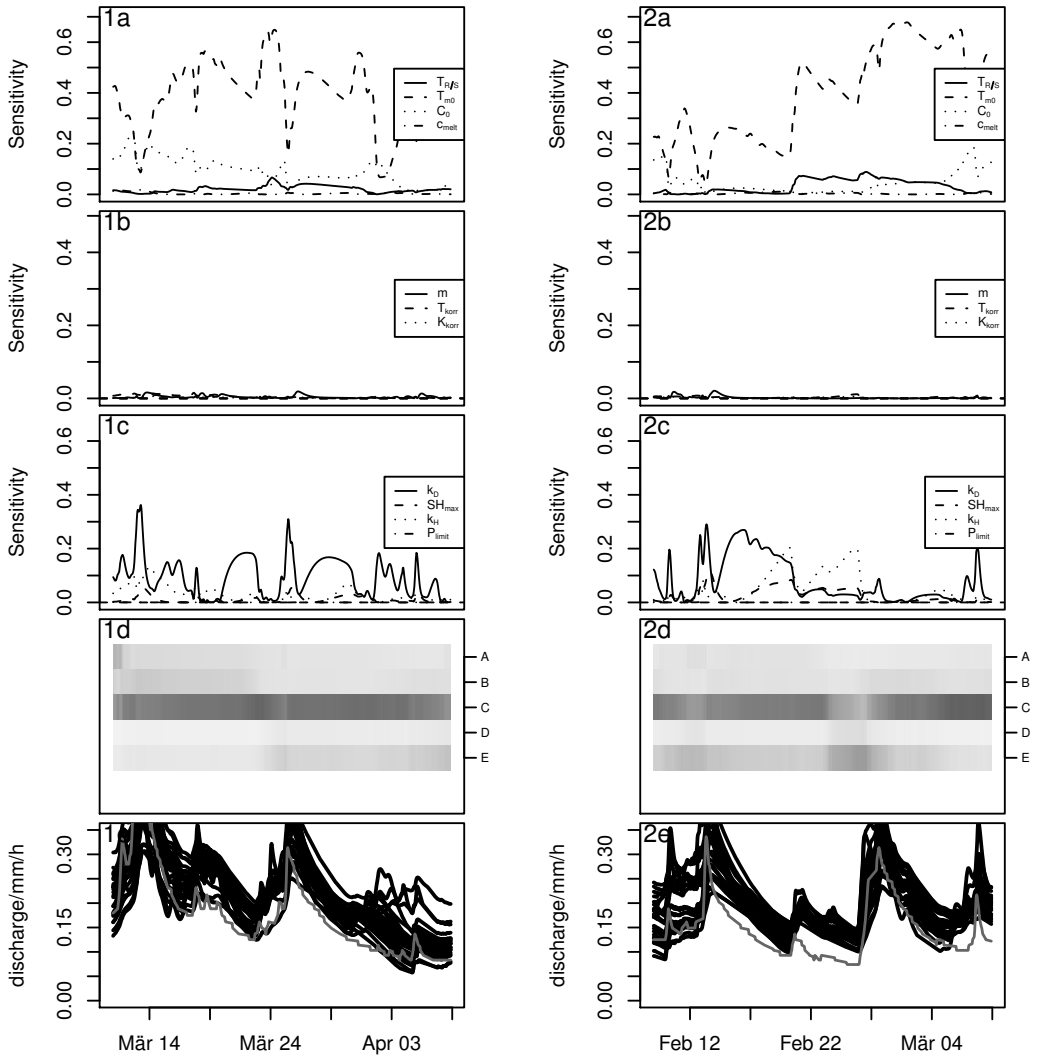


Figure 4.7: Details from Figure 4.5 for March 2001 (left plot) and February 2002 (right plot).

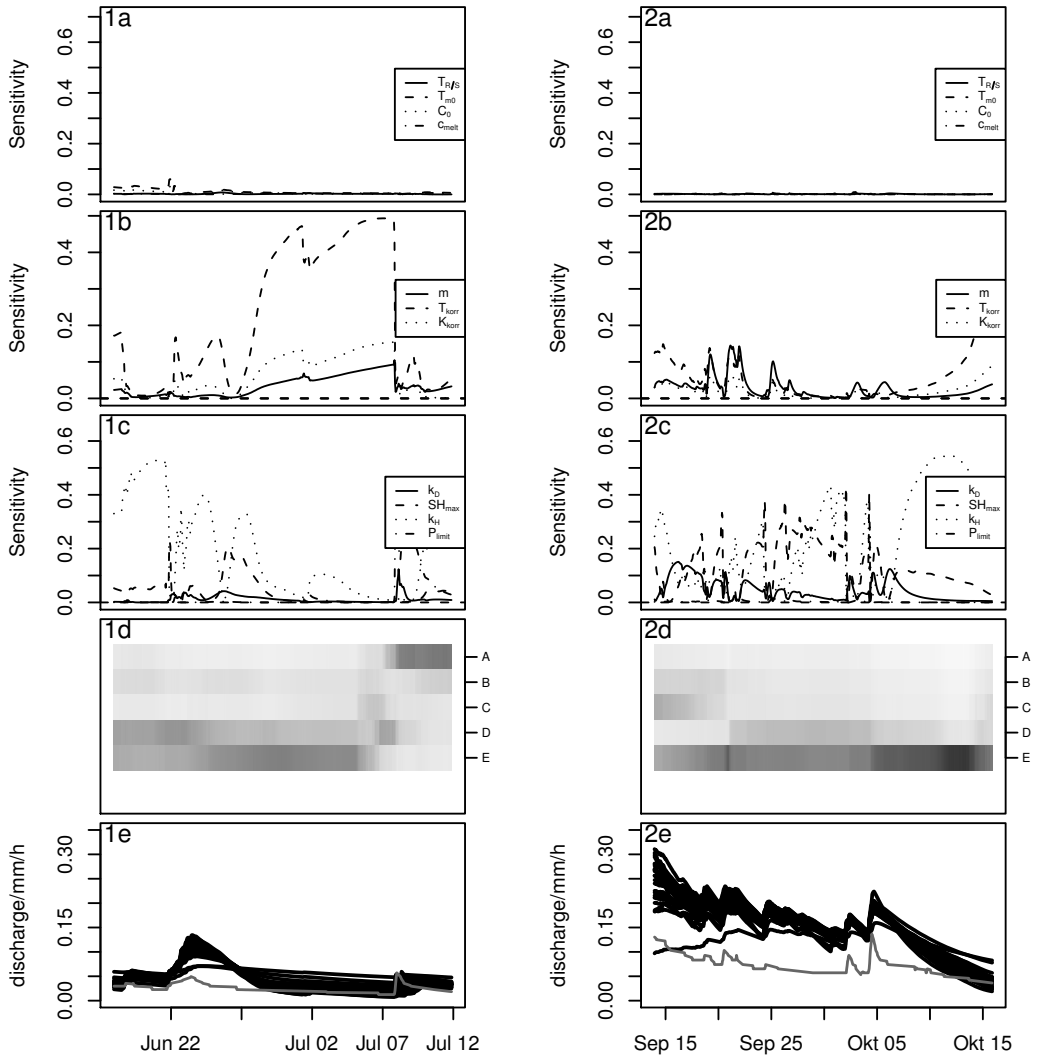


Figure 4.8: Details from Figure 4.5 for July (left plot) and September 2001 (right plot).

In addition to plausibility checks, the analysis of temporal dynamics of model parameters provides at least two further benefits. First, an understanding of the temporal dynamics of parameter sensitivity provides a valuable context for the calibration of model parameters. In general we would expect that periods of high parameter identifiability should coincide with periods of high parameter sensitivity. Our results suggest, therefore, that the fraction of melt water contributing to overland flow, c_{melt}, will hardly ever be well identifiable, because the sensitivity of simulated discharge for this parameter is always smaller than 2% despite the large range for c_{melt} of $20 \cdots 50\%$.

Second, it is possible to detect compensatory effects of parameters, indicated by a highly correlated sensitivity of the model output for multiple parameters. Correlated model parameters can be a major source for poor identifiability in hydrological modelling (Bárdossy, 2007). While Sieber and Uhlenbrook (2005); Cloke and Pappenberger (2009) used TEDPAS for plausibility checking, our results indicate that the TEDPAS approach can also be used to identify correlations among model parameters. In our case study, we observe correlation among the saturation deficit related parameters, which may complicate proper identification of these parameters during calibration.

4.4.2 Model performance (TIGER)

We found 5 clusters of model performance which we characterized with synthetic peak errors (Fig. 4.3). The Nash-Sutcliffe coefficients of efficiencies were greater than 0.47 for the 25 selected, model runs. The temporal pattern of model performance shows that acceptable agreement between model and observation (cluster A) occurs mainly during late spring and summer (Fig. 4.5 and Tab. 4.2). Four types of deviations (clusters B ... E) are observed during the other periods:

1) Completely missing peaks during snow season (cluster B). This will be further discussed below.

2) Major snow melt events are generally overestimated (cluster C).

3) At the start of the simulation we observe poor reproduction of dynamics and differences during low flow periods (cluster D). This indicates that the current model initialization (repeating a complete 2 year simulation with daily time steps until the saturation deficit storage stops changing and "compensating" for the starting conditions) can be further improved.

4) Strong overestimation combined with recession phases that are too fast in the model compared to the observation is observed thorough out the entire simulation period (cluster E).

4.4.3 Combination of model performance and parameter sensitivity

By combining the information regarding model performance with that about parameter sensitivity, we expect repeated patterns of similar error finger prints to point towards model structural deficits. Here, we discuss alternative possible explanations for the patterns observed in our study, and discuss strategies for distinguishing between alternatives.

1) The missing sensitivity of the modeled discharge to the temperature melt index during the periods with missing peaks (Example A in section 4.3.3) indicates that no snow melt is occurring in the model. This is in marked contrast with the observed increases in discharge. A check of temperatures (in the data record) shows them to be well below T_{m0} during these periods. Therefore, the catchment may be experiencing radiation induced melt events (process not included in the model) or the observed rises in discharge may be caused by backwater effects due to ice jams. To investigate the hypothesis that these peaks are due to radiation induced snowmelt events, we

made a quick "back of the envelope" calculation, checking if incoming radiation is sufficient to release the required amount of water under the assumption of clear sky and neglecting the effect of the forest cover. Only looking at short wave radiation, the energy is sufficient to melt the required amount of water. However, as soon as longwave radiation is included, total energy input into the snow cover is negative. This is supported by simulations of Kneis and Heistermann (2009) who used a hydrological model with an energy balance based snow module called Larsim for the same catchment. They did not find discharge to increase during these periods with their model. Also, rechecking data sources revealed that data is potentially influenced by ice at the gauging station.

2) Cluster B (missing peaks, peaks too early) occurs for some summer events. These are always short periods (period nr. 14, 15, 16) that are influenced by interflow and saturation deficit parameters. These events are always followed by Cluster E periods with strongly overestimated discharges that are influenced by interflow and saturation deficit parameters (k_h, T_{korr}, and $SH_{\max}$ – periods 40, 41, 42). This overestimation of discharge rises suggests that the model system does not retain enough water during interflow dominated periods. A check of the water balance revealed that the cumulative simulated discharge of 1310 - 1335 mm is 25% larger than cumulative observed discharge. This may be related to underestimation of evapotranspiration, and further analysis using calibrated model runs will be necessary to explore this hypothesis.

3) Cluster C periods consist of winter events with acceptable dynamics, but with an upwards shift of discharge leading to overestimation. This could be avoided if more water was stored in the system during snowmelt periods. Different values for the overall snowmelt indices are unlikely to resolve the problem, since the 25 selected parameter sets already provide relatively good values for the Nash-Sutcliffe efficiency.

   Overall, it is likely that the snow model may be too simple for this catchment. Complexity may be added by (for example) extending the model with a radiation induced melt component. This may lead to a higher estimate for T_{m0}, resulting in less water stored in the snow cover and released during the melt period. Alternatively, the model structure could be extended to use land-use related snowmelt indices instead of a single parameter, since melt is reported to generally occur slower in forested areas (Herbst and Casper, 2008; Winkler et al., 2005; Storck et al., 2002).

4) Cluster D is mainly a period of poor performance during the beginning of the simulation. We hypothesize that the current model initialization should be further improved (mainly by including a longer warm up period). This is further supported by the fact that discharge is much more sensitive to parameter m at the beginning of the simulation than during any other period. The remaining cluster D periods occur together with cluster E periods, however cluster D has lower discharge values. For cluster E (example C) we observe that the recession is too fast, while the discharge is sensitive to various parameters. Recession analysis (Fenicia et al., 2006; Wittenberg and Sivapalan, 1999) could be used to better understand the recession process during these periods. We suspect that a linear reservoir approach for interflow may be too simple.

4.5 Conclusions

The core idea of this study is to provide a novel diagnostic approach for a joined analysis of tem-

poral patterns of a) poor model performance and b) of dominant model components. More specifically, the idea is to work out whether certain types of model errors occur in coincidence with a) a certain context (snow melt, recession periods) and b) a high sensitivity of always the same model parameters. We suggest that coherence in the temporal patterns of error types and dominance of model components/parameter sensitivity allows a targeted identification of data errors and/or structural deficiencies of model components. This is a precondition for improving models to reduce occurrence of a certain error types in a targeted way.

Reduction of model structural uncertainty can be achieved in two ways. One approach is to test the model against several sets of independent target data. This is often be referred to as multi objective parameter estimation and means to increase the "information content" of the target data space. The other approach is to represents dominant processes and their controls such that characteristic behaviour can be reproduced in a more realistic manner, for instance resolving lateral flows and surface and subsurface flow paths, or reproducing subsurface storage volumes. This is often referred to as "process complexity" of the model and means to reduce the manifold of acceptable model structures. The use of more complex models implies that computational effort and simulation times increase considerably. The proposed approach is fast enough to be applied to models with increasing complexity because: a) it is not necessary to calibrate the model in advance, b) a highly efficient method is used to sample the parameter space, and c) all model runs are evaluated (to determine parameter sensitivity) while other Monte Carlo based methods often discard the 90% worst runs as a first step.

The case study shows that the method is able to enhance our understanding of the model's structural deficits with respect to the catchment. We expect the same model to show different structural deficits in different landscapes, and different model concepts to show different structural deficits in the same landscape. Consistent application of the proposed methodology could, in the long term, enable the development of basis for discriminating model/process concepts and landscapes into "compatible and incompatible sets" (in which the model/process can be expected to work with low structural/high structural deficits). Ultimately, it could help to reduce the overwhelming number of hydrological models to a minimum amount necessary for dealing with the richness of our landscapes.

Building upon the expected different structural deficits to be identified for different models, the approach presented may change the way model comparison is performed. The temporal dynamics of model performance allow to test if similar patterns of model performance are observed for a given hydrological context for different process descriptions. If the patterns of model performance do not differ, we can conclude that the process descriptions in the models are not distinguishable in terms of the process dynamics produced. This way, we might also be able to reduce the number of possible process representations into a small set of distinguishable formulations.

With respect to data, the approach is efficient in highlighting periods of possible data errors, for which additional checks are necessary. Also, specific conditions for which an improved understanding is necessary are highlighted by our method, which makes it possible to collect additional data in a more targeted way. Thus, the approach can be used to guide field experiments.

Future research may include application to different landscapes and model concepts, through testing with virtual landscapes and well-defined model deficiencies as well as the analysis of additional model output variables e.g. ground water levels or areal patterns of snow heights.

4.6 Acknowledgments

We would like to thank Hoshin Gupta and the anonymous reviewers for their comments on earlier versions of this manuscript whose comments helped to significantly improve it. Discussions with Bettina Schaefli were very helpful during the initial stage of this investigation. The study has been funded as part of OPAQUE (operational discharge and flooding predictions in head catchments), a project within the BMBF-Förderaktivität "Risikomanagement extremer Hochwasserereignisse" (RIMAX). We would like to thank Jenny Eckart for her support with the data preprocessing for WaSiM-ETH. A major part of the analysis was carried out with the open source statistical software R and contributed packages, we would like to thank its community.

Chapter 5

Low-cost monitoring of snow height and thermal properties with inexpensive temperature sensors. *

Small, self-recording temperature sensors were installed at several heights along a metal rod at five locations in a case study catchment. For each sensor, the presence or absence of snow cover was determined based on its insulating effect and the resulting reduction of the diurnal temperature oscillations. Sensor coverage was then converted into a time series of snow height for each location. Additionally, cold content was calculated. Snow height and cold content provide valuable information for spring flood prediction.

Good agreement of estimated snow heights with reference measurements was achieved and increased discharge in the study catchment coincided with low cold content of the snow cover. The results of the proposed distributed assessement of snow cover and snow state show great potential for a) flood warning, b) assimilation of snow state data, and c) modelling snowmelt process.

*Dominik Reusser, Erwin Zehe (2011), *Hydrological Processes*, in press, doi:10.1002/hyp.7937

5.1 Introduction

Comprehension of snowmelt induced floods requires a good understanding of the snow cover in terms of spatial distribution and temporal evolution. Two of the key parameters of the snow cover are the amount of water stored (snow water equivalents – SWE) and the cold content defined as the amount of energy necessary to trigger the melting process. Therefore, to make progress towards improved real time warning of snowmelt events, we are interested in the detection of 1) the amount of snow and 2) the required energy input to reach the melting point of the snow cover.

The amount of snow on large scales is commonly assessed through a combination of field measurements and remote sensing. The standard approach is to determine snow-covered area (SCA) and observe its change over time from remote sensing data (e.g. Durand et al., 2008; Kolberg and Gottschalk, 2006) or photography (Farinotti et al., 2010). Empirical relationships between SCA and SWE (so called snow depletion curves) are subsequently used to determine snow amounts (e.g. Liston, 1999; Durand et al., 2008; Kolberg and Gottschalk, 2006). Assimilation methods like Kallman filters may be used to combine model predictions with SCA information (Andreadis and Lettenmaier, 2006; Clark et al., 2006) or SWE data (Slater and Clark, 2006).

Reliable measurements of snow height or SWE are required since snow depletion curves need to be validated (Essery and Pomeroy, 2004; Pomeroy et al., 2004; Liston, 1999). Furthermore, SCA may be rather uniform at smaller extents (less than 100 km^2) and thus deemed as poor predictor for SWE and snow height. Manual measurements of snow courses are very labour intensive. Conventional equipment (snow pillows for SWE measurements and ultrasonic sensors for height measurements) is relatively expensive (>2000 Euro for one location) and thus allows sampling at a rather coarse spatial resolution.

Instead, inexpensive temperature sensors could be used at a higher spatial resolution with the same expenses. The measuring principle is based on the fact that the snow cover results in a strong reduction of daily temperature fluctuations. Lundquist and Lott (2008) demonstrated the characterization of snow patchiness and snow accumulation patterns with such inexpensive temperature sensors. For their measurements, single sensors were buried in the soil and the time of snow cover disappearance was recorded. The date the snow cover disappeared was converted to an estimation of the amount of snow that accumulated at the start of the melt season with a snowmelt model. The approach of Lundquist and Lott (2008) requires climatic data for the snow model and is based on the assumption that the snow model is representative. Also, their analysis can only be performed after the sensors are uncovered.

The main objective of our study is to obtain distributed data on snow height and snow temperature profiles by installing cheap temperature sensors at multiple locations. Our approach does neither depend on a snow model nor on climatic data for the determination of the snow height. We achieve this goal with multiple sensors installed at different elevations above ground at the same location.

The data from the simple, robust and cost effective temperature measurements in and above the snow cover will be assessed for their value for simultaneously obtaining information about snow height and temperatures, and the cold content. Related methodological issues to be solved are: 1) Can we find an algorithm to extract snow height information from temperature data? 2) How well do the estimated snow heights compare to reference measurements? 3) How do we calculate cold content from the temperature profile?

We describe the methods in section 2, results are presented (section 3) and discussed (section 4). Conclusions are drawn in section 5.

5.2 Methods

5.2.1 Measurement locations and experimental design

As a case study, we selected the upper catchments of the Wilde and Rote Weisseritz, situated in the eastern Ore Mountains close to the Czech-German border. Slopes are gentle with an average of 7°, 99% are <20°; calculated from a 90 m digital elevation model (SRTM, 2002). The area is 49.3 km^2 and 47.8 km^2 for the Wilde (gauge Ammelsdorf) and Rote Weisseritz (gauge Schmiedeberg), respectively (Figure 5.1a). About 50% of the area is forest and 40% is used for agricultural activities. There are only a few villages and towns in the upper catchment. Mean temperatures are 11°C and 1°C for the periods April - September and October - March, respectively. Annual precipitation is around 1000 mm/year. Discharge data for two gauging stations (Figure 5.1a) were obtained from the State Office for Environment and Geology (LfUG, 2007).

Sensors were placed at five locations in the upper part of the catchment (Figure 5.1 and table 5.1) where a snow cover of about 1 m is abundant for one to four months with high discharge during the snowmelt period. The sensor set placed at the lowest location was installed at around 500 m above sea level (table 5.1), the highest at 760 m a.s.l. and the catchments go up to about 900 m a.s.l. Gentle slopes on grass land were selected as measuring locations. All but one location had low exposure to wind (table 5.1).

To estimate the extent (Blöschl, 1999) of our measurements, the combined catchment area for the two rivers is relevant with an area of $\sqrt{A_{\text{catchment}}} \approx \sqrt{100\ \text{km}^2} = 10$ km, while the spacing is $\sqrt{A_{\text{catchment}}/n} \approx \sqrt{100\ \text{km}^2/5} = 4.5$ km The support is calculated from the measuring area of a sensor (r=(10 cm)$^2 * \pi$), resulting in a value of $\sqrt{A_{\text{sensor}}} = \sqrt{0.031\ \text{m}^2} = 0.18$ m.

Temperatures were measured and recorded with a Hobo pendant temperature data logger (Figure 5.1d). The logger has a size of 58 x 33 x 23 mm (about the size of a matchbox). Temperatures can be recorded in a range from -20 to +50°C with an accuracy of ± 0.47°C at 25°C and ± 0.8°C in the full measurement range. The data loggers are water tight and have a storage capacity to hold about one year's worth of ten minute data. Costs are around 20 Euro for each logger.

At each location, nine sensors were placed on a square metal rod with a spacing of 15 cm covering a range from 0 to 120 cm above ground (Fig 5.1c). We will refer to such a rod with nine sensors as a temperature sensor set.

A reference station was set up at an experimental station of Technical University (TU) Dresden and TU Freiberg located near Baerenfels (Fig 5.1a). The main purpose of the station is the measurement of air pollutants and meteorological variables. The station is at an elevation of 735 m above sea level. More details about the station and additional measurements are available on the station web page (Eichelmann, 2009).

Snow water equivalents were measured with a 3x3 m snow pillow made by the company Sommer from a stable PVC-sheet (Fig 5.1b). The snow pillow was installed on a level sand bed and filled with 600L of a water ethylene-glycol mixture (2:1) (IUPAC name: ethane-1,2-diol, obtained from Sigma-Aldrich). A pressure sensor DMP 331 (from BD sensors) measured the pressure inside the pillow in the range from 0-250 mbar relative to atmospheric pressure, giving a constant current signal (independent of the voltage) proportional to the pressure. The accuracy of the sensor is better than 0.1%.

The collected data were noisy, showing large short term variations. We were unable to identify the reason for the noisy measurements. Data quality was acceptable after applying a filter that ac-

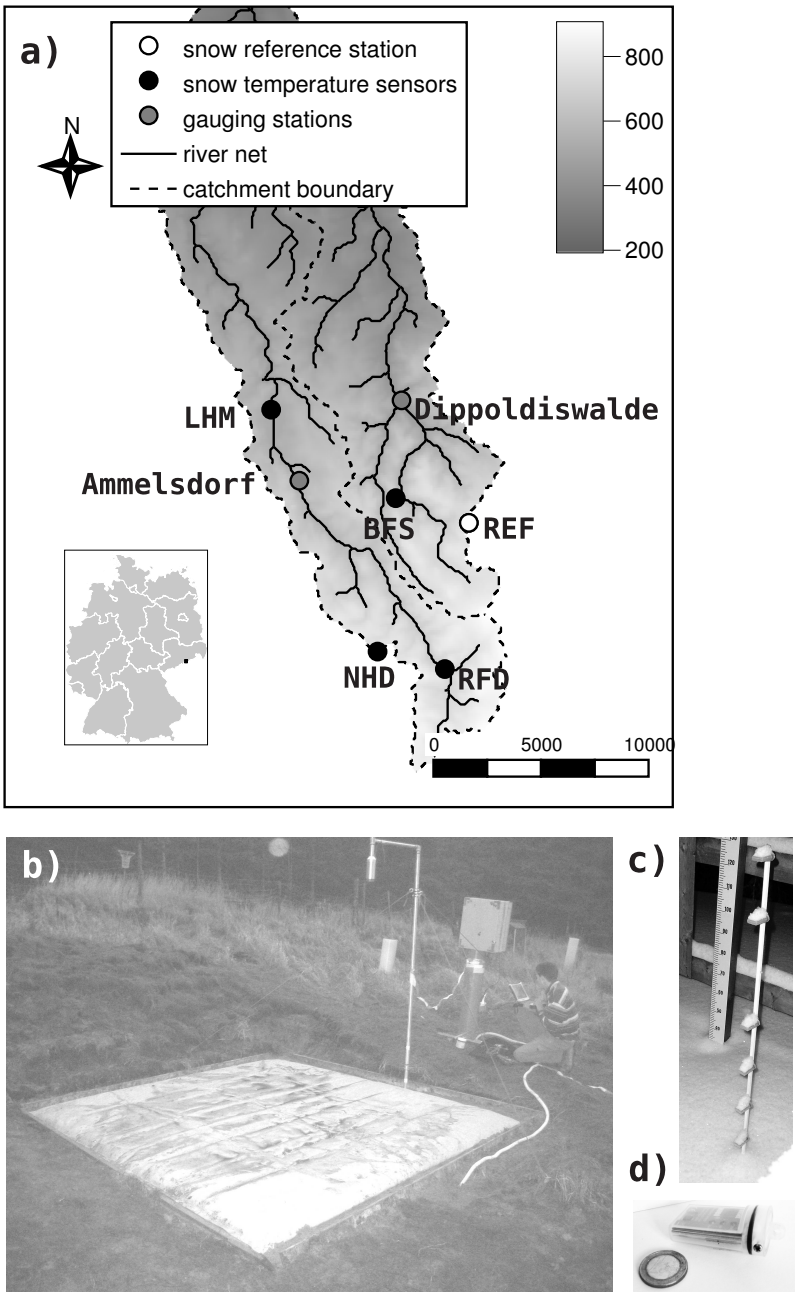


Figure 5.1: Weisseritz catchment: a) digital elevation model and location of temperature sensors and gauging stations (elevations and distances in meters) b) snow pillow and ultra-sonic sensor used for reference measurements c) installation of temperature sensors d) a single Hobo data logger and a 1 Euro coin as reference for the size.

Nr.	Location	Abbreviation	Elevation	Exposition	Description
1	Neuhermsdorf	NHD	760	west	clearing in a small valley
2	Oberbaerenburg	REF	735	leveled ground	clearing
3	Rehefeld	RFD	700	north east	open grassland, windy
4	Baerenfels	BFS	630	east	open grassland
5	Lehnmuehle	LHM	520	west	open grassland

Table 5.1: Location of sensors

cepted data within $\pm$ 0.2% around the local median, calculated from 151 10-minutes measurements ($\approx$24h).

A temperature sensor set (REF, see table 5.1) was installed about five meters from the snow pillow.

Snow height was recorded with an ultrasonic sensor SR50 (Campbell Scientific) with a measurement range of 0.5 to 10 m, a resolution of 0.1 mm and an accuracy of $\pm$ 0.4% of the distance to the target (at least $\pm$ 1 cm). The sensor was mounted at 1.8 meters directly over the snow pillow and the beam has a range of 22° resulting in an observed area with a diameter of about 0.7 m. The sensor failed to measure for some periods for reasons that we have not been able to identify. The sensor failures were clearly identifiable by a distance to ground of 0 m and were discarded.

Snow surface temperature was measured with an IRTS-P infrared temperature sensor (Campbell Scientific). The sensor has an accuracy of $\pm$0.3°C in the range from -10 to 55°C. It was also installed at a height of 1.8 m above ground. The 3:1 field of view results in an observed area with diameter 0.6 m. A correction for sensor temperatures was applied as advised by the supplier (Campbell, 2006).

A DL2e data logger (Delta-T Devices Ltd) was used to monitor and record data from all sensors at the reference station at an hourly interval. In March 2009 there was a logger failure and no data were recorded until the next field trip at the end of April.

Additional data included a snow report on the web page of the hotel SWF Skibahnhof located in Neuhermsdorf (Dietrich, 2009, refered to as web data set). We were in contact with the hotel staff throughout the research project and agreed, that snow height readings on the web page would be stored for this project. The snow height was read from the measurement pole shown in Figure 5.1c) and reported on the homepage. The page was downloaded daily and snow height information was extracted using standard Linux tools (grep, awk, vim). As shown in Figure 5.1c) one temperature sensor set (NHD) was installed within 1 m of the location of the reported snow height.

Manual snow depth measurements were made on fields close (within 500 m) to the temperature sensor sets during 5 campaigns on January the 16th and 30th, February 13th and 27th and March 26th. Snow depth was measured 60 times at each location using a sampling scheme with 1 m spacing as described by Jost et al. (2007).

5.2.2 Snow height estimation

The underlying idea is to use the reduction of the diurnal temperature variation caused by the insulating effect of snow to detect the height of the snow cover. Figure 5.2 (left) shows temperature data (using a grey scale) for a two day period at the end of February 2009 for different heights above ground. The figure clearly shows a very

constant temperature for all sensors at and below 75 cm. Above, a clear diurnal signal can be detected with highest temperatures around noon. The corresponding variance for the two day period for each sensor is shown in Figure 5.2 (right). A sharp drop of variance between the sensor at 90 cm and the sensor at 75 cm is clearly visible. The height of this maximum drop was determined for each day of the study period (using daily variances), giving an estimate of the height of the snow cover h_{est1} and the height of the lowest free sensor h_{free}.

Since h_{est1} is nearly random for snow free periods (we do not expect the maximum drop of the temperature variance to be observed at a certain, constant height), we included two additional conditions: 1) setting the height of the snow cover (h_{est2}) to 0 cm if temperatures above 2 °C were observed below height h_{est1} and 2) ignoring changes in snow height if the mean absolute change in snow height for five days was larger than 10 cm/day. This rate was determined empirically from the data set at the reference station.

h_{est2} has a vertical resolution corresponding to the spacing of the temperature sensors. However, during melting periods, this resolution can be increased the following way: Each time a temperature sensor is released from the snow cover, we observe a reduction of h_{est2} as a step function. The amount of snow melted between two such steps can be related to temperature with a simple temperature index model (TIM) (e.g. Ferguson, 1999):

$$\delta h = \begin{cases} ti * (T_{\text{air}} - T_{\text{lim}}) & \text{for } T_{\text{air}} > T_{\text{lim}} \\ 0 & \text{for } T_{\text{air}} \leq T_{\text{lim}} \end{cases} \quad (5.1)$$

δh: change in height in cm; ti: temperature index in cm/°C/time unit; T_{air}: air temperature; $T_{\text{lim}} = 0$°C: melting temperature

Equation 5.1 is used to interpolate the snow height (h_{est3}). Note that usually, TIMs are formulated in terms of changes of snow water equivalents while we use a formulation in terms of snow height. The density of the snow cover could be used to convert ti from our model to a ti for a standard TIM if settling of snow were not of importance.

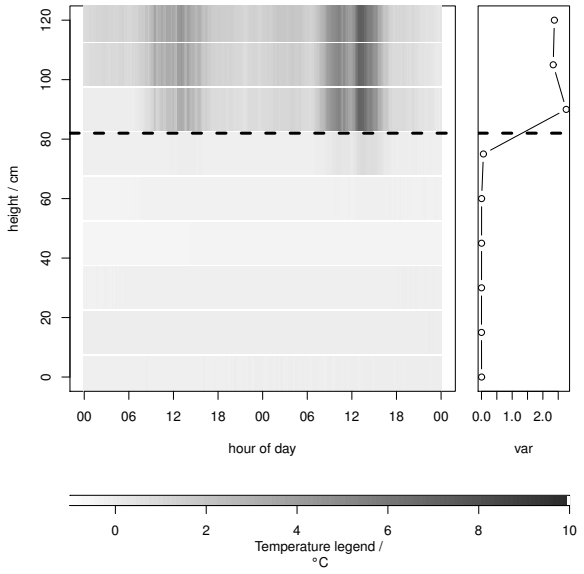


Figure 5.2: Core idea for the estimation of snow height from temperature data: the left hand plot shows the diurnal temperature variation (two days) as function of sensor height above ground. The right hand side plot shows the variance of the temperature data. The black dashed line indicates the estimated height of the snow cover.

5.2.3 Cold content of snow cover

The cold content Q_{cc} (Equation 5.2) defines the amount of energy required to bring the snow cover to a temperature of 0°C (Dingman, 2002). The snow cover starts melting if the cold content is 0 and additional thermal energy enters the snow cover. Knowledge about the onset of snowmelt is an important piece of information for the prediction of spring floods.

$$Q_{cc} = -c_i * \rho * h * (T_s - T_m) \quad (5.2)$$

$c_i = 2102$ J/kg/K the heat capacity of ice (Dingman, 2002), ρ density of the snow, h height of the snow cover, T_s average temperature of the snow cover, and $T_m = 0°$C melting temperature. Note that this approach neglects heating of liquid water in the snow pack.

With the temperature measurements and the height h_{est3}, data for all variables except for the density of snow are available. If we can make a reasonable assumption about snow density, we are able to determine Q_{cc}. The density of freshly fallen snow ranges from 4 to 340 kg/m^3 but an average relative density of $\rho_{ns} = 100$kg/m^3 is often assumed (Dingman, 2002). Well drained snow at melting point has a density near 350 kg/m^3 (Dingman, 2002) and (Anderton et al., 2004; Jonas et al., 2009; Lundberg et al., 2006) report relationships between snow depth and density. We used snow density obtained from the snow pillow and snow height measurement at the reference station.

The methods and models presented have been implemented as part of the R-package RHydro (Reusser and Buytaert, 2010). An example data set is also available in the package.

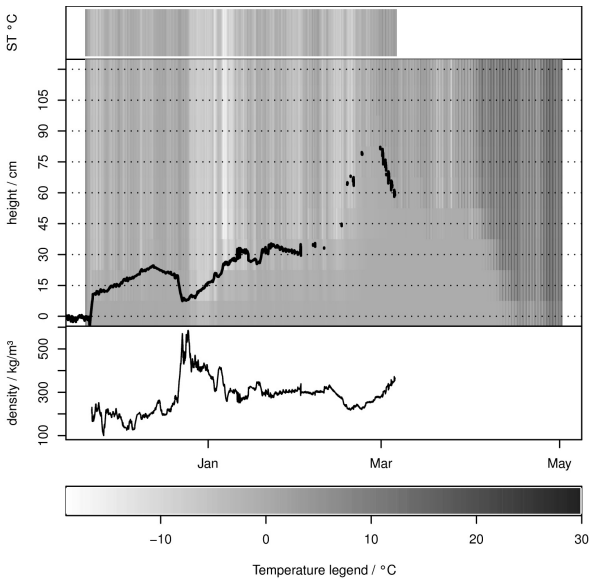


Figure 5.3: Measured time series for the reference station: snow surface temperature (infrared sensor) (top); observed temperatures as function of height above ground and snow surface height (ultrasonic sensor) marked by black points (middle); snow density (bottom)

5.3 Results

5.3.1 Measurements at the reference station

Figure 5.3 shows time series for measured data at the reference station, using grey scales for temperature values. The top plot shows snow surface temperatures. In the second plot, observed temperatures in and above the snow cover are shown (height above ground on the y axis) together with snow height from the ultrasonic sensor. The third plot contains snow densities calculated from the SWE (snow pillow) and the snow height.

Diurnal variation of air temperature is clearly visible in Figure 5.3, as well as the seasonal pattern of temperature with lowest temperatures recorded in January. The surface temperature closely follows air temperature for the recorded period. As expected we observe the strong reduction of the diurnal temperature variation due to the snow cover. The snow cover starts in December. Around new year, the snow height decreases, at the same time snow density increases from the low densities recorded for fresh snow to almost 500 kg/m^3 due to compaction during melting periods. From the beginning of January until the beginning of March, snow is further accumulated and at the beginning of March, the onset of snowmelt is recorded. As already reported, data from the reference station is not available after beginning of March due to a logger failure. Problems with the ultrasonic sensor resulted in the data gaps apparent in Figure 5.3.

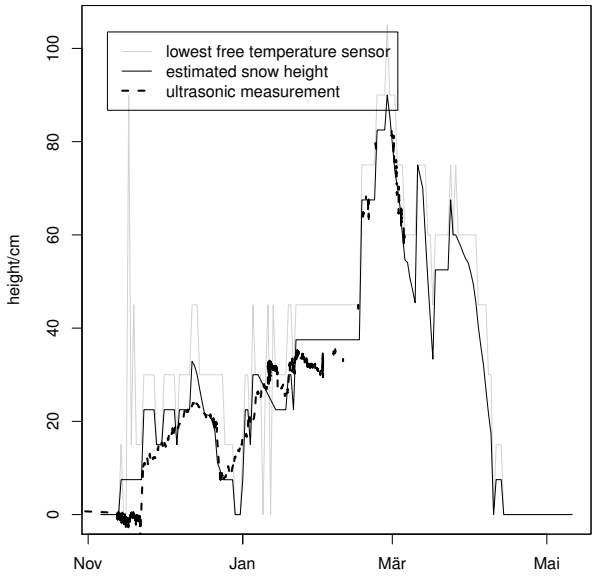


Figure 5.4: Snow surface height at the reference station: ultra-sonic reference, and h_{free} and h_{est3} from temperature measurements

5.3.2 Height estimation

Snow height estimation h_{est3} and h_{free} based on the algorithm described in section 5.2.2 are shown in Figure 5.4 for the reference station. The mean absolute error between interpolated results and ultrasonic measurements for the period from December 1$^{\text{st}}$ until end of January is 5.3 cm (see discussion for an expected value for the mean absolute error).

Estimated snow heights for the remaining four locations are presented in Figure 5.5. Measured temperatures for different heights are color coded. Again, the reduction of the diurnal variation due to the snow cover is visible. Snow height (h_{est3}) is white, while the web data set is shown as red dotted line for the Neuhermsdorf location (NHD). The mean absolute error between the web dataset and the corresponding h_{est3} is 6.0 cm.

In general, h_{est3} is within the measured variability from the manual snow depth measurements (shown in black in Figure 5.5). The only exceptions are the measurements at the Neuhermsdorf location (NHD) after the end of March (last two measurements). However, h_{est3} agrees well with the web data set during this period. The difference can be explained by the fact that the manual measurements were taken about 500 m away from the temperature sensor set on a field, which was more exposed to wind compared to the location of the temperature sensor set.

Temperature indices ti (Equation 5.1; table 5.2) were estimated for melting periods, during which two or more sensors were uncovered. Values are in the range from 0.4 to 5 mm snow height/day/°C. From the few data points, no seasonal trend can be determined.

As mentioned above, if settling of the snow cover was not an important process, density of the snow cover could be used to convert from a snow height TIM to a SWE-based TIM. In order to assess comparability of the ti values based on snow heights with ti values based on SWE, we fitted Equation 5.1 to both types of data for the reference station during the short melting period between December the 15$^{\text{th}}$ and 24$^{\text{th}}$ (Figure 5.6). The figure shows measurements (interrupted lines) of SWE, snow height (ultrasonic sensor), and h_{est3}. The TIM from Equation 5.1 was fitted to the melting periods of the three time series (shown with solid lines). Precipitation data are shown from the top of the figure.

Two melting phases can be observed with very different properties. No precipitation is observed for the first period and the index was found to be ti_2 =0.67 mm/day/°C (height based, ultrasonic sensor), while reduction of SWE (ti_4) is somewhat faster, as expected due to the density-factor.

During the second period, precipitation is observed (almost 20 mm in total). ti_3 (height based, ultrasonic sensor) is 3.4 mm/day/°C. During this event, the density of snow cover increases strongly, since the height is reduced to half of the initial value while SWE hardly decreases. Accordingly, the SWE-based index is found to be only ti_5=0.37 mm/day/°C. ti_1=3 mm/day/°C (height based, temperature sensors) is strongly influenced by the

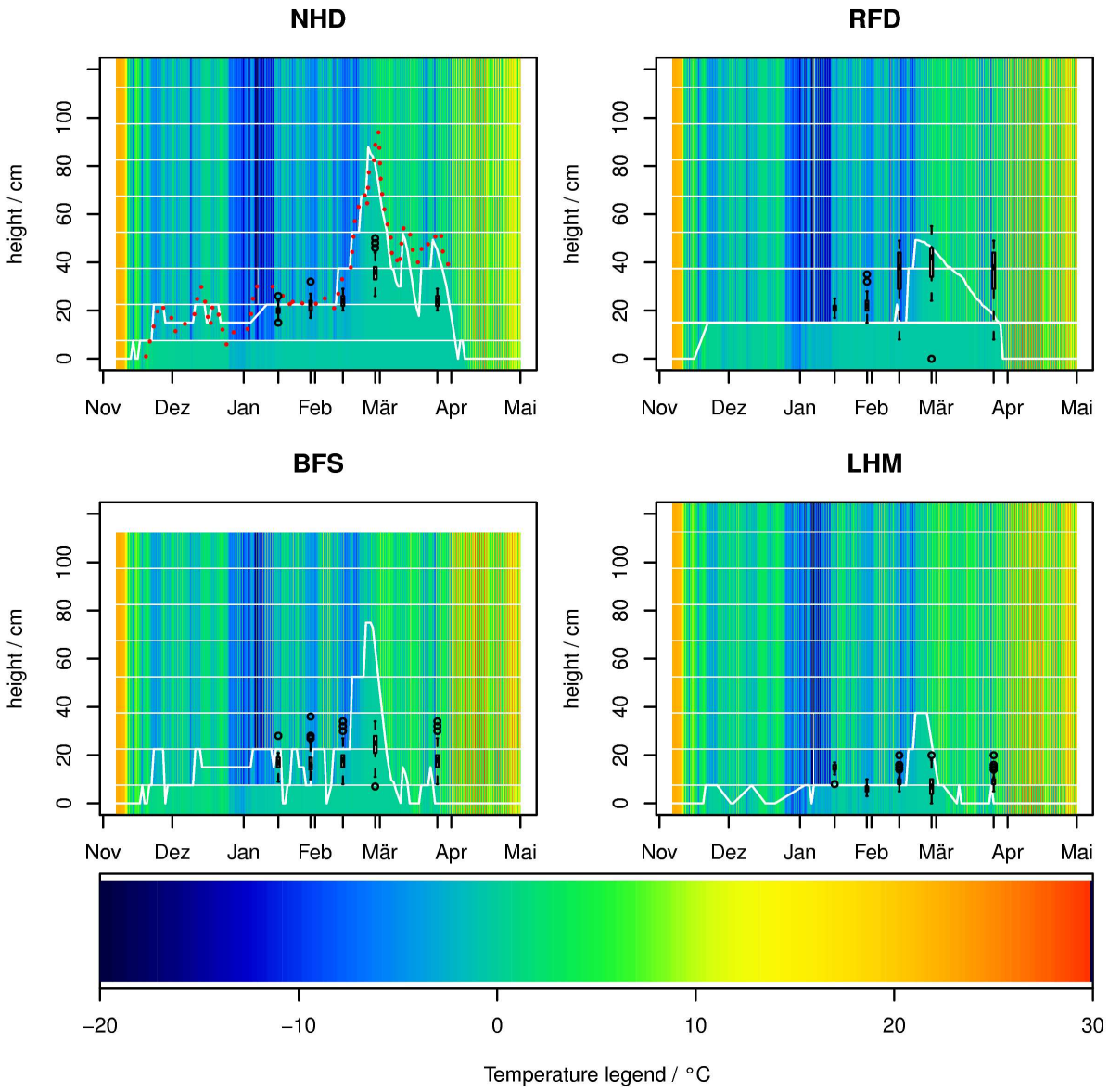


Figure 5.5: Temperature data (color coded) and estimated snow cover thickness (white) for all locations in the Weisseritz catchment. Manual reference measurements are shown in black. The red dots in the upper left panels refer to the web data set that consists of manual measurements of snow heights at the hotel SWF Skibahnhof located in Neuhermsdorf.

Period	REF	NHD	RFD	BFS	LHM
Late December	3.0				
Early March	2.2	2.6		1.2	
Mid March	3.7	2.3		1.3	
Late March	0.52	1.5		4.7	
March			0.42		

Table 5.2: Temperature indices ti (mm snow height/day/°C)

height reduction during this second event.

5.3.3 Cold content of the snow cover

Cold content describes how much energy is required to rise the temperature of the snow cover to 0°C and indicates when melting processes will start. It was determined according to Equation 5.2 and is plotted in Figure 5.7 (middle) along with the estimated snow height (h_{est3}, top graph). The cold content varies between 0 and 500 kJ/m^2 and as expected, is low during melting periods (<50 kJ/m^2 for periods with decreasing snow height)

Discharge data for the two gauging stations is also shown in Figure 5.7 (bottom), to allow a first assessment of the relevance of cold content for the prediction of flood events. Increases of discharge occur only during periods with a low cold content (<50 kJ/m^2) and an existing snow cover.

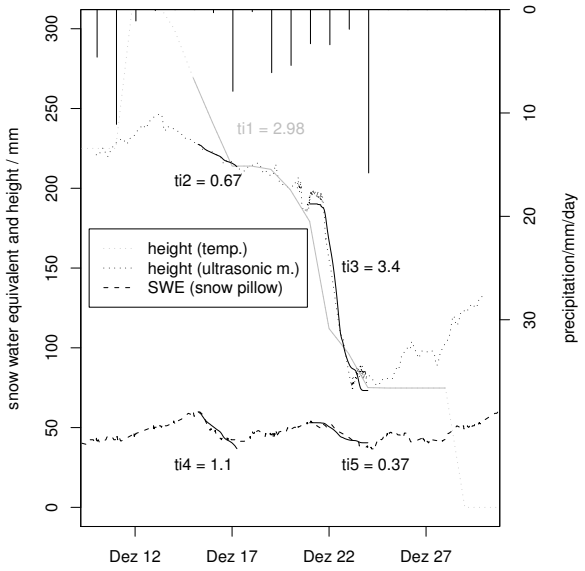


Figure 5.6: Snow height (measurements with ultrasonic sensor and estimated from temperature loggers) and snow weight (snow pillow) for the reference station. Snow melt/compaction modelled with TIM (Eq. 5.1) is shown in solid lines together with best estimates (least squares) for the ti-value. ti_1 through ti_3 are reported in mm snow height/day/°C while ti_4 and ti_5 are reported in mm SWE/day/°C. Rain fall data is shown from the top. To make the distinction of data easier, values derived from temperature measurements are shown in grey.

5.4 Discussion

5.4.1 Temperature measurements

The temperature accuracy and resolution of the Hobo sensors as well as a temporal resolution of ten minutes was sufficient for this application. The reduction of the diurnal variation was clearly detectable. Therefore, the data are well suited for the estimation of snow height and temperature profiles.

We observed that the snow cover is influenced by the sensors and the metal rod during the melting period, resulting in increased melting just around the sensor set. Our recordings did not allow us to

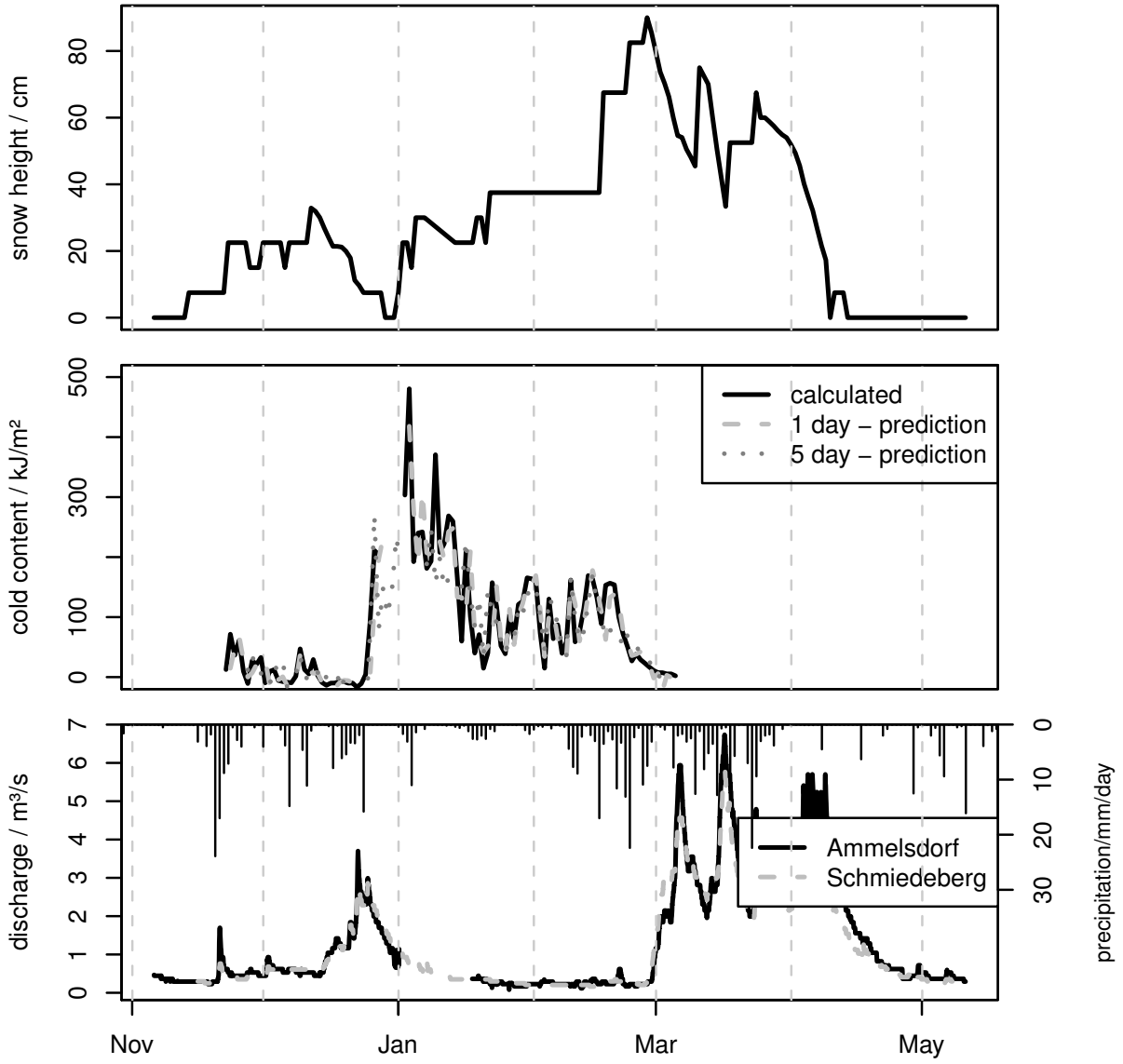


Figure 5.7: Snow height and cold content of the snow cover at Baerenfels-station as well as discharge observed at the two gauges. Cold content was calculated from observed temperatures and predicted one and five days ahead with a thermal diffusion model (Appendix).

fully assess this influence since manual reference measurements were not sufficiently frequent, the data logger at the reference station was not in operation, and the web data set ended towards the end of the snow season.

5.4.2 Snow height estimation

The maximum reduction of diurnal variance along the profile appears to be a robust indicator to estimate snow height for periods of snow cover. The resulting time series of snow height is in good agreement with the reference measurements with a mean absolute error of about 6 cm. This is the expected minimum error for a sensor spacing of 15 cm and the method presented.

For the estimation of the expected minimum error, the vertical resolution of the web data set was reduced to a 15 cm resolution (reduced web data set). This corresponds to perfect information about the snow coverage of each single temperature sensor. The reduced web data set was then used as h_{est1}-values in the normal snow height estimation procedure (section 5.2.2). The mean absolute error between the resulting snow height estimate h_{est3} and the original web data set was found to be 6 cm.

Small underestimations of snow height are likely, because the diurnal variation may be observed even though a thin snow cover is present. The thickness of the snow cover above the sensor that is required to sufficiently reduce the diurnal oscillation in order to detect the snow height is not known.

Location specific properties are apparent from the measurements: minor snow fall occurred in March, as shown for the NHD-station and confirmed by the web data set. It is very likely that snow also fell at the close-by RFD-station but windy conditions redistributed snow immediately away from the station, so that no increase in snow height was observed.

For snow free periods the maximum reduction of diurnal variance along the profile results in near-random, implausible snow heights. Setting a limit to the mean absolute change in snow height for a five day window turned out to be an objective, satisfying approach to remove impossible snow height dynamics during such periods.

For the melting periods, the TIM is a useful interpolation method in order to increase the horizontal resolution of the estimated snow heights. We observed values between 0.42 and 4.7 mm height/day/°C. Note that it is not possible to distinguish the reduction of height due settling of snow cover and due to melting. Therefore, derived ti values based on height can not be compared to ti values based on SWE since compaction of the snow cover is an important process.

Even though direct comparison to values reported in literature is not possible, we can still compare the range observed. Hock (2003) summarizes ti for snow from more than 10 studies and reports values between 2.5 and 11.6 mm SWE/day/°C. In comparison, values around 0.4 appear to be very low, since conversion from height to SWE would introduce a density factor, which is < 1. One possible reason for such differences are extended periods of thawing and refreezing cycles, a process that is not represented in the simple model.

5.4.3 Cold content of the snow cover

The measured temperature data in combination with the estimate for the snow height are well suited to the calculate cold content of the snow cover. We used densities obtained from the snow pillow at the reference station for our calculation. If no snow density measurements are available, density may be estimated as a function of snow depth (Anderton et al., 2004; Jonas et al., 2009; Lundberg et al., 2006). Such an estimate causes some additional uncertainty. Since cold content is a linear function of snow density, effects of density errors can easily be evaluated using error propagation.

Cold content seems to provide a reliable assessment of the potential for snow melt and may be of

value for flood predictions. We present and discuss a very simple model for the prediction of the cold content in the Appendix. A sensor network (Hart and Martinez, 2006; Xu, 2002; Barrenetxea et al., 2008; Lundquist et al., 2003), transmitting measured temperatures in real time, might be useful for flood prediction as the height derived from temperatures gives an indication about the amount, and the cold content gives an indication about the potential for melt. The real time data may also be useful to update the temperature state of process based snow models, and thus to improve their performance.

5.4.4 Improving the approach

Limitations of the proposed approach may be distinguished into potential error sources and issues related to the practical application. Possible error sources include the slight underestimation because of the undetectable thin snow cover as described above, and the influence of the sensors and the metal rod on the snow cover. We suggest further experiments in the lab to better understand the temperature reading for bearly covered sensors. Experimenting with different materieals and coatings could reduce the effect of the setup on the surrounding snow cover.

As temperature measurements are an indirect way of determining snow height, we are dependent on the algoritm presented. This introduces additional possibilities for improvement. The criterion for snow free periods is currently based on two empirical factors, the upper limit to acceptable snow height change rates and the upper limit to within-snow-cover temperatures. With these two empirical parameters, it is not guaranteed to reliably identify all snow free periods – depending on the exact value of the parameters, either snow free periods are not identified as such or periods with snow cover may be detected as snow free. In addition, snow heights can not be estimated correctly for periods of diurnal temperature variability below a certain limit. We are grateful for suggestions for an improved algorithm. Such an improvement may for example include information from recent days in the estimation of the snow height.

There are practical limits to the horizontal resolution. Estimation of the cold content requires information on the density as described above, which is not available from the temperature measurements and for a real time application, data transmission becomes an issue.

5.5 Conclusions

In this study, we presented an inexpensive method for simultaneous estimation of snow height, the temperature profile and cold content of snow. While an ultra sonic sensor with logger is about 1000 Euro and provides information about snow height only, the method presented in this study is based on 10 temperature sensors, which are available for about 200 Euro.

To estimate heights, the method exploits the insulating effect of the snow which reduces temperature fluctuations in the snow cover. The estimated heights agree well with reference measurements. Temperature, snow height and cold content are interesting properties for spring flood warning because snow height is an indicator for the amount of snow available while cold content tells us the amount of energy required to start snowmelt.

From the data, we also attempted a prediction of cold content with a very simple model (Appendix). While predictions are satisfactory, the model itself is an oversimplification and identifiability of the single parameter α is poor. From this study we suggest that a decision based on current height, estimated cold content and weather predictions may be as beneficial.

Following our study, installation of a sensor network that transmits observations in real time may be an interesting future step. The benefit for operational flood warning and updating of snow models could then be fully explored. Filtering techniques are most often used to assimilate snow cov-

ered area or snow water equivalents (Andreadis and Lettenmaier, 2006; Clark et al., 2006; Slater and Clark, 2006). The assimilation of snow cover temperature data is innovative and appealing, since the observed states and model states are directly commensurable. While our results indicate a great potential, certainly more data are needed to better understand the added value of such data to predict spring flood events based on estimated snow heights and cold content. Further experiments should also investigate the thickness of snow cover above the sensor required to sufficiently reduce diurnal variation of temperature. Further tests should minimize the influence of the measurement set-up on the snow cover, as melting occurred somewhat faster around the sensors.

Acknowledgements

This study has been funded as part of OPAQUE (operational discharge and flooding predictions in head catchments), a project within the BMBF-Förderaktivität "Risikomanagement extremer Hochwasserereignisse" (RIMAX). We would like to thank Sophie Baumann and Christian Rinner who contributed to the snow height estimation algorithm. Stefan Lüdtke, Andreas Bauer, Markus Morgner, Thomas Gräff, Jenny Eckart, Niko Bornemann, Mareike Eichler, Thomas Recknagel, Andreas Passing, Maximilian Semmling, David Kneis, David Torhorst, Helge Gross, Bettina Schäfli, Jessica Papke, Karen van der Merve, Antonius Golly, Daniel Stichs, Carsten Neumann, Christian Lehr, and Richard Jung, were involved in the field work. We wish to acknowledge valuable discussions with Theresa Blume, Markus Weiler and Uwe Ehret. Maik Heistermeister and Till Franke and two anonymous reviewers made helpful comments on earlier versions of the manuscript. We obtained great support during the installation of the equipment from Peter Eckart, Bernd Böhme and Timo Junkers from the Landestalsperren Verwaltung Sachsen (State Office for Reservoir Management).

Appendix: Simple model for the prediction of the cold content

We tested a parsimonious 1d thermal diffusion model for analysing our temperature profile data to determine how much additional information can be derived. A full snow process model would at least include air temperatures, input of energy by radiation, influence of wind and rain, and mass transport of water vapor in the snow cover and possibly transport of snow by wind and avalanches. We take the parsimonious approach for the sake of assessing how much information can be derived from the temperature measurements only. For our simple model, we need to assume that: a) heat transport into the snow cover by radiation is proportional to observed temperatures (similar to the often used temperature index approach for snowmelt modelling (Ferguson, 1999)) and that b) temperatures in the snow cover are not influenced by wind, rain and melting processes. While obviously, these are strong and erroneous assumptions, they allow us to formulate the problem of predicting cold content as a simple diffusion model for heat (e.g. Brandt and Warren, 1997). Fitting the diffusion model to observed data provides an estimation of the thermal conductivity of snow (for more details, see below), which can then be used to predict the cold content.

Thermal conductivity is commonly reported as effective thermal conductivity (ETC), which summarizes a number of complex transport processes including conductivity in ice and air spaces, as well as latent heat flow due to water vapor (Brandt and Warren, 1997). Literature values for ETC are for example reported by Fukusako (1990) and are dependent on density, temperature and snow microstructure, ranging from about 0.1 to 1.1 W/m/K. Measurements of ETC are frequently discussed (Aggarwal et al., 2009; Brandt and Warren, 1997; Sturm et al., 1997; Satyawali and Singh, 2008; Schneebeli and Sokratov, 2004) and are based on

one of three basic methods (e.g. Brandt and Warren, 1997): 1) attenuation and time lag of the natural temperature signal, 2) transient measurement with a so called transient-probe method, measuring the temperature response to a short heating pulse of a needle inserted into the snow cover, and 3) steady state method with a constant thermal gradient in the lab. The first method corresponds to solving a diffusion model of heat and has the problem mentioned above that some of the required assumptions generally do not hold (see discussion for more details).

Method

For the prediction of the cold content, we estimate future temperatures in the snow cover with the heat diffusion model described in Equation 5.3, which is based on a single model parameter α. Temperatures were calculated one to five days ahead with a time step equal to the measured frequency of the temperature data. Similar models for heat transfer in snow have been applied previously (e.g. Brandt and Warren, 1997).

$$\frac{dT(z)}{dt} = \alpha \frac{d^2T}{dz^2} \tag{5.3}$$

$\frac{dT(z)}{dt}$: change of temperature with time, and $\frac{d^2T}{dz^2}$: curvature of the snow temperature profile.

As initial conditions, we used an interpolation of measured temperatures at the first time step of the simulation period. We assumed perfect air temperature predictions, thus we used the measured temperatures of the top temperature sensor as upper boundary condition. For the lower boundary condition, measurements at the snow soil interface were used.

Solving the heat diffusion model described in Equation 5.3 requires an estimate of the model parameter α. We tested two estimation methods: 1) assuming a fixed value for α and 2) estimating α for each day by minimizing the difference between the measured and modelled (Eq 5.3) temperature of the previous two days (overlapping two day windows with a 1-day spacing).

For the first method we need an estimation of the upper bound for α. For our simple thermal diffusion model, α can be expressed as a function of three characteristics of the snow cover: $\alpha = \lambda/\rho/c_i$, where density ρ and heat capacity c_i were defined for Equation 5.2, and λ is the effective thermal conductivity (ETC). Assuming a mean density of $\rho = 200$ kg/m^3 and using the constant heat capacity of ice (the heat capacity of the enclosed air is negligible), the problem reduces to estimating upper bound for λ.

As an upper bound for ETC, the thermal conductivity for ice may be used $\lambda_i = 2.2$ W/m/K (e.g. Sturm et al., 1997; Dingman, 2002). While ETC especially increases with increasing water content, we did not find studies that present results for ETC for wet snow. Values for dry (cold) snow are for example summarized by (Fukusako, 1990) and lie below this upper bound.

From the estimated bound for ETC, we can estimate bounds for α using the density of ice ρ_i: $\alpha^u = \lambda_i/\rho_i/c_i = 2.2\text{W/m/K}/917\text{kg/m}^3/2102$ J/kg/K = 1.1*10^{-6}m^2/s.

For the second method of estimating α, the standard optimiziation algorithm in R, the statistical software package (Ihaka and Gentleman, 1996), for one dimensional problems was used and an upper limit of 5e-6 m^2/s was set to the optimization. The upper limit is five times higher than the thermal conductivity of pure ice. We could have used the thermal conductivity of ice as upper boundary, but we wanted to check how often values above this theoretical limit were found by the optimization algorithm. We used this information as a test for the validity of the model assumptions.

Results

We predicted cold content as an indicator for snowmelt with the thermal diffusion model (Eq 5.3). One and five day ahead predictions were calculated and are presented in Figure 5.7. We as-

sumed perfect prediction of air temperatures to test the "best-case" performance of the estimate. Predictions had a root mean square error (RMSE) of 24.4 and 36.1 kJ/m^2 for one and five days, respectively, with an intermediate α of 5*10^{-7} m^2/s.

As a simple reference, we used a persisting temperature profile, for which RMSE was 43 kJ/m^2 for the one day prediction.

The choice of α was not critical since results were not sensitive: RMSE was between 24.37 and 24.39 for six levels of α below the theoretical bounds ($\alpha^u = 1 * 10^{-6}$ m^2/s) for the one day prediction (36.0 to 36.2 m^2/s for the five day prediction).

Even though sensitivity for α was low, we also checked whether improved predictions could be achieved when the value for α was estimated with the data observed two days before the prediction time. However we did not see improved predictions compared to a fixed α (RMSE =24.5 and 36.3 kJ/m^2 for the one and five days, respectively). Compatible with this poor performance is the distribution of the optimized parameter values: only 29% of the α-values were below the estimated upper bound of 1*10^{-6} m^2/s, and 71% were above.

Discussion

Estimates of the current and predictions of future cold content may be of value for flood forecasting. With a RMSE of about 24-36 kJ/m^2, predictions seem sufficiently reliable for use in such a setting. It is surprising that the prediction of cold content is not very sensitive for α within the theoretical bounds. A possible explanation is that the sensor spacing of 15 cm is too large to observe sufficient variation in temperature since snow is a good thermal insulator. Accordingly, temperatures in the snow cover are always close to 0°C (Figures 5.3 and 5.5). Longer periods with very low temperatures would be required for the cold to travel far into the snow cover.

In our study we assumed perfect knowledge about future temperatures. We do not expect large errors because temperature forecasts are very reliable [†].

The limited temperature changes within the snow cover make estimations of α difficult. α is often estimated to be higher than the theoretical upper bound of pure ice, which indicates that the model is an oversimplification and processes other than the thermal diffusion are of importance. This complies with Brandt and Warren (1997) who report that estimated ETC from temperature measurements are generally not very reliable since disturbing processes are hard to avoid.

At least three disturbing processes are of importance: 1) radiation may increase temperature of sensors within the snow cover; 2) so called wind pumping, where air is pressed into the snow cover due to the pressure from the wind, affects temperatures; and 3) multiple mechanisms of heat diffusion within the snow cover due to its complex structure are lumped into a single parameter, which at the same time represents heat transport due to conductivity in ice and air spaces, as well as latent heat flow due to water vapor (Brandt and Warren, 1997). These modes of transport all depend on snow density, snow microstructure, temperature, and water content (Sturm et al., 1997, 2002; Fukusako, 1990). Increasing weathering generally leads to increased thermal conductivity, except for the formation of depth hoar at the base of a snow cover (Sturm et al., 2002).

Another process that is not included in the simple model but is certainly of importance (Figure 5.6) is energy input into the snow cover by rain. As an example, assuming rain with a temperature of 5°C, with the heat capacity of water c_w = 4.19 kJ/L/K (Dingman, 2002) we obtain 21 kJ/L that are available if the rain is cooled to 0°C. For our snow cover we observed a maximum cold content of about 300 kJ/m^2. To heat the entire snow cover to melting temperature, we

[†]Martin Göber, German Weather Service, personal communication, 29. March 2010: 90% of the two day predictions show an error smaller than ±2°C for the case study region

need 300 kJ/m^2 / 21 kJ/L = 14.3 L/m^2 = 14.3 mm. Events of this size are quite common in this catchment.

Due to these oversimplifications and the observed insensitivities of the heat diffusion model, a simpler warning scheme, based on current cold content and weather predictions is likely to be as effective for spring flood warning as the heat diffusion model. Nevertheless, we achieve a reduction of the RMSE of about 30% with the heat diffusion model, if compared to the persistence assumption.

Chapter 6

Spatial distribution of snow. *

The annual spring freshet of snow-dominated watersheds depends mainly on the snow melt volume, which is determined by the spatial distribution of both, the water stored in the snow pack (snow water equivalent; SWE) and melt rates. The aim of this study is to characterize SWE and melt rates with respect to topographic controls and land use at the watershed scale for the Wilde Weisseritz. In order to measure both, the variability at the local scale and the variability at the catchment scale an adjusted sampling designs was used. At the local scale, samples were collected on two perpendicular transects of 60 m length and analysed with geostatistical methods. At the watershed scale, locations of the extensive field campaign were selected according to a stratified sample design to capture the combined effects of elevation, aspect and land use.

At the local scale, the results of the snow surveys during the winter 2008/2009 show that the range of fitted variograms was within the range of sampling design for 80% of the plots. On the catchment scale, the snow height is mainly affected by the plot altitude. The expected influence of aspect and land use was not observed. A temperature-degree day model was applied to test whether the spatio-temporal variability of SWE can be represented by this simple model. The parameters were calibrated with a Bayesian approach. The degree-day model is capable to explain the temporal variability for plots with a continuous snow pack over the entire snow season, if parameters are estimated for single plots. However, processes described in the simple model are not sufficient to represent multiple accumulation-melt-cycles, as observed for the lower catchment. Thus, the combined spatio-temporal variability at the watershed scale is not captured by the model.

*Stefan Lüdtke, Dominik Reusser, Jörn Pagel, Erwin Zehe (short version of a Diplomarbeit, to be elaborated towards a full manuscript),

6.1 Introduction

Snow accumulation and melt are important hydrological processes controlling the water availability at mountainous watersheds. In winter, precipitation falls as snow and is stored until runoff is triggered during the melting period. Hydrological models, representing both, the accumulation and melt period are widely used for the prediction of snow melt (Ferguson, 1999). A general challenge in modelling is to find parsimonious models with sufficient complexity to represent all important aspects (Dunn and Colohan, 1999). Naturally, this is also an issue for snow models that range in their process description from completely empirical to fully physics based and spatial description from lumped to fully distributed (Ferguson, 1999). Models that are to complex suffer from missing parameter identifiability and high computational costs, while very simple models lack generality and transferability to other places (Dunn and Colohan, 1999).

The degree-day model is a parsimonious, empirical model widely used for snow modelling (Ferguson, 1999; Rango and Martinec, 1995). Its basic assumption is that snow melt increases linearly with temperature above the melting temperature, without additional influences. An underlying assumption of the approach is, that other energy sources (e.g. radiation and energy input from rain) are highly correlated to temperature. The proportionality constant is called degree-day factor and is often parametrized as changing slowly over time in order to allow for non-constant melting rates during an entire season. Hock (2003) reports that this model type works well over long time periods but that the spatial variability can not be captured since melt rates vitally depend on topographic properties.

In a previous study, we reported that WaSiM-ETH, a hydrological model that uses a degree-day model as a snow module consistently results in overestimated discharge for snow melt events for the Weisseritz catchment (Chapter 4). The first goal is thus to test whether the simple degree-day approach, besides resulting in overestimated snow melt discharge values, is also insufficient in the description of the spatial variance of snow. We would like to remove the deficiency in the process representation for snow melt events and thus need to better understand what additional factors are required to obtain an adequate process description. The second goal is therefore to test a number of additional factors for their ability to improve the description of the spatial variance of snow. In the current version of the manuscript – which is a short version of a Diplomarbeit – this question is not yet fully addressed. The ultimate goal would be to find a parsimonious model for the description of the spatio-temporal patterns of snow for the Weisseritz catchment that is sufficient to make spatial interpolations from few field observations and that give adequate discharges for snow melt events.

To assess the spatial variability of snow and to identify the most important processes, we need additional observations. The snow water equivalent (SWE) measures the water content of the snow pack and is the most important factor for the amount of melting water during the annual spring freshet. Since the measurement of SWE is time-consuming, a key challenge is to relate field measurements at the plot scale to the distribution of the snow pack at the watershed scale (Lundberg et al., 2010). Measurement of the spatial snow distribution is subject to ongoing research: Watson et al. (2006b) illustrated for a 300 km^2 study area in the Yellow Stone park that random effects on SWE were greatest at small scales and are superimposed by effects from radiation and vegetation at larger scales. Anderton et al. (2004) found redistribution of snow by wind to be the most important influence on snow distribution in a 0.32 km^2 catchment in the Spanish Central Pyrenees. Winkler et al. (2005) investigated the variability between different forest stands ranging from clear cuts to mature forests on an $\approx$1 km^2 study site in British Columbia and found that snow accumulation and melt differ significantly between different stand

types. Also in British Columbia in a 17.4 km^2 watershed, Jost et al. (2007, 2009) used a sampling design that is capable to quantify the variance at both, the plot and the catchment scale. The studies separated the observed spatial variability into effects by topographic and vegetative controls as well as local variability. They report that during a mild and snow rich winter, forests accumulated 39% and 27% less snow than clear-cuts, respectively, because of interception. López-Moreno and Latron (2008) and Hedstrom and Pomeroy (1998) observed different interception capacities over the snow season, depending on the time since the last snowfall, the initial canopy snow load and the different leaf area indices. Since snow in the canopy is exposed to solar radiation, wind and temperature variations, the full interception capacity may be available again during a subsequent snowfall event (Hardy et al., 2004). Local variability was also investigated by Watson et al. (2006a), who illustrated that random SWE variation is significant for short distances (< 10 metres), but decreases for longer distances (10 metres -100 metres) and becomes effectively zero for distances of between 100 metres-1000 metres. Therefore, to examine the autocorrelation structure of SWE, samples need to be spaced closer than 10 m and need to cover a range of close to 100 m. We used a sampling design similar to Jost et al. (2007) in order to assess both, the influence of topography (elevation and aspect) and land use, as well as local random variability on the spatial variability of SWE.

Information from measurements are often included into models with the calibration of parameters. Parameters in hydrological models are uncertain and much discussion is going on about the determination and propagation of such uncertainties (Pappenberger and Beven, 2006; Montanari et al., 2009). Without going into the detail of this discussion, we selected a Bayesian approach to obtain the distribution for each parameter conditional to the data. From these distributions, the uncertainty of the model output can be determined using a Monte Carlo simulation approach. The third goal is thus to make estimates about the level of (un)certainty that we may achieve from the data collected and our model structure. We could then use the framework to test how much field data is required to obtain a satisfying level of certainty about the snow distribution in the catchment.

6.2 Methods

The study consists of four major steps: 1) Field measurements using two perpendicular transects of 60 m length are used to capture local variability, while stratified sampling with elevation, aspect and land use as factors was used to select locations within the catchment. 2) Variability at the local scale was evaluated with empirical and theoretical variograms in an attempt to find factors influencing local random variability. 3) Box plots and linear regression models of mean SWE for each location were used to identify main factors affecting variability at the catchment scale. 4) Parameters for the simple degree-day model were estimated a) using data from each single plot only to assess the ability of the model to represent temporal variability (subsequently referred to as parameters estimated for single plots) and b) using data from all plots simultaneously to test if spatio-temporal variability at the catchment scale can be described (subsequently referred to as parameters estimated for all plots).

6.2.1 Study area

The Wilde Weisseritz watershed is located in the eastern part of the Ore Mountains in Saxony, Germany (Figure 6.1). Elevation ranges from 908 metres to 163 metres a.s.l. However, all plots investigated in this study are located in the upper part (above 400 m) of the watershed. Land use of the Weisseritz watershed consists of fields- and grasslands (45%), forest (34%), settlement (15%) and other usage such as infrastructure and bodies of water (6%). Forests (mainly spruce) with

patches of fields dominate the upper watershed (Figure 6.1), whereas the occurrence of settlements and grasslands increase with decreasing elevation (Pöhler, 2006). Two water reservoirs, Lehnmühle and Klingenberg, in the catchment of the Wilde Weisseritz are used for water supply and flood control. The climate is moderate with mean temperatures of 11°C and 1°C for the periods April - September and October - March, respectively. Annual precipitation for this catchment is around 1100 mm/year. During winter, the catchment usually has a snow cover of up to about 1 m for 1 to 4 months with high flows during the snow melt period. Data about land use and a digital elevation model was obtained from the state office for environment and geology (LfUG, 2007). Data from a web cam located in Holzhau† (Figure 6.1) was retrieved daily and stored in order to visually assess development of the snow cover.

6.2.2 Field measurements

Snow depth and weight were measured during five campaigns between January and March 2009 (15 - 16 January; 29 - 30 January; 12 -13 February; 26 - 27 February and 26 - 27 March). The study by Jost et al. (2007) guided the design of our measurement campaign.

At the local scale a sampling design with two perpendicular transects of 60 m length was used (Figure 6.2). At each site, snow depth measurements and the more time consuming snow weight measurements were performed 61 and 13 times with a spacing of 2 metres and 10 metres, respectively. Heights were measured with a plastic tube of 50 mm diameter and 1.5 m length. A balance with maximum weight of 5 kg and a resolution of 5 g (MH5K5 and HDB5K5 by Kern) was used for weight measurements.

At the catchment scale, a stratified sampling design was used to assess the influence of topography, aspect and land use. Plots were selected

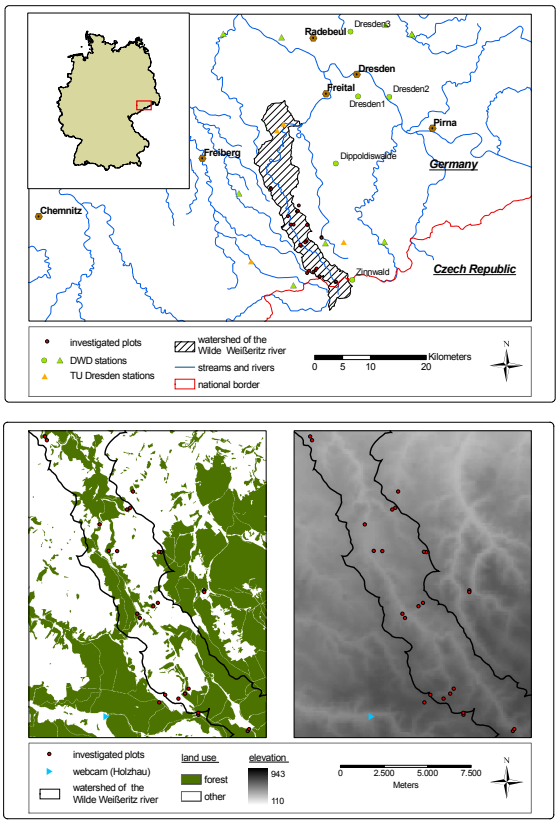


Figure 6.1: Top: location of the Wilde Weißeritz watershed including main streams, investigated plots (red circles) and meteorological stations. Green circles mark fully equipped stations (DWD), triangles mark stations measuring precipitation only (DWD and University of Dresden).
Bottom: location of snow survey plots (red circles) including land use (left hand side) and a DEM (right hand side). The location of a webcam is shown with a blue triangle and the border of the Weißeritz catchment with a black line. River net data, the DEM and the land use map were obtained from LfUG (2007).

†http://www.holzhau.de/webcam-holzhau.html

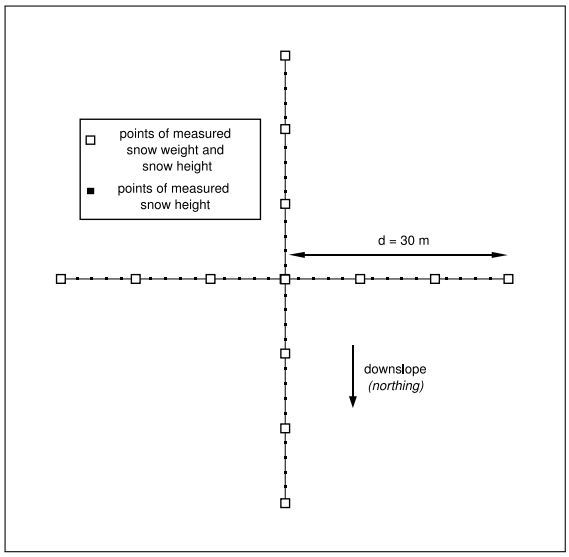


Figure 6.2: Sampling design at the plot scale as suggested by Jost et al. (2007). The black small squares indicate the points where snow heights were measured. Snow weight was measured in addition at location indicated by the bigger hollow squares.

from three elevation zones (zone 1: 400 m-600 m, zone 2: 601 m-700 m and zone 3: 701 m-800 m) and with differing aspect (north, east, south or west) and land use (forest and field). The aspect was converted to a northing index (Jost et al., 2007) with a value of one and zero for plots facing north and south, respectively. Plots facing either west or east have both a value of 0.5. A total of 25 sites were investigated. (Table 6.1 and Figure 6.1).

6.2.3 Post processing of field data

The statistical software R (Ihaka and Gentleman, 1996) was used for most calculations. The *zoo* package was used to work with time series (Zeileis and Grothendieck, 2005).

Mean and standard deviation were computed for the SWE for each plot and snow campaign. Snow weight/snow height-pairs of each snow campaign from all locations were used to calculate the snow

Table 6.1: Overview of the topographic properties of the snow survey plots. Elevation is reported in m a.s.l. and the aspect is 1 and 0 for plots facing north and south, respectively, while plots facing east and west have a value of 0.5 (section 6.2.2 for more details).

ID	Elev.	Land Use	Aspect
Elevation Zone 1			
1	460	forest	0.82
2	510	field	0.75
4	520	field	0.57
5	520	field	0.97
6	540	forest	0.59
7	550	field	0.54
8	550	forest	0.01
9	570	field	0.01
Elevation Zone 2			
10	630	forest	0.29
11	630	field	0.77
12	630	field	0.65
13	630	forest	0.65
14	630	field	0.18
15	650	forest	0.12
16	670	field	0.93
17	670	forest	0.75
Elevation Zone 3			
18	710	forest	0.87
19	700	field	0.96
20	760	forest	0.88
21	750	forest	0.01
22	760	field	0.17
23	780	field	0.97
24	750	field	0.75
25	780	forest	0.67
26	800	field	0.25

density, assuming a constant snow density for the entire catchment. Snow heights were then transformed to SWE using the mean snow density and the density of water of 1 kg/m^3.

To assess variability at the local scale, experimental variograms (Akin and Siemes, 1988) were calculated with a variable bin width of 2 m for short distances and about 5 m for long distances resulting in >58 data points in each bin. Theoretical variograms were calculated with the *autofitvariogram* function (automap package Hiemstra et al., 2008) in R. A number of variogram models (spherical, Gaussian or an exponential model) is fitted and the model with the lowest root mean square error (hereinafter RMSE) between experimental and theoretical variogram is selected. Nugget, sill and range were calculated from the theoretical variogram. The sill is the semivariance at which the variogram levels off, while the nugget represents the variability at distances smaller than the sample spacing. The range is related to the lag distance at which the variogram reaches the sill value and depends on variogram model. A nugget effect was assumed if the difference between nugget and sill was smaller than 2 kg/m^2, corresponding approximately to the measuring accuracy for SWE (1 cm resolution in snow height and an approximate density of 200 kg/m^3). A nugget effect was also assumed if the range was smaller than two metres, which is the sample spacing.

At the catchment scale, box plots and multiple linear regression models were used to investigate the influence of time, elevation, northing and land use on SWE. A two-sided t-test was used to check for significance of the parameters of the regression model.

6.2.4 Meteorological data

Meteorological data was obtained from the German Weather Service (DWD, 2007) and the University of Dresden (TU-Dresden, 2010). Daily sums of precipitation was available for stations maintained by the University of Dresden (triangles in Figure 6.1). Five DWD stations recorded air temperatures at different temporal resolutions, daily snow heights, SWE, and precipitation (circles in Figure 6.1) and six DWD stations (triangles) recorded precipitation only. All stations were equipped with heated rain gauges providing daily sums of precipitation over the entire year.

Data from the two DWD stations Zinnwald (877 m a.s.l.) and Dippoldiswalde (365 m a.s.l) is presented in figure 6.3. The top panel shows SWE and accumulated precipitation for both stations, Zinnwald in black and Dippoldiswalde in red. The bottom panel shows mean daily temperatures and rainfall - note that precipitation is not shown for temperatures below freezing. This representation was chosen to better illustrate the relation ship between snow melt and rain-on-snow events.

The figure shows that cumulative precipitation is only about 70% of the measured SWE at the Zinnwald station for the time of the snow maximum. It is also evident, that snow melt at Dippoldiswalde was accompanied by increasing air temperatures and rain-on-snow events. Finally, a significant elevation gradient exists for rain as expected (cumulative precipitation is 267.9 mm and 482.7 mm for the lower and higher station, respectively), while the temporal dynamics are similar. For temperature, an elevation gradient was also observed, thus, interpolation for the degree-day model was performed for both quantities height dependently. Temperature data from the five DWD stations was interpolated using a simple linear regression model for each day with mean daily temperatures depending on elevation. Similarly, precipitation from all fifteen stations was interpolated with a linear regression against elevation, if the correlation coefficient of the model was $R^2 \geq 0.6$. Subsequently, inverse distance weighted (IDW, with an exponent of 2) residuals were added to the results of the linear regression model. If $R^2 < 0.6$, the precipita-

tion for was interpolated with simple IDW. The algorithm is called "interpol v3" and was described in detail by Francke (2002) and Kneis and Heistermann (2009).

6.2.5 Degree-day model

The degree-day model is based on a linear relation between snow melt and temperature above the melting temperature. The proportionality constants the called degree-day factor $c0$ in mm/day $^\circ$ C. For the accumulation period, an additional, dimensionless parameter $a0$ is introduced as a constant, multiplicative factor for precipitation. An error model for the Bayesian estimation is introduced that draws an error term ϵ from a normal distribution with zero mean and standard deviation σ, an additional parameter of the error model required for the Bayesian parameter estimation.

$$S_t^{sim} = S_{t-1}^{sim} \begin{cases} +a0 * P_t & \text{if } T_t \leq T_{crit} \\ -c0 * (T_t - T_{crit}) & \text{if } T_t > T_{crit} \end{cases}$$
$$S_t^{sim} = S_t^{obs} + \epsilon_t \tag{6.1}$$

where *t* denotes the time step, T_t the temperature , P_t the precipitation and S_t^{sim} the modelled snow water equivalent at time step *t*. $T_{crit} = 0°C$ is the melting temperature, S_t^{obs} is the observed SWE and ϵ_t is the model error at time t.

Parameters are estimated twice, first separately for single plots, to check if the temporal variability for a single plot is captured by the simple model and second for all plots at the same time, testing if spatio-temporal variability of the entire catchment can be explained.

As initial condition, SWE was set to zero on November 1^{st} 2008. In addition to the data from the measurement campaigns, SWE was set to zero for the last ten days of the simulation period (ending on April 30^{th}), based on webcam pictures from a station at 600 m.a.s.l showing no snow between April 10^{th} and 30^{th}. A time step of one day was used.

6.2.5.1 Bayesian parameter estimation

The objective of Bayesian parameter estimation is to determine how a probability distribution for a vector of model parameters $\vec{\theta}$ changes given the data D (Gelman et al., 2003). Prior knowledge about the parameter vector is expressed as marginal distribution $P(\vec{\theta})$. The relationship between the parameters and the data is given as conditional distribution $P(D|\vec{\theta})$, which is called a likelihood function if it is regarded as function returning the probability of D dependent on the parameters $\vec{\theta}$. Bayes' theorem states how the desired posterior distribution $P(\vec{\theta}|D)$ is related to the likelihood function and the prior (Gelman et al., 2003):

$$P(\vec{\theta}|D) = \frac{P(D|\vec{\theta}) * P(\vec{\theta})}{P(D)} \tag{6.2}$$

The prior probability or marginal probability of the data $P(D)$ acts as normalizing constant and is generally only determined after the parameter estimation by normalizing $P(\vec{\theta}|D)$ to one.

In the current application, D includes the observed SWE, while $\vec{\theta}$ consists of the two model parameters *a0*, *c0* and the parameter σ of the error model.

To define $P(\vec{\theta})$, normal distributions with zero mean and standard deviation of 10 were used as non-informative prior distributions for $a0$ and $c0$ while for σ the prior probability was set to 0 for all values.

After updating with the observed SWE, the posterior distribution $P(\vec{\theta}|D)$ has an expectation of $E(\vec{\theta}|D)$, reflecting the information gain from the data. Similarly the variances Var$[\vec{\theta}]$ and Var$[\vec{\theta}|D]$ indicate the uncertainty about the parameter before and after updating. Change of the variance may be regarded as a reduction of uncertainty caused by the data.

The posterior distribution $P(\vec{\theta}|D)$ is analytically intractable. Therefore we used Markov Chain Monte Carlo techniques based on the Metropolis-Hastings algorithm (Chib and Greenberg, 1995)

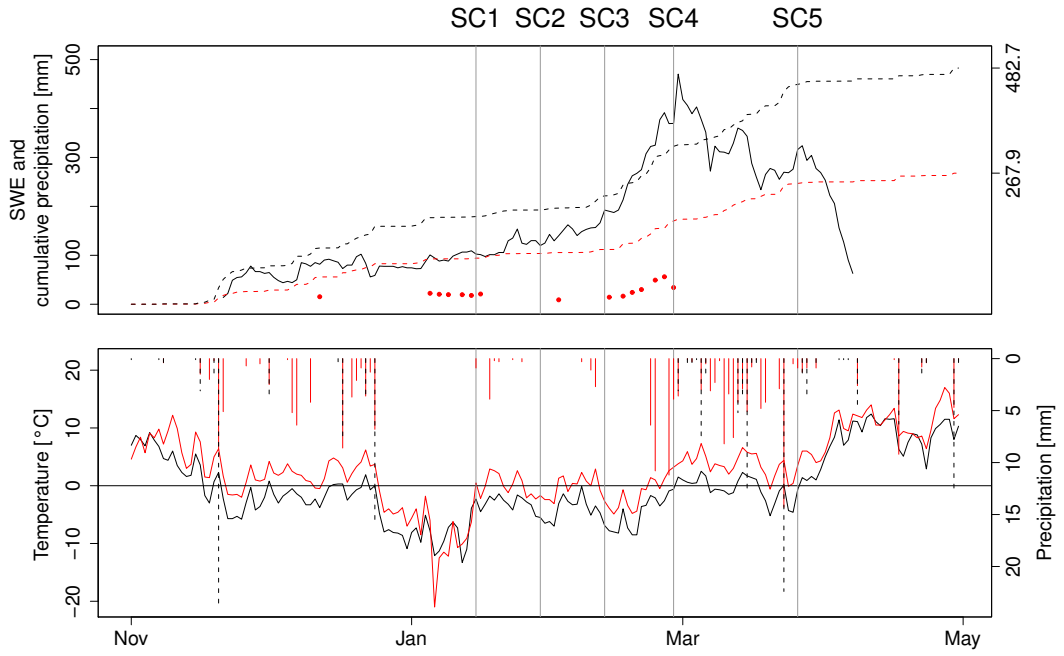


Figure 6.3: Data from the DWD stations in **Dippoldiswalde** (red lines and points) and **Zinnwald** (black lines). The vertical grey lines refer to the dates were the snow surveys took place. The continuous black line and the red points in the upper panel show the measured SWE. The dashed lines show the cumulative precipitation, both SWE and precipitation are given in mm. Both continuous lines in the lower panel show mean daily temperatures. Daily sum of precipitation for days with temperatures $> 0°C$ are shown from the top (right hand side y-axis). All data from DWD (2007).

for the sampling. Chain convergence was tested with the trace plot for each parameter as well as the Gelman-Rubin diagnostic (Gelman et al., 2003; Plummer et al., 2009). This diagnostic defines the potential scale reduction factor (*PSRF*) as the ratio between the between chain variance and the mean within chain variance W, which is close to 1 for converging chains:

$$PSRF = \sqrt{\frac{\overline{var}(m)}{W}} \tag{6.3}$$

with $\overline{var}(m)$ being a weighted mean of the within chain and the between chain variance. Chains with a value above 1.1 were considered as non-converged (Gilks and Richardson, 1995).

6.2.5.2 Model evaluation

As no additional data was available, the model evaluation was performed with the same data as used for the calibration from winter 2008/2009, using two objective functions. First, the RMSE between the measured and the modelled SWE using the distribution mean for $a0$ and $c0$ was determined. Second, a Monte Carlo simulation with 1000 runs was performed by sampling from the posterior parameter distributions for $a0$ and $c0$. Subsequently a random error from a normal distribution with zero mean and standard deviation σ was added. Confidence bands (CB_{MC}) were calculated from the central 80% of the Monte Carlo runs.

The mean and standard deviation of the measurements at each plot were used to calculate a measurement confidence band (CB_{obs}) including one standard deviation around the mean. An overlap of the two confidence bands CB_{MC} and CB_{obs} was counted as successful prediction. The rate of successful predictions was determined.

Because parameters estimated for all plots did not successfully represent spatio-temporal dynamics (see section 6.3.2) and no further data on SWE at single plots was available, an additional, formal model validation was not conducted.

6.3 Results

6.3.1 Snow variability at the plot scale

Exemplary SWE data for plot 24 during 3 campaigns are shown in figure 6.4 (top). The green (x-axis) and blue (y-axis) line correspond to the horizontal and vertical transect, respectively. The data shows the short range, random variability as well as a developing non-random spatial structure over time. The red circles indicate locations of snow weight measurements. The same data points are also highlighted in red in the bottom panel, which illustrates the relationship between measured snow height and snow weight for the three campaigns. Density estimates are summarized in Table 6.2. Density is increasing until the fourth snow campaign and decreasing slightly to the fifth snow survey. The correlation coefficients is always ≥ 0.89.

To assess the local variability of SWE, experimental variograms were calculated and used to match theoretical variograms. We will not show all variograms, but illustrate important findings using three exemplary plots (with ID 9, 21 and 26) for all five campaigns (Figure 6.6).

For all three plots (and in general) we observe increasing variance over the snow season and the highest values are observed for the fourth or fifth snow campaign. A pure nugget effect (difference between nugget and sill smaller measurement resolution of the tube or range smaller than the sampling interval), as present for plots 9 and 21 during the first campaign, was generally observed more often at lower elevations and at the beginning of the snow season. In total, a pure nugget effect was observed for 15 % of the variograms. For plot 8, variograms could not be estimated (not shown) since logging prohibited use of the local sampling design and a random sampling pattern was used instead. Figure 6.6 also shows, that the type of the best matching theoretical variogram, as well as the

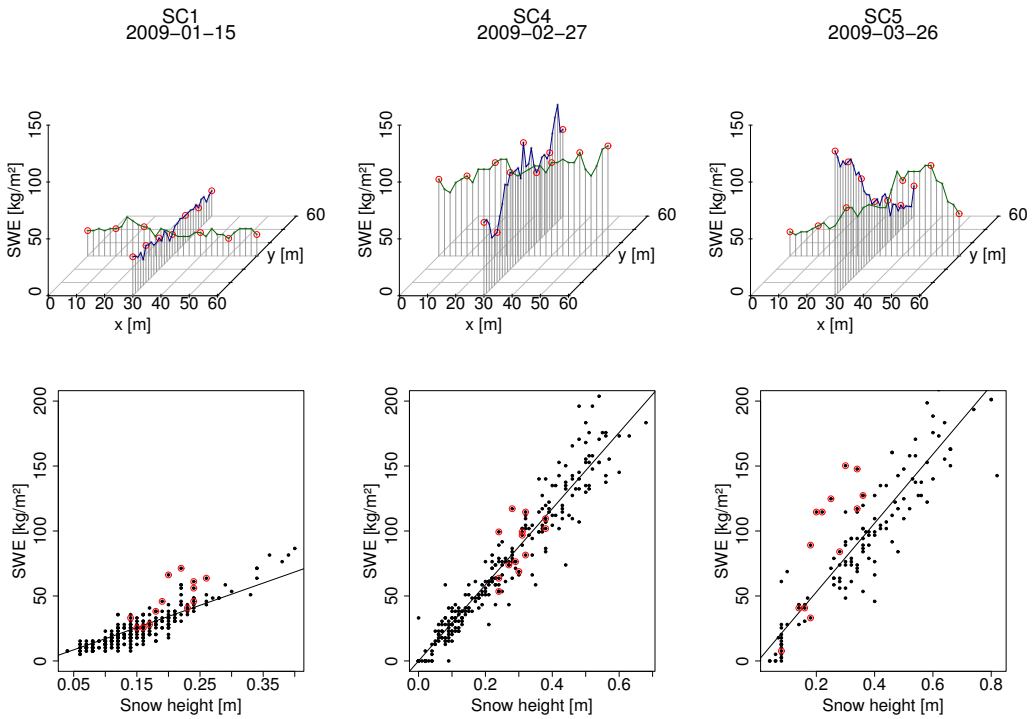


Figure 6.4: Upper panel: SWE data (in kg/m^2) at plot 24 for three snow campaigns. Horizontal (x-axis) and vertical (y-axis) transects are indicated by the green and blue line, respectively. Lower panel: linear regressions used to estimate snow density including data from all plots. Red circles indicate data from plot 24 (also shown in the upper panel).

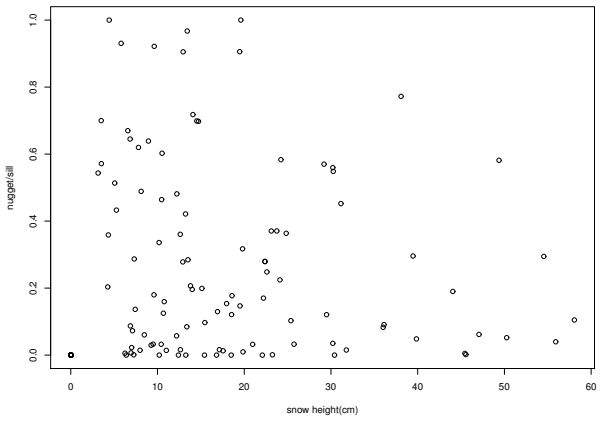


Figure 6.5: Nugget-to-sill-ratio for spherical variograms for all plots and all campaigns plotted as a function of snow height.

estimated range vary considerably and without apparent pattern over time. Ranges calculated from the theoretical variogram models are smaller than 60 metres (the extent of the sampling design) for 65 % of all plots (or 80% if plots with a pure nugget effect are included). For about 20 % of the investigated plots, a range > 60 m was found.

A nugget-to-sill-ratio of close to one (pure nugget) was observed for snow heights < 20 cm only (Figure 6.5. The Figure shows the nugget-to-sill-ratio for all plots and all campaigns plotted as a function of snow height. For this analysis, parameters were determined for spherical variograms for all plots in order to increase comparability.

Excluding cases where the range is >60 m, we find anisotropy for >60% of the cases if we define anisotropy as a ratio of the range in x direction and the range in y direction of >1.2 or <0.83 (data not shown). The same holds true if the ratio of the sill is considered, indicating that we are observing zonal anisotropy. The direction of the anisotropy appears to vary randomly.

6.3.2 Snow variability at the catchment scale

Mean SWE are generally increasing with increasing elevation (Table 6.2 – plot IDs are ordered by altitude). No measurements were taken at plot 13 during the first campaign because of logging. While plots at lower elevations were snow free again during the fourth snow campaign, maximum SWE at higher elevation was not reached until that time. Similarly, a continuous snow pack existed over the entire season in Zinnwald, while recurring short term snow accumulation and melting cycles can be observed for the lower station Dippoldiswalde.

Boxplots (Figure 6.7) are used to assess variability of snow at the catchment scale. Dependence on elevation, northing and land use was tested. Forest and field plots are shown in green and red, respectively. Differences tend to be significant if interquartile ranges (IQR, a measure for variability) in boxplots do not overlap. For elevation zone 1 (400 to 600 m a.s.l.) non overlapping IQR's for forest and field plots occur during the first and third snow survey. No snow was observed for the fifth campaign and the SWE for surveys two and four were slightly lower compared to the previous surveys. For elevation zone 2 (601 to 700 m a.s.l.) we find higher median SWE compared to zone 1 (note that the range of the y-axes are different). SWE are also lower for forest plots compared to fields. The IQR () is increasing over time for both land uses until the third and fourth measurement campaign. For elevation zone 3 (701 to 800 m a.s.l.), SWE are again higher compared to lower elevations. A difference in magnitude between field and forest plots is not apparent for the first three surveys. For the fourth and fifth campaign, SWE are somewhat higher for forest plots, which is opposite to the effect of land use at lower elevations. The IQR is higher for forest plots. While SWE monotonically increases for higher elevations until the fifth campaign, it fluctuates over time non-monotonically for lower elevations. For Zinnwald (Figure 6.3),

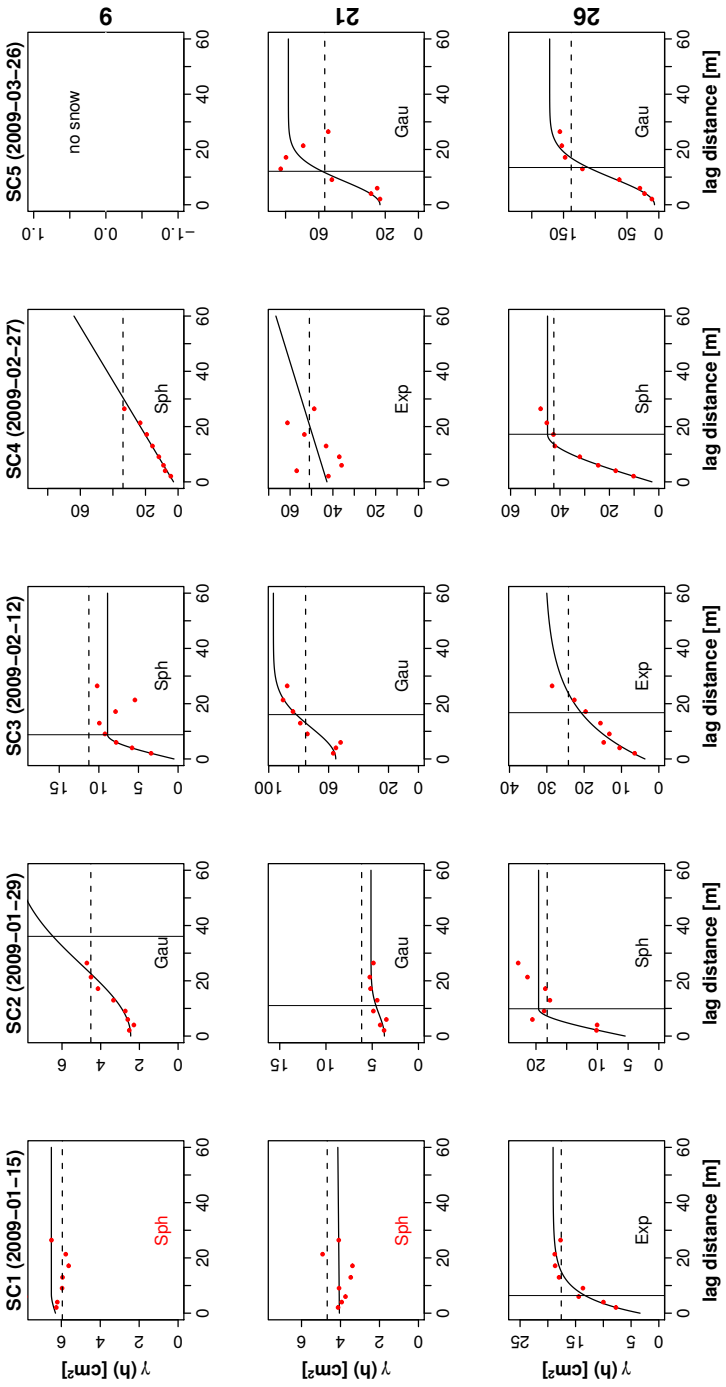


Figure 6.6: **Empirical** variograms (red points), the fitted variogram **models** (black line) and the variance at the plot scale (dashed line). The vertical line refers to the **range** of the theoretical variogram. The model type is indicated in the plot: "Gau" = Gaussian, "Sph"= spherical and "Exp"= exponential. Model names in red indicate plots with a **pure nugget** effect. The columns refer to the snow survey (with the dates in parentheses), the rows to the plot *ID* appended on the right hand side.

Table 6.2: Mean SWE in kg/m^2 for each measurement campaign. Densities in kg/m^3 were calculated from snow height/snow weight pairs from the entire catchment. Plot IDs are ordered by increasing altitude.

ID	Mean1	Mean2	Mean3	Mean4	Mean5
1	11	0	9	0	0
2	21	22	21	0	0
4	18	9	18	0	0
5	25	12	19	21	0
6	18	12	19	26	0
7	22	0	14	0	0
8	12	0	13	0	0
9	23	26	24	21	0
10	26	25	38	42	21
11	29	34	36	71	18
12	23	20	22	25	17
13	NA	11	25	45	9
14	23	21	26	22	14
15	12	7	15	13	8
16	32	35	59	65	25
17	17	16	26	20	19
18	31	51	86	147	154
19	36	44	71	117	121
20	24	34	46	112	80
21	22	28	57	133	105
22	34	45	47	106	84
23	34	47	49	88	81
24	24	31	38	74	53
25	40	45	61	145	144
26	32	45	58	138	148
Density	171.21	199.21	196.17	293.02	265.08
R^2 (Density)	0.89	0.91	0.94	0.96	0.91

the first three snow surveys took place during the accumulation period, the fourth survey was conducted around peak snow accumulation and the fifth survey during the melting period. No influence of northing on SWE could be detected from box plots (results not shown).

To complement the results from the box plots, multiple linear regressions between SWE and elevation, northing and land use were calculated. Results between the two methods are in agreement: While the influence of elevation was always significant (Table 6.3), northing and land use never had a significant effect. Correlation coefficients of the multiple linear regression models were between 0.64 and 0.76.

6.3.3 Degree-day model

6.3.3.1 Convergence of the MCMC parameter estimation

Convergence of the MCMC parameter estimation for the degree-day model needs to be checked. The potential scale reduction factor (PSRF) as convergence criterion is shown in table 6.4. Based on the PSRF 13 out of 25 parameters estimated for single plots as well as parameters estimated for all plots converge. However, trace plots (e.g. Kass et al., 1998) indicate, that chains did not converge for plots 6 and 11 despite a low PSRF value. Ten out of the remaining 11 plots with converging chains are higher than 670 metres a.s.l. The standard deviations (SD) of the posterior distributions as a measure of remaining uncertainty is also presented in table 6.4. As expected, SD is much smaller (mean(SD($a0$))=0.03, mean(SD($c0$))=0.22) for converging chains compared to non converging chains (mean(SD($a0$))=0.83, mean(SD($c0$))=5.6) A possible reason for the non convergence of the parameter estimation in the lower catchment is presented in the discussion and is based on the observation that all plots with non converging chains show a high mean value for degree-day factor > 10 mm/d $^\circ$ C, except for plots 10 and 16. Since pa-

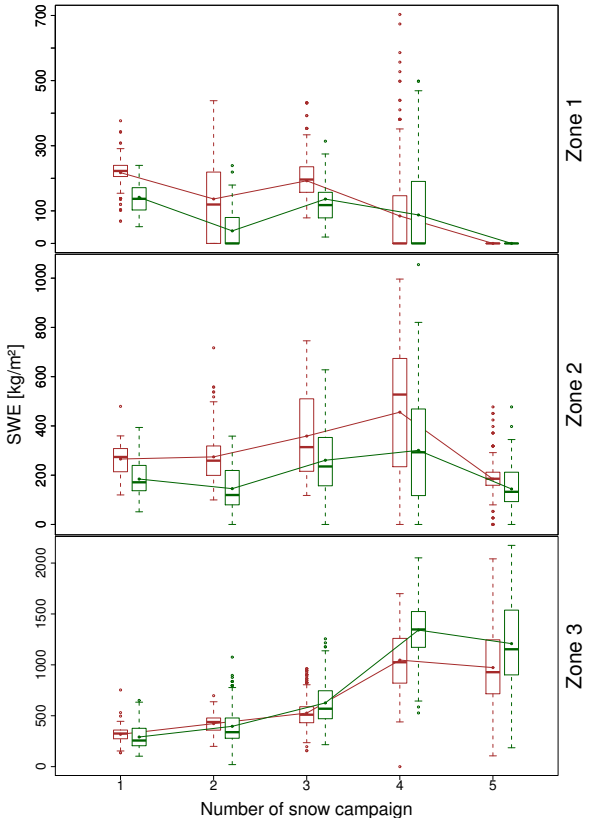


Figure 6.7: Box plots of SWE for three different elevation zones (zone 1: 400 m-600 m, zone 2: 601 m-700 m and zone 3: 701 m-800 m; table 6.1) and the two different land uses (forests green, fields in dark red). Mean SWE for each group are indicated by points.

Table 6.3: Parameters of the multiple linear regressions between SWE and topographic controls (**elevation (a)**, **northing (b)** and **land use (c)**). For each snow campaign (**SC**) the estimated coefficient (**Est**) and its standard error (**Error**) are listed. Significant factors (two sided t-test; $\alpha \leq 0.01$) are printed in bold. The last row shows the correlation coefficients of the regression model

	SC1		SC2		SC3		SC4		SC5	
Predictor	Est1	Error1	Est2	Error2	Est3	Error3	Est4	Error4	Est5	Error5
Intercept	-10	7	**-62**	12	**-75**	19	**-248**	38	**-254**	45
Elevation (a)	**0.05**	0.01	**0.13**	0.02	**0.16**	0.03	**0.45**	0.06	**0.44**	0.07
Northing (b)	5.7	3.0	9.3	5.4	14.0	8.3	21	17	18	20
Landuse (c)	-5.1	2.0	-6.7	3.6	0.7	5.4	8.9	11.0	9.1	13.0
R^2	0.59		0.70		0.59		0.73		0.63	

rameter estimates of non converging chains are not reliable, we will focus on the plots in the upper catchment with converging chains for the remaining presentation of the results. Note that non convergence is an indicator that model formulation is not appropriate – for possible extensions, see discussion.

6.3.3.2 Parameter dependence

We tested whether estimates for $a0$ or $c0$ showed any dependence on elevation, northing or land use with visual plots and linear regression models. As expected no such dependence structure was apparent with respect to elevation, indicating that interpolation of meteorological data is sufficient to represent the observed elevation gradients. For aspect, no dependence was found as well, which for $c0$ is against expectation because differences in radiation input should result in different melting factors. For land use, forest plots show between 13% and 30% higher accumulation parameters $a0$ compared to adjacent field plots (table 6.4 – the same subscripts indicate adjacent plots). Similarly, for the degree-day factor, consistently higher values are found in forests compared to adjacent field plots, differences ranging from 3 to 43%. Plots with small differences between the accumulation factors do not necessarily show small differences between the degree-day factors and vice versa.

6.3.3.3 Model evaluation

The model is only evaluated for plots with converging chains. The RMSE are reported in table 6.4. RMSE range from 1.2 to 17 for parameters estimated for single plots, while values go up to 91 for parameters estimated for all plots. As expected, the model error σ and the RMSE show high correlation R^2=99.9%.

We found an overlap of the confidence bands (CB_{MC}) and (CB_{obs}) for all five campaigns for 9 out of 11 plots with converging chains (table 6.4). Confidence bands are also illustrated in figure 6.8 for plots 6 (non-converging), 24 (best case), and 26 (typical). For plot 6, the very wide confidence bands indicate the non-convergence during the parameter estimation. The upper limit is constantly increasing (not shown) with a melting factor $c0 < 0$. For the other two plots, the model represents temporal dynamics sufficiently as indicated by overlapping confidence bands for all 5 campaigns. The pattern observed for plot 26 was also apparent for multiple other plots: the SWE is overestimate for the first three snow survey, while it was underestimated for the last two snow surveys.

Results are shown also for parameters estimated

Table 6.4: Results for parameter estimation and model evaluation. Chain convergence is assessed with the potential scale reduction factors (PSRF) and the standard deviation of the posterior distribution (SD). Gray coloured rows indicate convergence of chains (PSRF and trace plot). Mean values for the estimated parameter distribution are presented independent of convergence. The root mean square error (RMSE) and **P** value (fraction of overlapping confidence bands of model and measurements) are presented for model evaluation. P and the RMSE were only computed for converging chains, non convergence is indicated by empty entries. The first two column show plot ID (Table 6.1) and land use. The subscripts (a, b, c, d) indicate plots situated close to one another. *ALL* stands for parameters estimated for all plots.

ID	Land use	Convergence						Parameters			Evaluation			
		$a0$		$c0$		σ		$a0$	$c0$	σ	single		all	
		PSRF	SD	PSRF	SD	PSRF	SD				RMSE	P	RMSE	P
1	forest	1	0.09	1.16	5.71	1	0.18	0.61	12.21	1.69			4.75	100
2	field	2.03	6.55	1.34	6.99	1.28	0.23	-4.57	10.47	7.92			14.18	100
4	field	1.06	3.11	1.18	5.28	1.01	0.23	0.13	17.77	4.36			11.13	100
5	field	1.23	0.22	1.26	3.67	1.03	0.19	1.25	14.06	4.63			12.36	100
6	forest	1.03	0.25	1.02	4.56	1.01	0.19	0.64	4.75	4.02			9.86	100
7	field	1	0.01	1.78	4.35	1	0.18	1.05	27.78	0.22			12.26	100
8	forest	1.02	0.05	1.15	3.55	1	0.18	0.72	19.71	1.14			9.97	100
9	field	1.31	0.32	1.33	8.6	1.03	0.2	0.9	16.09	7.33			14.08	100
10	forest	1.88	0.16	1.98	4.16	1.17	0.21	0.42	1.99	6.73			18.02	100
11	field	1.08	0.21	1.08	6.56	1	0.19	0.97	11.96	5.78			24.96	100
12	field	2.42	0.15	2.44	6.47	1.02	0.2	0.93	25.79	5.23			12.54	100
13	forest	1	0.02	1.03	0.12	1	0.2	0.4	0.9	1.8	3.23	100	9.63	100
14	field	1.09	0.27	1.11	11.33	1.03	0.25	0.89	25.85	5.2			13.12	100
15	forest	2.24	0.11	2.03	5.87	1	0.2	0.48	16.86	3.32			10.82	100
16	field	2.23	0.08	5.04	0.89	1.2	0.2	0.57	1.44	7.75			30.46	80
17	forest	1	0.03	1	0.11	1	0.19	0.18	0.24	4.16	7.1	100	12.6	100
18_a	forest	1	0.05	1	0.21	1	0.19	0.79	0.86	9.94	17.01	100	91.42	40
19_a	field	1	0.02	1	0.1	1	0.19	0.64	0.61	5.14	8.78	80	70.86	0
20_b	forest	1	0.03	1	0.22	1	0.19	0.62	2.23	3.31	5.57	100	40.58	60
21_c	forest	1.01	0.06	1.01	0.46	1	0.19	0.69	1.84	8.14	13.72	80	58.23	60
22_c	field	1	0.01	1	0.12	1	0.19	0.59	1.79	1.81	3.04	100	42.03	60
23_b	field	1.01	0.02	1.01	0.18	1	0.19	0.47	1.38	2.43	4.08	100	31.14	100
24_b	field	1.01	0.01	1.01	0.04	1	0.19	0.43	1.27	0.73	1.23	100	23.81	100
25_d	forest	1.01	0.05	1.02	0.45	1	0.19	0.69	1.5	7.14	12.07	100	68.1	40
26_d	field	1	0.04	1	0.43	1	0.18	0.6	1.27	8.41	14.38	100	59.73	60
ALL		1.01	0.02	1	0.29	1	0.03	0.31	1.46	19.01				
mean											8.3		28.3	

for all plots (figure 6.8, right hand side). It is apparent that confidence bands are wider compared to parameters estimated for single plots, which is caused by the larger width for the distribution of the error model (σ). It becomes also clear, that in the upper catchment, the model is able to represent the dynamics for one location (24) and at the same time fails to represent the dynamics for the other location (26), indicating missing processes (see discussion). SWE for plots at higher elevations are often underestimated (not shown).

6.4 Discussion

6.4.1 Snow density

Snow density is expected to increase with time as a consequence of compression and metamorphic processes caused by weather and the weight of the snow cover (e.g. Gabel, 2000). Thawing and refreezing cycles further increase the density of snow over time. In agreement with these expectations, we observed increased snow density during the accumulation period.

Jonas et al. (2009) and Anderton et al. (2002) reported increasing snow density with increasing snow depth. In contrast, we assumed snow density to be independent of snow height, since we were unable to obtain sufficiently accurate weights in cases of low snow masses (< 100 g). The reason was the low resolution (5 g) of the balance used in this study since balances with higher resolution did not have a sufficiently large maximum measurable weight. Using a more precise balance would make it possible to compute snow height dependent densities to compute the SWE.

6.4.2 Snow variability at the plot scale

The sampling design was expected to capture the variability at the plot scale. Jost et al. (2007) used the same design and found, that besides a few exceptions the extend is sufficient to estimate reliable local means. In our study, for 80% of all investigated plots the method is covering the autocorrelation structure, as indicated by a range $<$60 m or a pure nugget effect. The increasing variance over time, that we found for most plots indicates the ripening of the snow cover, as small scale local effects are likely to accumulate over time. We were able to identify anisotropy in more than half the plots, which is accordance with expectations as processes differ depending on the direction (gravitational forces acting in the down slope direction). We were not able to identify patterns affecting the anisotropy.

An initial goal of the study was to find factors influencing properties of the local variability of snow, such as the range, the anisotropy or the nugget-to-sill-ratio. However, we were not able to identify any such factor. Reasons for differences of snow variability at the plot scale have been reported before, and some might be of importance in our catchment: For example, Hiemstra et al. (2006) reported differences in snow heights between 0.1 and 7 metres over a span of a few metres caused by wind redistribution. We attempted to record indicators of wind transportation during the snow surveys, but since no uniform criteria had been defined before, the data was not consistent for all plots and snow surveys. The snow distribution in forests is influenced by shading effects and the interception capacity, which is affected by canopy density and leaf area index (Hedstrom and Pomeroy, 1998; López-Moreno and Latron, 2008). Future investigations should therefore select locations with homogeneous crown closure and forest types, which was not alway given for our plots. In addition, not enough care was taken to find exactly the same location for each measuring campaign, often the center of the sample design was shifted by a few meters. Comparison between campaigns thus also includes effects from such shifts. Nugget only effects for plots with little snow is likely to be a result of surface discontinuities. If the latter are high compared to the snow height, the variance of SWE is superimposed by irregularities of

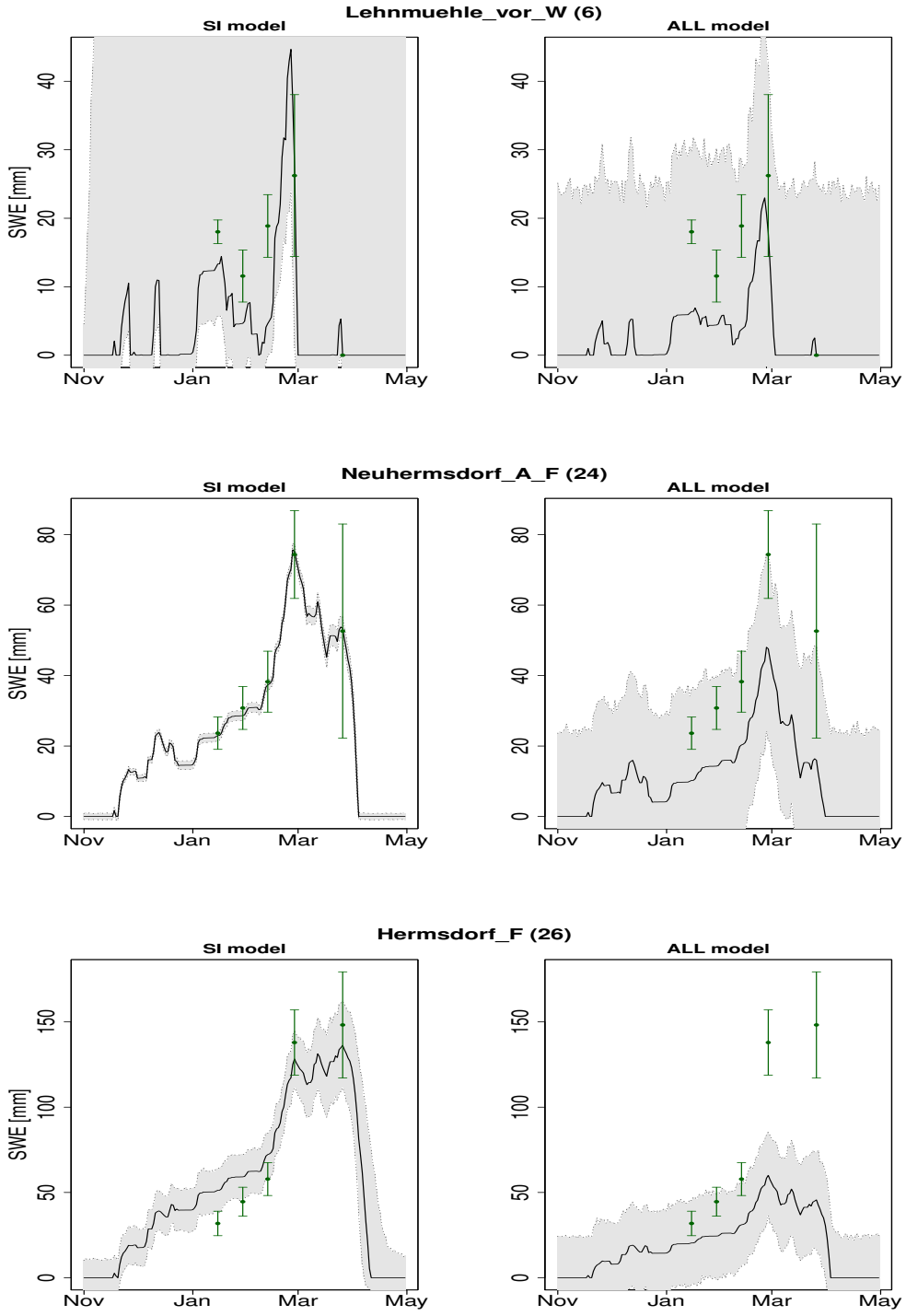


Figure 6.8: Time series of the simulated **SWE [mm]** for the plots 6 (non converging chains), 24 (best fit) and 26 (representative example). The black solid line shows the SWE estimates based on the mean parameter value, while the grey area represents the confidence band (10- and 90% percentile) of the modelled SWE, including the term from the error model. The green error bars show the mean SWE measured at the plot scale and the associated standard deviation. Results for parameters estimated for single plots are shown in the left column, while results for parameters estimated for all plots are shown on the right hand side.

the surface. Three dimensional laser scans at the plot scale before and during the snow season may help to quantify the relation between surface discontinuities and variability of SWE as reported by Hood and Hayashi (2010). In addition, estimated variograms for the assessment of the local variability will certainly improve using such high resolution data.

We did not expected SWE to exceed the cumulative precipitation. In contrast, we found that cumulative precipitation was only 70% of the SWE in Zinnwald, at the time of maximum SWE. We found two possible explanations: (a) The measurement error of rain gauges is higher for snow fall and high wind speeds, resulting in underestimations of the true precipitation amount (Richter, 1995). For future studies, the error of the rain gauge should be corrected. (b) Strong winds and patterns from wind ablation and redistribution were observed during the snow surveys, indicating that snow was possibly transported to the Zinnwald station. Further investigation should therefore quantify the magnitude of wind redistribution and ablation.

6.4.3 Snow variability at the catchment scale

Jost et al. (2007) reported plot elevation as the most important factor for SWE at the catchment scale and we expected similar results. Box plots as well as the statistical model confirmed elevation to be the most influential factor. The influence, is increasing until the time of peak snow accumulation.

Because of the snow interception in the canopy, lower SWE were expected for plots in forests compared to fields for the accumulation period. During melting periods, we expect lower melting in forest because of shading and lower wind speeds (Ferguson, 1999). Thus the difference from the accumulation period is expected to level out or even turn around, showing more snow at forested plots. As expected, less snow was observed in forests compared to fields during the accumulation period for zone 1 and 2. For the same zones, decreasing differences during the melting period was also in agreement with expectations.

The expected differences between forest and fields was not observed at elevation Zone 3 (Figure 6.7). This is not only against our expectation, but also stands in contrast to results from Jost et al. (2007), who reported forests to have SWE reduced by 39% compared to nearby field sites. The only possible explanation we could imagine is redistribution by wind from fields to forests and depressions, reducing SWE on field plots, while the snow pack on adjacent plots in the forest remains almost unaffected. Therefore, we suggest to include the effects of wind redistribution in further investigations, for example using the approach of Dunn and Colohan (1999) which combines the degree-day model with a function accounting for wind redistribution.

Different effects of land use for different elevation zones are likely to cause the missing effect in the multiple linear regression model.

We expected some effect of plot aspect on SWE for melting periods, because of differing radiation inputs. However, the observed patterns appear to be random. We found two possible explanations: (a) The potential solar radiation input depends on the influence of fog, the plot aspect and slope, but this study only considered aspect, since measurements of slopes were not performed with sufficient accuracy. Therefore the potential solar radiation input is not quantified adequately. Subsequent studies should therefore pay more attention to obtaining detailed radiation data. (b) Anderton et al. (2004) report, that the spatial distribution of SWE at the start of the melting season is the primary control on patterns of snow disappearance, and that representation of spatial variability in melt rates is of minor importance. As a consequence, effects like wind redistribution during the accumulation period may superimpose effects like the plot aspect during the melting period.

6.4.4 Degree-day model

6.4.4.1 Chain convergence

While PSRF was a useful indication of chain convergence, we found two cases where chains appeared to have converged according to the PSRF, while the evaluation of the trace plots and target distributions indicated non convergence. The different interpretation between PSRF and visual inspection may be caused by the overdispersion of the prior distribution compared to the posterior distribution, that is required by the Gelman-Rubin diagnostic (Cowles and Carlin, 1996; Gelman et al., 2003; Plummer et al., 2009). We used random sampling for the starting values for each chain, sampling a low number of starting values, which makes it impossible to assure that the sampled values span the range of the given initial distribution. This in turn can cause the Gelman-Rubin diagnostic to be inappropriate, since we can not guarantee for the required overdispersion. A possible alternative would be to start with fixed values for each chain, covering the entire parameter space. Based on our experience, it is mandatory to examine the trace plots and posterior distributions for each parameter. In addition, future studies should make use of the possibility to include prior knowledge about feasible parameter ranges, for example by making negative and implausibly high values improbable.

High values for the degree-day factor for plots with non-converging chains result in the modelled SWE to be to zero after short times of melting conditions, (almost) independent of the amount of accumulated snow. Apparently, the posterior distributions estimated for plots at lower altitudes are strongly influenced by the snow surveys where no snow was observed.

6.4.4.2 Parameter dependence

We found strong elevation gradients for temperature and precipitation, which are the driver for increasing SWE with elevation. While interpolation included the elevation gradients, we did not explicitly evaluate the interpolated meteorological input data. Thus, statements about the quality of the input data for the snow model would be speculative.

Since the elevation-gradients were included during the interpolation and to obtain a closed water balance, we expected an accumulation factor $a0$ of about one for field plots with no dependence on elevation. As expected, no dependence of $a0$ on elevation was found. However, we observed values between 0.19 to 0.97 with the majority around 0.6, indicating that precipitation had to be reduced in order to successfully model snow melt dynamics.

We found two possible explanations for the low $a0$ values. First, an insufficient representation of the spatial variability (intensity and distribution) of precipitation during the interpolation, since rainfall is temporally and spatially highly heterogeneous. We suggest to use cross validation or a comparison with daily sums of radar rainfall as quality assessment of the interpolation methods. Second, reduction of snow height by wind transport from exposed locations to depressions and wind sheltered locations – which implies that our measurement locations were generally rather wind exposed. Note that effect of wind was not considered during the stratification of the plots. The unexpected effect of land use on snow height (more snow in forests) was already discussed above (section 6.4.3) and is also apparent in the $a0$ values.

Strong parameter interactions (Chapter 3) between $a0$ and the transition temperature between rain and snow could be an additional reason for the low $a0$ values. Thus, a different melting temperature might lead to values for $a0$ closer to 1. Martinec and Rango (1986) report threshold temperature between 3 $^\circ$ C and 0.75 $^\circ$ C, depending on the time of the year. Apparently a threshold temperature of 0 $^\circ$ C is lower. However, using higher threshold temperatures would increase the fraction of precipitation falling as snow, which would result in even lower values for the accumulation factor $a0$.

The accumulation factor $a0$ was not found to

depend on northing. This was expected because accumulation does not depend on radiation, for which northing is a proxy.

Values for degree-day factors $c0$ were expected between 2.8 and 5.3 mm/d $^{\circ}$ C (Hock, 2003). We found $c0$ between 0.24 and 2.23 for plots with converging chains. We are surprised by this result and do not have an explanation. In contrast to our initial expectations, no relationship between northing and the degree-day factor could be identified.

Degree-day factors for open fields are reported to be twice as high as for forests (Rango and Martinec, 1995). In contrast, we found higher $c0$ values for plots in the forest, compared to nearby fields, very similar to results for the accumulation factor. We found two possible causes for this difference to results from literature: 1) higher accumulation factors for forested sites may in turn require higher melt rates to melt the higher amounts of accumulated snow. 2) The shielding function of the forest canopy (Hardy et al., 2004) may be superimposed by other effects, such as wind redistribution.

6.4.4.3 Model evaluation

The results of the model evaluation reveal both, the advantages and the limitation of the degree-day model. On the one hand, for plots with a continuous snow pack, it is capable to explain most of the temporal variability over the snow season. While confidence bands of measurements and simulations overlapped, we often found a small overestimation of the first three and small underestimation for the last two campaigns, indicating that at the beginning of the snow season some process reduces the snow accumulation (remember also the surprisingly low accumulation factor $a0$), while accumulation suddenly increases before the fourth campaign. No reasons have been found that could explain this observation.

On the other hand, we were not able to find parameter sets that describe the pattern observed for lower elevations. Plots at low elevations showed recurrent accumulation and melting cycles with alternating snow free periods and periods with low SWE. The snow season also finishes earlier at lower elevations, while peak snow accumulation was not even reached by this time for plots at higher altitudes. This was not only observed for our measurements, but also for the data from the meteorological station situated in Dipplodiswalde, where the snow cover melts around the fourth snow survey. The snow free periods resulted in implausibly high degree-day factors compared to literature values (Hock, 2003; Martinec and Rango, 1986). In the results section, we highlighted that rain-on-snow may be an important process. Hence, we expect better estimations of snow dynamics in the lower catchment if the cold content of the snow pack and the heat input by rain would be quantified.

Since parameters could not be determined successfully for plots in the lower catchment, we did not expect parameters estimated for all plots to be meaningful. Accordingly the parameter set is not capable to explain the spatial variability at the catchment scale. The low value for the accumulation factor $a0$ leads to high RMSE values for the upper catchment as not sufficient precipitation is accumulated as snow. An extension of the model appears to be necessary to describe spatio temporal dynamics at this catchment. From our tests we we can conclude that the topographic factors elevation and northing are not a suitable extension of the model. The two factors are probably superimposed by other effects, mainly redistribution of snow by wind during the accumulation period and heat input by rain-on-snow during the melting period at the lower catchment.

An alternative approach to an extension of the model would be to estimate parameters for the data from a single elevation zone. Especially results for the upper catchment are expected to improve with this approach. However, results for the lower catchment are unlikely to improve – model extension seems more promising.

6.5 Summary and Conclusions

The variability of SWE was investigated at two scales in this study, the plot scale and the catchment scale. The local sampling design captured the variability at the plot scale in most cases. For less than 20 % of the plots the range estimated from fitted theoretical variograms extended beyond the range of the sampling design. At the catchment scale, a stratified nested sampling was used to capture SWE patterns with respect to elevation, exposition and land use. The results show increasing SWE with increasing elevation, however, the expected influence of land use and exposition has not been observed. All together, topographic controls and land use were capable to explain 74% of the observed variability at the catchment scale with multiple linear regression models.

A degree-day model was calibrated for each single plot. Confidence bands for model and measurements overlapped for plots at altitudes above 670 metres. For plots at lower elevations, the calibration showed non converging chains with the MCMC algorithm, which is a strong indicator for model deficiencies. Recurrent accumulation and melting periods as well as the heat input of rain are possible reasons for this result. With the simple degree-day model a continuous snow pack appears to be necessary in order to explain the temporal variability at the plot scale. As expected, results are not satisfactory for the simple model for parameters estimated for all plots.

We reported several facts indicating that redistribution of snow by wind could be a very important effect. 1) In Zinnwald, accumulated precipitation over the season only accounted for 70% of the reported SWE, 2) accumulation factors $a0$ of the degree-day model were significantly lower than 1, 3) for the upper catchment, we did not observe lower snow heights in forests compared to fields as reported in other studies and 4) we hypothesized that the effect of exposition, which we expected but did not observe, may be superimposed by wind transport.

The possibility to find a parsimonious model that explains the spatio-temporal variability for the entire catchment is more likely after inclusion of additional processes such as heat input by rain-on-snow and redistribution of snow by wind.

Suggestions for further investigations:

- Quantify the impact of wind redistribution that superimposes effects of land use and exposition.
- Apply the model to single elevation zones in order to account for the different characteristics of the snow season controlled by temperature and precipitation.
- Quantify the heat input of rain-on-snow.

Chapter 7

Summary and conclusions

7.1 Summary of achievements

Using the Weisseritz case study to demonstrate the different steps of the learning cycle lead to a number of scientific achievements. The major achievements are briefly summarized with respect to guiding questions as summarized in section 1.6:

How to assess (poor) model performance? Defining good and poor model performance in an objective way is challenging, because every performance measure is only sensitive for a certain number of differences between model and observation (Chapter 2). The same problem is also reported in the context of calibration, where multi-objective approaches have emerged. In order to find a generic approach, instead of selecting a small number of performance measures targeted at a single purpose, a large number (near to comprehensive) of performance measures has been assembled and implemented in R (Reusser, 2009). A complete model diagnostic procedure includes assessment of model performance and evaluation of the representation of processes by answering the three questions 1) when a model is performing acceptably/poor, 2) of what kind deviation are and 3) whether the relevant process conceptualisations are active during the right period. Temporally resolved analysis of model performance was developed in this thesis to answer the first two questions. This is the first part of TIGER (Chapter 2) and an innovation for hydrology, since temporally resolved model performance has only been assessed for single peaks before.

How is it possible to identify temporal patterns and context dependence in model performance? Since catchment behavior is strongly context dependent (rainfall driven, energy driven, snow influenced) it is informative to relate temporal patterns of model performance to context dependence of hydrology. Temporal patterns in model performance are visually easy to detect for short periods, including only few events. However, as longer series are analysed, visual inspection starts to fail because either details are not visible anymore if looking at an overview or there is a danger of loosing track and being overwhelmed by all the details if looking at zoomed views. As second part of TIGER, I developed a possibility to assess long simulation periods based on a meaningful data reduction method. The data reduction allows to cluster periods of similar model performance. Two cluster interpretation tools were developed and applied to better understand the significance of the clusters: cluster-wise boxplots of model performance and interpretation with synthetic peak errors.

Can we identify relevant model components (for computationally expensive models)? Relevant model components are detectable using temporal dynamics of parameter sensitivity (TEDPAS). The procedure allows to make a qualitative model validation by testing if processes are dominant during periods as we expect it. While TEDPAS has been used for shorter periods by Sieber and Uhlenbrook (2005), they reported computational problems to prevent analysis of longer time periods (e.g. complete hydrological years). My thesis is the first report of application of TEDPAS to long time series. In order to achieve this goal, the Fourier amplitude sensitivity test (FAST) is implemented in R (Reusser, 2008). This new implementation of FAST is compared to existing implementations of FAST and Sobol's method in Simlab, and is found to be computationally more efficient.

What are the limitations of WaSiM-ETH as representation of the Weißeritz catchment? Limitations of WaSiM-ETH are identified using a combination of TIGER and TEDPAS. With this combination, repeatedly occurring patterns of poor model performance are related to dominant model parameters, indicating deficiencies in process representations. For WaSiM-ETH I found that snow

melt periods are generally overestimated which may be caused by missing processes of snow removal or missing representation of spatial variability of driving forces. Recession periods during dry periods are not matched well, indicating that simple linear reservoir recession or the topmodel approach may not fit the landscape very well. In addition, some peaks are missed in winter, which has been identified to be caused by erroneous winter data (ice jams). Not including this data is important for a correct calibration. Additional measurements have focused on snow processes and we collected information about temporal and spatial variability.

How much information can be obtain from measurements with inexpensive temperature sensors? Temperature was measured in and above snow cover with inexpensive loggers. As part of this thesis, a new algorithm was developed to determine snow height from these measurements. Estimated snow heights agree well with reference measurements at the same location. The vertical resolution of the estimated snow height can be increased during melting periods, since a simple temperature index model allows to interpolate between single sensors. The data is also well suited to estimate cold content. Going from observations to predictions of cold content, a simple temperature diffusion model reduces RMSE by ≈30% compared to the persistence assumption, despite strong oversimplifying assumptions.

What are temporal and spatial structures of the snow in the catchment? In a Diploma Thesis (Chapter 6), measurements were used to separate snow variability at the local scale (few meters) to variability at the catchment scale (km). At the local scale, variograms were estimated and the range determined from fitted variograms was within the range of the sampling design for 80% of the plots. Consistent with Jost et al. (2007) no variables explaining the structure of the local variability could be detected. On the catchment scale, the snow height is mainly affected by the plot altitude. The expected influence of aspect and land use was not observed.

Temporal structures are different for the upper and the lower catchment. For the upper catchment, we can identify an accumulation and a melt phase with a persisting snow cover during the entire winter. In the lower catchment, the snow cover was determined by recurrent snow-fall and melting cycles, resulting in multiple short term snow covers over time.

What processes are required to describe the new measured data and what are the resulting updates to the model? A temperature-degree day model was applied to test whether the spatio-temporal variability of SWE can be represented by this simple model. The degree-day model is capable to explain the temporal variability for plots with a continuous snow pack over the entire snow season, if parameters are estimated separately for single plots. However, processes described in the simple model are not sufficient to represent multiple accumulation-melt-cycles, as observed for the lower catchment. Thus, the combined spatio-temporal variability at the watershed scale is not captured by the model. Since WaSiM-ETH is based on this approach, the result confirms the short-comings of this model as a conceptualisation of snow processes in the Weisseritz catchment.

The analysis indicates, that snow on rain and redistribution by wind or a full energy balance model will have to be tested as a next steps. Time did not permit to find a conclusive answer.

7.2 Discussion and future research questions

Discussion and future research questions are organized by methods.

7.2.1 Temporal dynamics of model performance

TIGER is a step towards the resolution of two issues with model diagnostics. First, a single criterion is not sufficient for diagnosis of current environmental models, as each objective function is sensitive for certain aspects of deviation only. Sensitivity of objective functions for certain aspects is demonstrated with the synthetic peak errors (Chapter 2). To avoid negligence of some aspects of deviation, multiple diagnostic signatures should be derived from theory which check that important system behaviours are reproduced by the model (Gupta et al., 2008). Combination of multiple measures provides a better characterization of the performance compared to any single measure, which agrees with the basic idea of multi-objective calibration. Second, catchments work very different in different contexts (rainfall driven, energy driven, snow dominated, ...). Thus, looking at temporal dynamics of model performance provides additional, context dependent information.

TIGER has been used to identify performance clusters for two different models, Catflow, a physically based model and WaSiM, a more conceptual model. From these applications I am confident that the method is generic and can be applied to a wide range of rainfall-runoff models.

Some of the performance measures are highly correlated. After applying the method to several models, the same set of performance measures appears to be sufficient to describe the difference between model and observation. Future research should investigate this more systematically, possibly providing a minimal list of comprehensive performance measures.

While the results of TIGER are not very sensitive for the selection of the time window size, an automatic peak detection and identification of appropriate window size would remove one subjective selection of the method. The rank transformation and use of self-organizing maps make the method robust for extreme values in performance measures (which occur for the short time windows).

The criterion applied for the selection of cluster size may result in very few clusters (2 or 3), which will not reveal a great amount of temporal dynamics. As with every data aggregation method, some information is lost during the data reduction. However, selecting a higher number of clusters that does not strictly minimize the selection criterion may be a better choice. Subjectivity at the end of the analysis can not be avoided, since the subsequent interpretation of the clusters necessarily remains subjective. For the interpretation of the clusters, the two tools proved to be valuable: cluster wise box plots and synthetic peak errors.

Future research may include 1) multi model comparison with TIGER, 2) using virtual experiments to further test how well TIGER can detect deficits in model structure, 3) box plots might be extended to include state variables such as discharge, antecedent precipitation index and similar. 4) Continue development of synthetic peak errors. Combined errors could be used as a bench mark for performance measures, similar to the error response groups discussed in chapter 2. 5) more effort towards feature based objectives, which are "signature indices that measure theoretically relevant system process behaviors" Gupta et al. (2008). 6) Methods from Brun et al. (2001) or Chu and Hahn (2009) may be used to find representative subsets from all performance measures, replacing the current, correlation-based selection. 7) A smarter assessment of performance clusters should be able to indicate what the missing processes are. This problem is in some way similar to problems encountered during fault detections of large power grid systems*.

*Hoshin Gupta, Mai 2010, personal communication

7.2.2 Temporal dynamics of parameter sensitivity

So far, not many applications of TEDPAS have been reported (Sieber and Uhlenbrook, 2005; Cloke et al., 2008). TEDPAS can be used as powerful tool for 1) the identification of relevant model components for each period, 2) for the selection of the most informative time periods for calibration, and 3) detection of parameter interactions from TEDPAS correlations.

Qualitative validity for WaSiM-ETH can be established with TEDPAS (e.g. no snow during summer, right order of different recession components). Future research should demonstrate the benefit of TEDPAS for calibration, by using periods of high parameter sensitivity only for the calibration of a given parameter. Also, the detection of parameter interactions should be further developed with virtual experiments and by extending the approach called "practical identifiability analysis“ (Brun et al., 2001) with global instead of local sensitivity analysis.

The underlying sensitivity analysis could be extended. For FAST, checks against benchmark tests (Saltelli and Bolado, 1998, e.g.) for SA should be run. Methods for total order SA could complement FAST to provide “a fairly complete and parsimonious description of the model in terms of its global SA properties” Saltelli et al. (2006).

7.2.3 Method combination

The combination of TIGER and TEDPAS was useful for the identification of weaknesses of WaSiM-ETH for representing the Weisseritz catchment and highlighted a number of deficiencies (Chapter 4). The combination of TIGER and TEDPAS is helpful for the reduction of model structural uncertainty. It is possible to use an “extended multi objective” approach by testing the model against multiple objective functions calculated for several sets of independent target data. The underlying idea is to increase the "information content" of the calibration data space (Gupta et al., 2008). A second approach for the reduction of model structural uncertainty is to represents dominant processes and their controls such that characteristic behaviour can be reproduced in a more realistic manner, for instance resolving lateral flows and surface and subsurface flow paths, or reproducing subsurface storage volumes. This is often referred to as "process complexity" of the model and means to reduce the manifold of acceptable model structures. The use of more complex models implies that computational effort and simulation times increase considerably. The combination of TIGER and TEDPAS is fast enough to be applied to models with increasing complexity because: a) it is not necessary to calibrate the model in advance, b) a highly efficient method is used to sample the parameter space, and c) all model runs are evaluated (to determine parameter sensitivity) while other Monte Carlo based methods often discard the 90% worst runs as a first step.

Future development of the method could include virtual experiments, having one researcher introduce a model deficiency and a second researcher apply the combination of TIGER and TEDPAS to identify the deficiencies. The approach could also be further developed as a means to catchment classification by searching sets of compatible and incompatible catchment - model pairs. We used a very simple method for matching of error clusters and dominant parameter sensitivities. Putting some effort into an improved combination might enhance interpretability. One possibility would be to include parameter sensitivities while estimating the self organizing maps. When using model diagnostics to guide field experiments it is not possible to provide a generally applicable recipe, because the interpretation of the results remains subjective. Work related to finding a minimal set of measurements for ungauged basins within the PUB initiative may be relevant for progress on this topic (Seibert and Beven, 2009; Winsemius et al., 2009; Blume et al., 2008b,a).

7.2.4 Snow temperatures

Hobo temperature sensors are sufficiently accurate and their resolution with respect to temperature and time is high enough to clearly detect the reduction of diurnal variation, which is necessary for the estimation of snow heights. Estimated heights are in good agreement with reference measurements. The observed error corresponds to the theoretical expectation from the sensor spacing. A slight underestimation is possible, because a thin snow cover on the sensor may not be detected.

As random height estimates occur during snow free periods, this can be used as a criterion to identify snow free periods. Estimates for degree day factors from the interpolation during melting periods are not comparable to literature values, since compaction of snow and melting can not be distinguished from the temperature data alone. Further development is necessary to reduce the influence of the sensors and the metal rod on the snow cover.

The estimation of the cold content requires an assumption about the snow density, which introduces additional uncertainty. Jonas et al. (2009) use height, location and time for a regression based density estimate for locations in the alps, which should be tested for its applicability to the Weisseritz catchment. Alternatively history of temperature might be used instead of time of year.

If going from estimation of the cold content towards prediction, strong simplification assumptions are necessary to create a simple diffusion model for the prediction of the cold content. The estimated thermal conductivity, which is the only model parameter, is often outside the theoretical range. Despite the error reduction of 30% with respect to the persistence assumption (assuming perfect temperature predictions), the model is not sufficient to make reliable predictions.

Future research might include the use of sensor networks to obtain real time information. Also, the algorithm should be adapted to include information from previous time steps about snow height, e.g. a Kalman filtering approach or similar could be used.

7.2.5 Spatial variability of the snow cover

The sampling scheme used for the assessment of local variability is sufficient. Covering 60 m for the assessment at local scale was found to be sufficient, which agrees with (Deems et al., 2006, 2008) who report a break in fractal scale at a length of 15-40 m for other catchments.

Improvements are certainly possible with new measurement techniques. Terrestrial laser scanning (TLS) could provide much richer data sets (Hood and Hayashi, 2010; Schaffhauser et al., 2008; Prokop, 2008). Possibilities for the evaluation could be strongly increased using TLS as a non-intrusive method, which would allow to measure at exactly the same location for each campaign. Also considering irregularities of the soil surface appears to be very important to understand variabilities for low snow heights. This could be assessed with TSL measurements during snow free conditions.

Overall, the problem with local variability is similar in structure as issues also encountered in the OPAQUE project for the measurement of soil moisture as reported by Zehe et al. (2010). Large short scale variabilities are present, however, dynamics are strongly correlated.

At the catchment scale, the elevation was the only influencing factor that was able to contribute to the explanation of variability of SWE. No effect was visible for land use and exposition. For future campaigns, more effort is necessary to better quantify topographic characteristics (the available 20 m DEM does not have sufficient resolution). Estimates of the exposition to wind should also be included in a future sampling scheme.

A simple degree day model was used to reproduce the data at the catchment scale. However, Monte Carlo Markov Chains did not converge for sites situated in the lower catchment. These sites are characterized by multiple snow accumulation and melting periods. For plots in the up-

per catchment, parameter estimation for individual plots was successful and the model was able to reproduce the data in a satisfactory way. For conservation of mass, we would expect accumulation factors of 1, however, much lower factors were generally observed. One reason may be insufficient interpolation of meteorological data, that leads to uncertain input data. Also effects from wind, interception, sublimation, and other local characteristics may cause accumulation factors below 1. Also the degree day factors are often below the range reported by Hock (2003) (2.5-11.6 mm/d/°C). No reasons has been found for this difference.

The effect of forest is not consistent for all places and differs between the lower and the upper catchment. In the lower catchment less snow is observed in forests as reported in other studies (Jost et al., 2007; Pomeroy et al., 1998). For the upper catchment, snow accumulation and melt factors are higher in forests compared to field plots, which is opposite to results reported for example by (Jost et al., 2007). I do not have an explanation for this differences, but factors that make the influence of forest complex include interception capacity varying with meteorological conditions and tree stand density (Winkler et al., 2005; Winkler and Moore, 2006). Wind transport of snow into the forest may explain higher accumulation in the forests. It is uncertain whether such small scale effects need to be fully resolved for flood prediction or may be accounted for by some empirical integrative factor. But even if resolving these variabilities is not necessary, identification of representative sites will remain an important issue (Molotch and Bales, 2005), making better understanding of snow processes necessary.

7.3 Conclusion

Models used for flood prediction need to represent bio-physical processes in a realistic way in order to reduce mistakes caused by the extrapolation to unobserved system states. This requires a constant learning about the functioning of the catchment under investigation and its representation in the model. This is achieved with a learning cycle which starts with the model as currently best representation of the relevant processes in a catchment. Model diagnostic is then used to identify deficiencies in the catchment representation. From the deficiencies, a set of field measurements is derived. These measurements are the base for a revision of the process concepts. The overall rational behind this thesis is to provide new tools to better facilitate such an iterative learning cycle for the case of flood predictions.

A key step in this learning cycle is the model diagnostics, for which we identified 3 questions to be answered: 1) during which periods the model is or is not reproducing observed quantities and dynamics; 2) What is the nature of the error in times of poor model performance, and 3) which components of the model are causing this error. Answering these three question will highlight the relevant components of the model to be revisited, enabeling us to improve the model in a very targeted way.

The first two questions are related to poor model performance. Since hydrological functioning is strongly context dependent (rain driven, energy driven, snow dominated, ...), average performance of a model is only a first order assessment. To improve rainfall-runoff models in a much more targeted way requires time dependent performance measures. This is achieved with moving time windows in order to resolve the temporal dynamics of the model performance.

In order to identify and characterize poor model performance, a large number of objective functions is used in the newly developed TIGER method (TIme series of Grouped ERrors). The important aspects are captured by these measures and a better characterization of the performance is possible compared to any single measure, similar to multi objective calibration. The approach is consistent with the diagnostic evaluation approach (Gupta et al., 2008). Their idea of multiple diagnostic signatures is very similar to using the large number of

performance measures in TIGER.

However, data reduction techniques are necessary to handle the resulting large amount of data. From the data reduction, patterns of error repetition are of special interest, as these highlight recurrent differences between model and observation. The clustering algorithm used in TIGER results in meaningful clusters. To reveal the meaning of each cluster, interpretation tools for performance clusters have been developed. The two tools used in this study are synthetic peak errors and box plots of performance measures. TIGER has been applied successfully to multiple models and catchments. This demonstrates the wide range of research areas and modelling approaches to which the approach can be applied.

The third question of the model assessment, is to identify which components of the model are causing a difference between observation and model. This analysis is based on sensitivity analysis (SA). More precisely, the Fourier amplitude sensitivity test (FAST) was used as SA method, because the FAST method is a very efficient method for the calculation of first order partial variance global sensitivities - which is state of the art for the identification of dominating model components. A reimplementation of FAST has been compared to other sensitivity methods with a lumped, computationally inexpensive model. Great differences in computational expenses exist. My reimplementation of FAST results in a improved efficiency of a factor of about 50 compared to SIMLAB 3.4.6 for 10000 SA evaluations.

So far, temporal dynamics of parameter sensitivity (TEDPAS) has been applied only in few cases in hydrology despite its great potential (Cloke et al., 2008; Sieber and Bremicker, 2006). In this thesis, relevant processes for the grid-based, computationally expensive model WaSiM-ETH for the Weisseritz have been identified with TEDPAS, demonstrating that such an analysis is possible for long time series and complex models. This provides a part of the information required for the model diagnostics and is a way for qualitative model evaluation. Combination of TIGER and TEDPAS allows to answer all three questions relevant for model diagnostic. The combined diagnostics is applied without prior calibration, constituting a large advantage for computationally expensive models.

As a vision, such a diagnostic could be performed for multiple hydrological models and for multiple catchments, providing a very powerful approach for model comparison. We expect the same model to show different structural deficits in different landscapes, and different model concepts to show different structural deficits in the same landscape. Consistent application of the proposed methodology could, in the long term, enable the development of a basis for discriminating model/process concepts and landscapes into "compatible and incompatible sets" (in which the model/process can be expected to work with low structural/high structural deficits). Ultimately, it could help to reduce the overwhelming number of hydrological models to a minimum amount necessary for dealing with the richness of our landscapes.

From the in depth model analysis, gaps in the observation data set can be identified. Research for cost-effective measurements allows us to obtain higher resolved data since more sensors can be installed for the same cost. We developed and applied methods resolving temporal and spatial variability of the snow cover. To observe temporal variability, inexpensive temperature sensors provide a cost efficient way for snow height monitoring without additional information requirements (such as meteorological data). A new algorithm has been presented which allows automatic extraction of snow height estimates from temperature profile measurements. The determined heights agree well with reference measurements. Snow height and cold content are simultaneously calculated and provide important information for flood warning, model evaluation and model state updating.

The spatial variability of snow heights and SWE

is characterized with a data set collected during the winters 2008/2009 and 2009/2010 for the Weisseritz catchment. This data serve to test and improve the models describing the spatial variability as well as for reference for the remote sensing snow cover characterisation, which is part of a related project within OPAQUE. As expected, height a.s.l. is the most influential factor on the catchment scale. Contrary to the expectations, no significant effects of exposition and land use could be identified from the data.

As a case study, WaSiM-ETH was used as a hypothesis for the functioning of the Weisseritz catchment. The combined diagnostics revealed important deficiencies requiring improvements to the model: 1) overestimation during winter snow melt periods can possibly be reduced with a land use dependent snow melt index based on land use dependent measurements of snow cover patterns. 2) improvement of the model spin-up are necessary. 3) Topmodel may not be the right approach to conceptualize the functioning of the Weisseritz catchment. 4) the validity of the simple linear reservoir recession should be checked with a recession analysis. Also, errors in data caused by ice jams were identified, thus making it possible to reduce the influence of data errors during model calibration by excluding this data.

Using the newly collected data, the suitability of the snow module of WaSiM-ETH and the effect from the land use dependent snow melt index as proposed above is tested. To this end, a simple degree day factor model as it is used in WaSiM-ETH is tested for its capability to describe the spatio temporal patterns of the snow cover. It is sufficient to represent the temporal dynamics of the snow cover for locations where a continuous snow cover is present. No dependence of the snow melt index on land use could be detected, indicating that this is not the relevant model improvement to deal with the overestimation during winter snow melt periods. However, snow accumulation has to be corrected with a factor <1, indicating that not all snow remained at the measurement locations. Both, sublimation and relocation by wind to depressions and more wind protected areas may be important influences as reported for example by MacDonald et al. (2009, 2010) and may alter the water balance during melt events.

The model does not capture the dynamics at locations with repeated accumulation and melting periods, as observed in the lower catchment. Snow melt factors estimated for these sites are far above the range reported in literature. Thus, processes relevant at such sites such as energy input from rain on snow may need to be included in addition to the influence by wind.

Before closing the learning cycle with an updated model, more investigations are necessary. For example, tests are necessary to see if rain on snow and redistribution by wind will improve the performance of the degree day factor approach. Also, spatial patterns of the snow cover should be further evaluated. A simple way for such an evaluation is to correlate the ranks of the distributed observations with the ranks of the simulated snow cover at these locations. A high correlation would indicate that the major processes causing spatial variability are represented in the model, while a low correlation would indicate that there are still missing processes.

My analysis shows how a constant learning process for flood forecasting helps to achieve good knowledge about the relationship between the catchment and its representation in the model. While in practical applications, constant learning is occurring in any case, the tools developed in this thesis may help to perform such a learning in a more structured and reproducible way.

List of Figures

List of Tables

Bibliography

Abramowitz, G., Leuning, R., Clark, M. P., Pitman, A., Nov. 2008. Evaluating the Performance of Land Surface Models. Journal Of Climate 21 (21), 5468–5481.

Aggarwal, R. K., Negi, P. S., Satyawali, P. K., 2009. New Density-based Thermal Conductivity Equation for Snow. Defence Science Journal 59 (2), 126–130.

Akin, H., Siemes, H., 1988. Praktische Geostatistik: eine Einführung für den Bergbau und die Geowissenschaften. Springer.

Anderton, S. P., White, S. M., Alvera, B., Nov. 2002. Micro-scale spatial variability and the timing of snow melt runoff in a high mountain catchment. Journal of Hydrology 268 (1-4), 158–176.

Anderton, S. P., White, S. M., Alvera, B., 2004. Evaluation of spatial variability in snow water equivalent for a high mountain catchment. Hydrological Processes 18 (3), 435–453.

Andreadis, K. M., Lettenmaier, D. P., 2006. Assimilating remotely sensed snow observations into a macroscale hydrology model. Advances In Water Resources 29 (6), 872–886.

Andreadis, K. M., Storck, P., Lettenmaier, D. P., 2009. Modeling snow accumulation and ablation processes in forested environments. Water Resources Research 45, —-.
URL http://dx.doi.org/10.1029/2008WR007042

Bárdossy, A., 2007. Calibration of hydrological model parameters for ungauged catchments. Hydrology And Earth System Sciences 11 (2), 703–710.
URL http://www.hydrol-earth-syst-sci.net/11/703/2007/

Barrenetxea, G., Ingelrest, F., Schaefer, G. L., Vetterli, M., 2008. The hitchhiker's guide to successful wireless sensor network deployments. In: SenSys '08: Proceedings of the 6th ACM conference on Embedded network sensor systems. ACM, New York, NY, USA, pp. 43–56.

Bartelt, P., Lehning, M., Nov. 2002. A physical SNOWPACK model for the Swiss avalanche warning Part I: numerical model. Cold Regions Science And Technology 35 (3), 123–145.

Bastola, S., Ishidaira, H., Takeuchi, K., Aug. 2008. Regionalisation of hydrological model parameters under parameter uncertainty: A case study involving TOPMODEL and basins across the globe. Journal of Hydrology 357 (3-4), 188–206.
URL http://linkinghub.elsevier.com/retrieve/pii/S0022169408002187

Benke, K. K., Lowell, K. E., Hamilton, A. J., 2008. Parameter uncertainty, sensitivity analysis and prediction error in a water-balance hydrological model. Mathematical And Computer Modelling 47 (11-12), 1134–1149.

Beven, K. J., 1997. TOPMODEL: a critique. Hydrological Processes 11 (9), 1069–1085.

URL http://earth.boisestate.edu/home/jmcnamar/hydanalysis/Notes/topmodel_beven.pdf

Beven, K. J., 2001. Rainfall-runoff modelling : the primer. Wiley.

Beven, K. J., Oct. 2002. Towards a coherent philosophy for modelling the environment. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 458 (2026), 2465–2484.
URL http://rspa.royalsocietypublishing.org/cgi/doi/10.1098/rspa.2002.0986

Beven, K. J., Binley, A., 1992. The future of distributed models: Model calibration and uncertainty prediction. Hydrological Processes 6 (3), 279–298.
URL http://dx.doi.org/10.1002/hyp.3360060305

Beven, K. J., Freer, J. E., Aug. 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. Journal of Hydrology 249 (1-4), 11–29.

Beven, K. J., Goetzinger, J., Montanari, A., More, 2010. How can we separate and identify input observation error versus model structural error? Advances In Water Resources, submitted.

Beven, K. J., Kirby, M., 1979. A physically based variable contributing area model of basin hydrology. Hydrol. Sci. Bull 24 (1), 43–69.

Beven, K. J., Lamb, R., Quinn, P., Romanowicz, R., Freer, J. E., 1995. Topmodel. Water Resources Publications, Colorado, pp. 627–668.

Bezdek, J. C., 1981. Pettern Recognition with Fuzzy Objective Function Algorithms. Plenum, New York.

Blöschl, G., 1999. Scaling issues in snow hydrology. Hydrological Processes 13 (14-15), 2149–2175.

Blume, T., 2008. Hydrological processes in volcanic ash soils - Measuring, modelling and understanding runoff generation in an undisturbed catchment. Ph.D. thesis, University of Potsdam.

Blume, T., Zehe, E., Bronstert, A., 2007. Rainfall runoff response, event-based runoff coefficients and hydrograph separation. Hydrological Sciences Journal 52 (5), 843–862.

Blume, T., Zehe, E., Bronstert, A., 2008a. Investigation of runoff generation in a pristine, poorly gauged catchment in the Chilean Andes II: Qualitative and quantitative use of tracers at three spatial scales. Hydrological Processes 22 (18), 3676–3688.
URL http://www3.interscience.wiley.com/journal/117904047/abstract

Blume, T., Zehe, E., Bronstert, A., Jul. 2009. Use of soil moisture dynamics and patterns at different spatio-temporal scales for the investigation of subsurface flow processes. Hydrol. Earth Syst. Sci. 13 (7), 1215–1233.
URL http://www.hydrol-earth-syst-sci.net/13/1215/2009/http://www.hydrol-earth-syst-sci.net/13/1215/2009/hess-13-1215-2009.pdfDO-10.5194/hess-13-1215-2009

Blume, T., Zehe, E., Reusser, D. E., Iroume, A., Bronstert, A., Aug. 2008b. Investigation of runoff generation in a pristine, poorly gauged catchment in the Chilean Andes I: A multi-method experimental study. Hydrological Processes 22 (18), 3661–3675.

Bourgouin, P., Oct. 2000. A Method to Determine Precipitation Types. Weather and Forecasting 15 (5), 583–592.

URL http://journals.ametsoc.org/doi/abs/10.1175/1520-0434(2000)015<0583:AMTDPT>2.0.CO;2

Boyle, D. P., Gupta, H. V., Sorooshian, S., 2000. Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods. Water Resources Research 36 (12), 3663–3674.

Brandt, R. E., Warren, S. G., 1997. Temperature measurements and heat transfer in near-surface snow at the South Pole. Journal Of Glaciology 43 (144), 339–351.

Braun, L., 1985. Simulation of snowmelt - runoff in lowland and lower alpine regions of switzerland. Zürcher Geographische Schriften 21.

Bronstert, A., Creutzfeldt, B., Graeff, T., Hajnsek, I., Heistermann, M., Itzerott, S., Jagdhuber, T., Kneis, D., Lück, E., Reusser, D. E., Zehe, E., 2010. Multi-scale soil moisture observations and their potential use for flood forecasting in mountainous headwater catchments. Natural Hazards, submitted.

Brun, R., Reichert, P., Künsch, H., 2001. Practical identifiability analysis of large environmental simulation models. WATER RESOURCES RESEARCH 37 (4), 1015–1030.
URL http://www.agu.org/journals/wr/v037/i004/2000WR900350/

Buytaert, W., Beven, K. J., Nov. 2009. Regionalization as a learning process. Water Resources Research 45, —-.
URL http://dx.doi.org/10.1029/2008WR007359

Buytaert, W., Célleri, R., De Bièvre, B., Cisneros, F., Wyseure, G., Deckers, J., Hofstede, R., Nov. 2006a. Human impact on the hydrology of the Andean páramos. Earth-Science Reviews 79 (1-2), 53–72.
URL http://linkinghub.elsevier.com/retrieve/pii/S0012825206000808

Buytaert, W., DECKERS, J., WYSEURE, G., Feb. 2006b. Description and classification of nonallophanic Andosols in south Ecuadorian alpine grasslands (páramo). Geomorphology 73 (3-4), 207–221.
URL http://linkinghub.elsevier.com/retrieve/pii/S0169555X05002606

Buytaert, W., Reusser, D. E., Krause, S., Renaud, J.-P., 2008. Why can't we do better than Topmodel? Hydrological Processes 22 (20), 4175–4179.

Buytaert, W., Wyseure, G., De Bièvre, B., Deckers, J., Dec. 2005. The effect of land-use changes on the hydrological behaviour of Histic Andosols in south Ecuador. Hydrological Processes 19 (20), 3985–3997.
URL http://doi.wiley.com/10.1002/hyp.5867

Campbell, 2006. IRTS-P Precision Infrared Temperature Sensor.

Chib, S., Greenberg, E., 1995. Understanding the metropolis-hastings algorithm. American Statistician 49 (4), 327–335.
URL http://www.jstor.org/stable/2684568

Choi, H. T., Beven, K. J., 2007. Multi-period and multi-criteria model conditioning to reduce prediction uncertainty in an application of TOPMODEL within the GLUE framework. Journal Of Hydrology 332 (3-4), 316–336.

Christiaens, K., Sep. 2002. Use of sensitivity and uncertainty measures in distributed hydrological modeling with an application to the MIKE SHE model. Water Resources Research 38 (9), —-.
URL http://www.agu.org/pubs/crossref/2002/2001WR000478.shtml

Chu, Y., Hahn, J., Jul. 2009. Parameter Set Selection via Clustering of Parameters into Pairwise Indistinguishable Groups of Parameters. Industrial & Engineering Chemistry Research 48 (13), 6000–6009.
URL http://pubs.acs.org/doi/abs/10.1021/ie800432s

Clark, M. P., Slater, A. G., Barrett, A. P., Hay, L. E., McCabe, G. J., Rajagopalan, B., Leavesley, G. H., Aug. 2006. Assimilation of snow covered area information into hydrologic and land-surface models. Advances In Water Resources 29 (8), 1209–1221.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., Hay, L. E., 2008. Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. Water Resources Research 44, W00B02.

Cloke, H., Pappenberger, F., 2008. Evaluating forecasts of extreme events for hydrological applications: an approach for screening unfamiliar performance measures. Meteorological Application 15, 181–197.

Cloke, H., Pappenberger, F., 2009. Ensemble flood forecasting: a review. Journal of Hydrology In Press,, —-.
URL http://www.sciencedirect.com/science/article/B6V6C-4WH2M7P-8/2/0a51bb69fa203edd0e042aa64c2d6d64

Cloke, H., Pappenberger, F., Renaud, J.-P., 2008. Multi-Method Global Sensitivity Analysis (MMGSA) for Modelling Floodplain Hydrological Processes. Hydrological Processes.

Corbari, C., Ravazzani, G., Martinelli, J., Mancini, M., Nov. 2009. Elevation based correction of snow coverage retrieved from satellite images to improve model calibration. Hydrology and Earth System Sciences 13 (5), 639–649.
URL http://www.hydrol-earth-syst-sci.net/13/639/2009/

Cottrell, M., de Bodt, E., 1996. A Kohonen map representation to avoid misleading interpretations. In: 4th European Symposium on Artificial Neural Networks.
URL http://www.dice.ucl.ac.be/esann/proceedings/papers.php?ann=1996

Cowles, M., Carlin, B., 1996. Markov Chain Monte Carlo convergence diagnostics: a comparative review. Journal of the American Statistical Association 91 (434).

Cukier, R. I., Fortuin, C. M., Shuler, K. E., Petschek, A. G., Schaibly, J. H., 1973. Study Of Sensitivity Of Coupled Reaction Systems To Uncertainties In Rate Coefficients .1. Theory. Journal Of Chemical Physics 59 (8), 3873–3878.

Cukier, R. I., Levine, H. B., Shuler, K. E., 1978. Non-Linear Sensitivity Analysis Of Multi-Parameter Model Systems. Journal Of Computational Physics 26 (1), 1–42.
URL http://www.sciencedirect.com/science/article/B6WHY-4DD1NST-H8/2/f280917ea9b610ef28755bbe55c507be

Cukier, R. I., Schaibly, J. H., Shuler, K. E., 1975. Study Of Sensitivity Of Coupled Reaction Systems To Uncertainties In Rate Coefficients .3. Analysis Of Approximations. Journal Of Chemical Physics 63 (3), 1140–1149.

Cullmann, J., Mishra, V., Peters, R., 2006. Flow analysis with WaSiM-ETH – model parameter sensitivity at different scales. Advances in Geosciences 9, 73–77.

Davies, J., Studer, R., Sure, Y., Warren, P. W., 2005. Next generation knowledge management.

BT Technology Journal 23 (3), 175–190.
URL http://dx.doi.org/10.1007/s10550-005-0040-3

Dawson, C. W., Abrahart, R. J., See, L. M., 2007. HydroTest: A web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. Environmental Modelling & Software 22 (7), 1034–1052.

DeBeer, C. M., Pomeroy, J. W., 2010. Simulation of the snowmelt runoff contributing area in a small alpine basin. Hydrology and Earth System Sciences Discussions 7 (1), 971–1003.
URL http://www.hydrol-earth-syst-sci-discuss.net/7/971/2010/

Deems, J. S., Fassnacht, S. R., Elder, K. J., Apr. 2006. Fractal distribution of snow depth from lidar data. Journal Of Hydrometeorology 7 (2), 285–297.

Deems, J. S., Fassnacht, S. R., Elder, K. J., 2008. Interannual Consistency in Fractal Snow Depth Patterns at Two Colorado Mountain Sites. Journal of Hydrometeorology 9 (5), 977–988.
URL http://dx.doi.org/10.1175/2008JHM901.1

Deflandre, A., Williams, R. J., Elorza, F. J., Mira, J., Boorman, D. B., 2006. Analysis of the QUESTOR water quality model using a Fourier amplitude sensitivity test (FAST) for two UK rivers. Science Of The Total Environment 360 (1-3), 290–304.

Demaria, E. M., Nijssen, B., Wagener, T., 2007. Monte Carlo sensitivity analysis of land surface parameters using the Variable Infiltration Capacity model. Journal Of Geophysical Research-Atmospheres 112 (D11), D11113.

Dietrich, R., 2009. Snow Report of the "Skisportzentrum, Wanderheim und Sporthotel, Freizeitanlagen GmbH", a hotel in Hermsdorf-Rehefeld.
URL http://www.saechsische-schweiz-touristik.de/swf/www-cms2.pl?template=winter.html

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., and Andreas Weingessel, 2008. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien.

Dingman, S. L., 2002. Physical hydrology. Waveland Press, Long Grove, Ill.

Dunn, S. M., Colohan, R. J. E., Sep. 1999. Developing the snow component of a distributed hydrological model: a step-wise approach based on multi-objective analysis. Journal of Hydrology 223 (1-2), 1–16.
URL http://www.sciencedirect.com/science/article/B6V6C-3XG1T11-1/2/9c3f1576064e0287d53b9651ef031788

Durand, M., Molotch, N. P., Margulis, S. A., 2008. A Bayesian approach to snow water equivalent reconstruction. Journal Of Geophysical Research-Atmospheres 113 (D20), D20117.

DWD, 2007. {Deutscher Wetter Dienst} (German Weather Service) Climatological data for 11 cllimate stations around the Weisseritz catchment. data.

Dyrrdal, A. V., 2009. An evaluation of Norwegian snow maps: simulation results versus observations. Hydrology Research 41 (1), 27–37.
URL http://www.iwaponline.com/nh/041/nh0410027.htm

Eckart, J., 2008. Flächendifferenzierte Beschreibung der Schneeschmelze im Einzugsgebiet der Weißeritz: Einfluss der räumlichen Variabilität von Eingangsdaten und Modell- parametern Jenny Eckart. Diplomarbeit, Technische Universität Dresden.

Egli, L., Jonas, T., Meister, R., 2009. Comparison of different automatic methods for estimating snow water equivalent. Cold Regions Science and Technology 57 (2-3), 107–115.
URL http://www.sciencedirect.com/science/article/B6V86-4VR242V-1/2/a188ee746640ee808f6576248cb7e6a6

Eichelmann, U., 2009. TUD - Research Station Oberbärenburg.
URL http://tu-dresden.de/die_tu_dresden/fakultaeten/fakultaet_forst_geo_und_hydrowissenschaften/fachrichtung_wasserwesen/ifhm/meteorologie/forschung/stationen/station_obb

Eichler, M., Francke, T., D. Kneis, Reusser, D. E., 2009a. The GOLM-database standard- a framework for time-series data management based on free software. In: Data Management Workshop, Köln.

Eichler, M., Franke, T., Kneis, D., Reusser, D. E., 2009b. The GOLM-database standard- a framework for time-series data management based on free software. In: Geophysical Research Abstracts. Vol. 11. EGU General Assembly 2009, pp. EGU2009–8070.

Elder, K. J., Cline, D., Goodbody, A., Houser, P., Liston, G. E., Mahrt, L., Rutter, N., 2009. NASA Cold Land Processes Experiment (CLPX 2002/03): Ground-Based and Near-Surface Meteorological Observations. Journal Of Hydrometeorology 10 (1), 330–337.

Elder, K. J., Dozier, J., Michaelsen, J., 1991. Snow Accumulation and Distribution in an Alpine Watershed. Water Resources Research 27 (7), 1541–1552.
URL http://dx.doi.org/10.1029/91WR00506

Erickson, T. A., Williams, M. W., Winstral, A., Apr. 2005. Persistence of topographic controls on the spatial distribution of snow in rugged mountain terrain, Colorado, United States. Water Resources Research 41 (4), W04014.

Essery, R. L. H., Blyth, E., Harding, R., Lloyd, C., 2005. Modelling albedo and distributed snowmelt across a low hill in Svalbard. Nordic Hydrology 36 (3), 207–218.

Essery, R. L. H., Pomeroy, J., 2004. Implications of spatial distributions of snow mass and melt rate for snow-cover depletion: theoretical considerations. Annals Of Glaciology, Vol 38, 2004 38, 261–265.

Fang, S. F., Gertner, G. Z., Shinkareva, S., Wang, G. X., Anderson, A., Aug. 2003. Improved generalized Fourier amplitude sensitivity test (FAST) for model assessment. Statistics And Computing 13 (3), 221–226.

Faria, D. A., Pomeroy, J. W., Essery, R. L. H., 2000. Effect of covariance between ablation and snow water equivalent on depletion of snow-covered area in a forest. Hydrological Processes 14 (15), 2683–2695.

Farinotti, D., Magnusson, J., Huss, M., Bauder, A., 2010. Snow accumulation distribution inferred from time-lapse photography and simple modelling. Hydrological Processes 24 (15), 2087–2097.
URL http://doi.wiley.com/10.1002/hyp.7629

Fenicia, F., Savenije, H. H. G., Matgen, P., Pfister, L., 2006. Is the groundwater reservoir linear? Learning from data in hydrological modelling. Hydrology And Earth System Sciences 10 (1), 139–150.

Fenicia, F., Savenije, H. H. G., Matgen, P., Pfister, L., 2008. Understanding catchment behavior through stepwise model concept improvement. Water Resources Research 44 (1), W01402.

Ferguson, R. I., 1999. Snowmelt runoff models. Progress In Physical Geography 23 (2), 205–227.

Foglia, L., Hill, M. C., Mehl, S. W., Burlando, P., 2009. Sensitivity analysis, calibration, and testing of a distributed hydrological model using error-based weighting and one objective function. Water Resources Research 45, —-.
URL http://dx.doi.org/10.1029/2008WR007255

Francke, T., 2002. Kurzbeschreibung zum Interpolationsmodul interpol v3.

Frey, H. C., Patil, S. R., 2002. Identification and review of sensitivity analysis methods. Risk Analysis 22 (3), 553–578.

Fukusako, S., 1990. Thermophysical Properties Of Ice, Snow, And Sea Ice. International Journal Of Thermophysics 11 (2), 353–372.

Gabel, B. L. K., 2000. Lawinenhandbuch, 7th Edition. Tyrolia Verlag.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B., 2003. Bayesian Data Analysis, Second Edition (Texts in Statistical Science), 2nd Edition. Chapman & Hall/CRC.
URL http://www.worldcat.org/isbn/158488388X

Gilks, W. R., Richardson, S., 1995. Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics (Chapman & Hall/CRC Interdisciplinary Statistics Series), 1st Edition. Chapman & Hall/CRC.
URL http://www.worldcat.org/isbn/0412055511

Graeff, T., Zehe, E., Reusser, D. E., Lück, E., Schröder, B., Bronstert, A., Wenk, G., John, H., 2009. Process identification through rejection of model structures in a mid-mountainous rural catchment: observations of rainfall-runoff response, geophysical conditions and model intercomparison. Hydrological Processes 23 (5), 702–718.

Gupta, H., 2003. Reply to comment by K. Beven and P. Young on "Bayesian recursive parameter estimation for hydrologic models". Water Resources Research 39 (5), 1–5.
URL http://www.agu.org/pubs/crossref/2003/2002WR001405.shtml

Gupta, H. V., Beven, K. J., Wagener, T., 2005. Encyclopedia of Hydrological Sciences. John Wiley & Sons, Ch. 131: Model, pp. 1–17.

Gupta, H. V., Kling, H., Yilmaz, K. K., Martinez-Baquero, G. F., 2009. Decomposition of the Mean Squared Error & NSE Performance Criteria: Implications for Improving Hydrological Modelling. Journal of Hydrology In Press,, —-.
URL http://www.sciencedirect.com/science/article/B6V6C-4WYDN5X-3/2/3d97c1407d6209e598d2056bd2a3267b

Gupta, H. V., Sorooshian, S., Yapo, P. O., Apr. 1998. Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. Water Resources Research 34 (4), 751–763.

Gupta, H. V., Wagener, T., Liu, Y. Q., Aug. 2008. Reconciling theory with observations: elements of a diagnostic approach to model evaluation. Hydrological Processes 22 (18), 3802–3813.

Hardy, J. P., Melloh, R., Koenig, G., Marks, D., Winstral, A., Pomeroy, J. W., Link, T., 2004. Solar radiation transmission through conifer canopies. Agricultural And Forest Meteorology 126 (3-4), 257–270.

Hart, J., Martinez, K., 2006. Environmental Sensor Networks: A revolution in the earth system science? Earth-Science Reviews 78 (3-4), 177–191.

URL http://dx.doi.org/10.1016/j.earscirev.2006.05.001

Haykin, S., 1999. Neural networks - A comprehensive foundation, 2nd Edition. Prentice-Hall, Ch. 9. Self-or.

Hedstrom, N. R., Pomeroy, J. W., 1998. Measurements and modelling of snow interception in the boreal forest. Hydrological Processes 12 (10-11), 1611–1625.
URL http://dx.doi.org/10.1002/(SICI)1099-1085(199808/09)12:10/11<1611::AID-HYP684>3.0.CO;2-4

Herbst, M., Casper, M. C., 2008. Towards model evaluation and identification using Self-Organizing Maps. Hydrology And Earth System Sciences 12 (2), 657–667.

Herpertz, D., 2001. Schneehydrologische Modellierung im Mittelgebirgsraum. Phd thesis, Friedrich-Schiller-Universitaet Jena.

Hiemstra, C. A., Liston, G. E., Reiners, W. A., 2006. Observing, modelling, and validating snow redistribution by wind in a Wyoming upper treeline landscape. Ecological Modelling 197 (1-2), 35–51.

Hiemstra, P. H., Pebesma, E. J., Twenhöfel, C. J. W., Heuvelink, G. B. M., 2008. Real-time automatic interpolation of ambient gamma dose rates from the Dutch Radioactivity Monitoring Network. Computers & Geosciences.

Hock, R., Nov. 2003. Temperature index melt modelling in mountain areas. Journal Of Hydrology 282 (1-4), 104–115.

Hood, J. L., Hayashi, M., Jun. 2010. Assessing the application of a laser rangefinder for determining snow depth in inaccessible alpine terrain. Hydrology and Earth System Sciences 14 (6), 901–910.
URL http://www.hydrol-earth-syst-sci.net/14/901/2010/

Hornberger, G. M., Spear, R. C., 1981. An Approach To The Preliminary-Analysis Of Environmental Systems. Journal Of Environmental Management 12 (1), 7–18.

Horsburgh, J. S., Tarboton, D. G., Maidment, D. R., Zaslavsky, I., 2008. A relational model for environmental and water resources data. Water Resources Research 44, W05406.
URL http://dx.doi.org/10.1029/2007WR006392

Hungerbühler, D., Steinmann, M., Winkler, W., Seward, D., Egüez, A., Peterson, D., Helg, U., Hammer, C., Jan. 2002. Neogene stratigraphy and Andean geodynamics of southern Ecuador. Earth-Science Reviews 57 (1-2), 75–124.
URL http://linkinghub.elsevier.com/retrieve/pii/S001282520100071X

Ibbitt, R., Henderson, R., Copeland, J., Wratt, D., Dec. 2000. Simulating mountain runoff with meso-scale weather model rainfall estimates: a New Zealand experience. Journal of Hydrology 239 (1-4), 19–32.
URL http://linkinghub.elsevier.com/retrieve/pii/S0022169400003516

Ihaka, R., Gentleman, R., 1996. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics 5 (3), 299–314.
URL http://www.amstat.org/publications/jcgs/

Iroumé, A., 2003. Transporte de sedimentos en una cuenca de monta\'na en la Cordillera de los Andes de la Novena Reg\'ion de Chile. Bosque 24 (1), 125–135.

Jachner, S., van Den Boogaart, K. G., Petzoldt, T., 2007. Statistical Methods for the Qualitative Assessment of Dynamic Models with Time Delay (R Package qualV). Journal of Statistical Software 22, 1–30.

Jackson, B., Schaefli, B., Gupta, H., More, 2010. What is the information content of hydrologic data? Advances In Water Resources, submitted.

Jonas, T., Marty, C., Magnusson, J., Nov. 2009. Estimating the snow water equivalent from snow depth measurements in the Swiss Alps. Journal of Hydrology 378 (1-2), 161–167.
URL http://linkinghub.elsevier.com/retrieve/pii/S0022169409005848

Jost, G., Dan Moore, R., Weiler, M., Gluns, D. R., Alila, Y., Sep. 2009. Use of distributed snow measurements to test and improve a snowmelt model for predicting the effect of forest clear-cutting. Journal of Hydrology 376 (1-2), 94–106.
URL http://linkinghub.elsevier.com/retrieve/pii/S0022169409004090

Jost, G., Weiler, M., Gluns, D. R., Alila, Y., 2007. The influence of forest and topography on snow accumulation and melt at the watershed-scale. Journal of Hydrology 347 (1-2), 101–115.
URL http://www.sciencedirect.com/science/article/B6V6C-4PN060S-4/2/abf3a42a8688f61ea05912b4c39be8b2

Kalteh, A. M., Hjorth, P., Berndtsson, R., 2008. Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application. Environmental Modelling & Software 23 (7), 835–845.
URL http://www.sciencedirect.com/science/article/B6VHC-4R5G3M5-2/2/5a6d4a37f8c65e2b6fe7bd0fd0352a43

Kass, R., Carlin, B., Gelman, A., Neal, R., 1998. Markov chain monte carlo in practice: A roundtable discussion. The American Statistician 52 (2).
URL http://www.questia.com/PM.qst?a=o&se=gglsc&d=5001349516

Kavetski, D., Kuczera, G., Franks, S. W., 2006a. Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. Water Resources Research 42 (3), W03407.

Kavetski, D., Kuczera, G., Franks, S. W., 2006b. Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. WATER RESOURCES RESEARCH 42 (3), W03408.

Kirchner, J. W., 2006. Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. Water Resources Research 42 (3).
URL http://www.agu.org/pubs/crossref/2006/2005WR004362.shtml

Klaus, J., Zehe, E., Apr. 2010. Modelling rapid flow response of a tile-drained field site using a 2D physically based model: assessment of 'equifinal' model setups. Hydrological Processes 24 (12), 1595–1609.
URL http://doi.wiley.com/10.1002/hyp.7687

Kleinn, J., Frei, C., Gurtz, J., Luthi, D., Vidale, P. L., Schar, C., 2005. Hydrologic simulations in the Rhine basin driven by a regional climate model. Journal Of Geophysical Research-Atmospheres 110 (D4), D04102.

Kleinn, J., Frei, C., Gurtz, J., Vidale, P. L., Schär, C., 2003. Klimaänderungen und extreme Flusswassermengen. Klima – Wasser – Flussgebietsmangement – im Lichte der Flut. Forum für Hydrologie und Wasserbewirtschaftung 04.03, 43–50.

Kneis, D., Heistermann, M., 2009. Quality assessment of radar-based precipitation estimates with the example of a small catchment. Hydrologie und Wasserbewirtschaftung/Hydrology and Water Resources Management-Germany 53 (3).
URL http://www.csa.com/partners/viewrecord.php?requester=gs&collection=ENV&recid=10276795

Kohonen, T., 1995. Self-Organizing Maps. In: Series in Information Sciences, second ed. Edition. Vol. 30. Springer, Heidelberg.

Kolberg, S. A., Gottschalk, L., 2006. Updating of snow depletion curve with remote sensing data. Hydrological Processes 20 (11), 2363–2380.

Kolberg, S. A., Rue, H., Gottschalk, L., 2006. A Bayesian spatial assimilation scheme for snow coverage observations in a gridded snow model. Hydrology And Earth System Sciences 10 (3), 369–381.

Krause, P., Boyle, D. P., Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Advances in Geosciences 5, 89–97.
URL http://www.adv-geosci.net/5/89/2005/

Kuczera, G., Kavetski, D., Franks, S., Thyer, M., Nov. 2006. Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. Journal of Hydrology 331 (1-2), 161–177.
URL http://www.sciencedirect.com/science/article/B6V6C-4KBVX50-2/2/b066c6156e2b79860d5e07dd57111ff2

Lee, H., Sivapalan, M., Zehe, E., 2007. Predictions of rainfall-runoff response and soil moisture dynamics in a microscale catchment using the CREW model. Hydrology and Earth System Sciences 11, 819–849.

Lehning, M., Bartelt, P., Brown, B., Fierz, C., Satyawali, P. K., Nov. 2002. A physical SNOWPACK model for the Swiss avalanche warning Part II: Snow microstructure. Cold Regions Science And Technology 35 (3), 147–167.

LfUG, 2007. {Landesamt für Umwelt und Geologie Sachsen (State office for environment and geology)}, Data about land use, soils, discharge, and the digital elevation model. data.

Lindenmaier, F., Zehe, E., Dittfurth, A., Ihringer, J., 2005. Process identification at a slow-moving landslide in the Vorarlberg Alps. Hydrological Processes 19 (8), 1635–1651.

Liston, G. E., 1999. Interrelationships among snow distribution, snowmelt, and snow cover depletion: Implications for atmospheric, hydrologic, and ecologic modeling. Journal Of Applied Meteorology 38 (10), 1474–1487.

Liu, Y., Others, 2010. Dynamic Parameter Analysis for Hydrological and Environmental Model Diagnostics and Improvement. Advances in Water Resources, submitted.

López-Moreno, J. I., Latron, J., 2008. Influence of canopy density on snow distribution in a temperate mountain range. Hydrological Processes 22 (1), 117–126.
URL http://dx.doi.org/10.1002/hyp.6572

Ludwig, K., Bremicker, M., 2007. The Water Balance Model LARSIM. In: Ludwig, K., Bremicker, M. (Eds.), Freiburger Schriften zur Hydrologie. Vol. 22. Institut für Hydrologie der Universität Freiburg.

Lundberg, A., Granlund, N., Gustafsson, D., 2010. Towards automated 'Ground truth' snow measurements-a review of operational and new measurement methods for Sweden, Norway, and

Finland. Hydrological Processes 9999 (9999), n/a–n/a.
URL http://www3.interscience.wiley.com/journal/123410780/abstract

Lundberg, A., Richardson-Näslund, C., Andersson, C., 2006. Snow density variations: consequences for ground-penetrating radar. Hydrological Processes 20 (7), 1483–1495.
URL http://www3.interscience.wiley.com/journal/112221721/abstract

Lundquist, J. D., Cayan, D., Dettinger, M., 2003. Information Processing in Sensor Networks. Vol. 2634 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg.
URL http://www.springerlink.com/content/8gcqv9dmbaj9km6v

Lundquist, J. D., Lott, F., 2008. Using inexpensive temperature sensors to monitor the duration and heterogeneity of snow-covered areas. Water Resources Research 44, —-.
URL http://dx.doi.org/10.1029/2008WR007035

MacDonald, M. K., Pomeroy, J. W., Pietroniro, A., Aug. 2009. Parameterizing redistribution and sublimation of blowing snow for hydrological models: tests in a mountainous subarctic catchment. Hydrological Processes 23 (18), 2570–2583.
URL http://doi.wiley.com/10.1002/hyp.7356

MacDonald, M. K., Pomeroy, J. W., Pietroniro, A., Feb. 2010. Hydrological response unit-based blowing snow modelling over an alpine ridge. Hydrology and Earth System Sciences Discussions 7 (1), 1167–1208.
URL http://www.hydrol-earth-syst-sci-discuss.net/7/1167/2010/

Martinec, J., Rango, A., 1986. Parameter values for snowmelt runoff modelling. Journal of Hydrology 84, 197–219.

Matott, L. S., Babendreier, J. E., Purucker, S. T., 2009. Evaluating uncertainty in integrated environmental models: A review of concepts and tools. Water Resources Research 45, —-.
URL http://dx.doi.org/10.1029/2008WR007301

McIntyre, N. R., Wagener, T., Wheater, H. S., Chapra, S. C., Apr. 2003. Risk-based modelling of surface water quality: a case study of the Charles River, Massachusetts. Journal Of Hydrology 274 (1-4), 225–247.

McRae, G. J., Tilden, J. W., Seinfeld, J. H., 1982. Global sensitivity analysis - a computational implementation of the Fourier amplitude sensitivity test ({FAST}). Comput. Chem. Eng. 6, 15–25.

Merz, B., 2006. Hochwasserrisiken-Möglichkeiten und Grenzen der Risikoabschätzung. Schweizerbart'Sche Verlagsbuchhandlung.
URL http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Hochwasserrisiken+-+Möglichkeiten+und+Grenzen+der+Risikoabschätzung#1

Molotch, N. P., Bales, R. C., Nov. 2005. Scaling snow observations from the point to the grid element: Implications for observation network design. Water Resources Research 41 (11), W11421.

Molotch, N. P., Bales, R. C., 2006. SNOTEL representativeness in the Rio Grande headwaters on the basis of physiographics and remotely sensed snow cover persistence. Hydrological Processes 20 (4), 723–739.

Montanari, A., 2007. What do we mean by 'uncertainty'? The need for a consistent wording about

uncertainty assessment in hydrology. Hydrological Processes 21, 841–845.

Montanari, A., Shoemaker, C. A., van de Giesen, N., 2009. Introduction to special section on Uncertainty Assessment in Surface and Subsurface Hydrology: An overview of issues and challenges. Water Resources Research 45, —-.
URL http://dx.doi.org/10.1029/2009WR008471

Nash, J. E., Sutcliffe, J. V., Apr. 1970. River flow forecasting through conceptual models part I – A discussion of principles. Journal of Hydrology 10 (3), 282–290.
URL http://www.sciencedirect.com/science/article/B6V6C-487FF7C-1XH/1/75ac51a8910cad95dda46f4756e7a800

Niehoff, D., Bronstert, A., 2001. Influences of land use and land cover conditions on flood generation: A simulation study. Advances In Urban Stormwater And Agricultural Runoff Source Controls 6, 267–278.

Niehoff, D., Fritsch, U., Bronstert, A., 2002. Land-use impacts on storm-runoff generation: scenarios of land-use change and simulation of hydrological response in a meso-scale catchment in SW-Germany. Journal of Hydrology 267 (1-2), 80–93.
URL http://www.sciencedirect.com/science/article/B6V6C-46HBKF8-2/2/e7d43db548caa8d7c0ee195052aa4e98

Pappenberger, F., Beven, K. J., 2006. Ignorance is bliss: Or seven reasons not to use uncertainty analysis. WATER RESOURCES RESEARCH 42 (5), W05302.

Pappenberger, F., Beven, K. J., Ratto, M., Matgen, P., 2008. Multi-method global sensitivity analysis of flood inundation models. Advances In Water Resources 31, 1–14.

Pappenberger, F., Iorgulescu, I., Beven, K. J., 2006. Sensitivity analysis based on regional splits and regression trees (SARS-RT). Environmental Modelling & Software 21 (7), 976–990.

Pebesma, E. J., Switzer, P., Loague, K., 2005. Error analysis for the evaluation of model performance: rainfall-runoff event time series data. Hydrological Processes 19 (8), 1529–1548.

Pepe, A., Mayernik, M. S., Borgman, C. L., de Sompel, H. V., 2009. Technology to Represent Scientific Practice: Data, Life Cycles, and Value Chains. CoRR abs/0906.2.

Plummer, M., Best, N., Cowles, K., Vines, K., 2009. coda: Output analysis and diagnostics for MCMC.

Pöhler, H. A., 2006. Anpassung von WaSiM-ETH und die Erstellung und Berechnung von Landnutzungs und Klimaszenarien für die Niederschlag-Abfluss-Modellierung am Beispiel des Osterzgebirges. Phd thesis, TU Freiberg.

Pomeroy, J., Essery, R. L. H., Toth, B., 2004. Implications of spatial distributions of snow mass and melt rate for snow-cover depletion: observations in a subarctic mountain catchment. Annals Of Glaciology 38, 195–201.

Pomeroy, J. W., Parviainen, J., Hedstrom, N. R., Gray, D. M., 1998. Coupled modelling of forest snow interception and sublimation. HYDROLOGICAL PROCESSES 12, 2317–2337.

Prokop, A., Nov. 2008. Assessing the applicability of terrestrial laser scanning for spatial snow depth measurements. Cold Regions Science and Technology 54 (3), 155–163.
URL http://linkinghub.elsevier.com/retrieve/pii/S0165232X08001018

R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. R

Foundation for Statistical Computing, Vienna, Austria.
URL http://www.r-project.org

Rango, A., Martinec, J., Aug. 1995. Revisiting The Degree-Day Method For Snowmelt Computations. Water Resources Bulletin 31 (4), 657–669.

Reichert, P., Feb. 2006. A standard interface between simulation programs and systems analysis software. Water Science & Technology 53 (1), 267.
URL http://www.iwaponline.com/wst/05301/wst053010267.htm

Reichert, P., Mieleitner, J., 2009. Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters. Water Resources Research 45, —-.
URL http://dx.doi.org/10.1029/2009WR007814

Reusser, D. E., 2008. Implementation of the Fourier Amplitute Sensitivity Test (FAST).

Reusser, D. E., 2009. {TIGER}: Analysing TIme series of Grouped ERrors.

Reusser, D. E., Blume, T., Schaefli, B., Zehe, E., 2009. Analysing the temporal dynamics of model performance for hydrological models. Hydrology And Earth System Sciences 13, 999–1018.

Reusser, D. E., Buytaert, W., 2010. RHydro: Classes for hydrological modelling and analysis.

Reusser, D. E., Buytaert, W., Zehe, E., 2010a. Temporal dynamics of model parameter sensitivity for computationally expensive models with FAST (Fourier Amplitude Sensitivity Test). WRR, submitted.

Reusser, D. E., Francke, T., 2008. wasim: Helpful tools for WaSiM-ETH.

Reusser, D. E., Lüdtke, S., Pagel, J., Zehe, E., 2010b. Spatio-temporal variability of snow in the Weisseritz catchment. dontknow, inPrep.

Reusser, D. E., Zehe, E., 2010a. Inferring model structural deficits by analyzing temporal dynamics of model performance and parameter sensitivity. WRR, submitted.

Reusser, D. E., Zehe, E., 2010b. Low-cost monitoring of snow height and thermal properties with inexpensive temperature sensors. Hydrological Processes, submitted.

Richter, D., 1995. Ergebnisse methodischer Untersuchungen zur Korrektur des systematischen Meßfehlers des Hellmann-Niederschlagsmessers.

Saltelli, A., 2002. Making best use of model evaluations to compute sensitivity indices. Computer Physics Communications 145 (2), 280–297.

Saltelli, A., Bolado, R., 1998. An alternative way to compute Fourier amplitude sensitivity test (FAST). Computational Statistics & Data Analysis 26 (4), 445–460.

Saltelli, A., Ratto, M., Tarantola, S., Campolongo, F., 2006. Sensitivity analysis practices: Strategies for model-based inference. Reliability Engineering & System Safety 91 (10-11), 1109–1125.

Saltelli, A., Tarantola, S., Campolongo, F., Nov. 2000. Sensitivity analysis as an ingredient of modeling. Statistical Science 15 (4), 377–395.

Saltelli, A., Tarantola, S., Chan, K. P.-S., 1999. A Quantitative Model-Independent Method for Global Sensitivity Analysis of Model Output. Technometrics 41 (1), 39.
URL http://www.jstor.org/stable/1270993?origin=crossref

Satyawali, P. K., Singh, A. K., Aug. 2008. Dependence of thermal conductivity of snow on mi-

crostructure. Journal Of Earth System Science 117 (4), 465–475.

Schaefli, B., Gupta, H. V., 2007. Do Nash values have value? Hydrological Processes 21, 2075–2080.

Schaffhauser, A., Adams, M., Fromm, R., Jorg, P., Luzi, G., NOFERINI, L., SAILER, R., Nov. 2008. Remote sensing based retrieval of snow cover properties. Cold Regions Science and Technology 54 (3), 164–175.
URL http://dx.doi.org/10.1016/j.coldregions.2008.07.007

Schaibly, J. H., Shuler, K. E., 1973. Study Of Sensitivity Of Coupled Reaction Systems To Uncertainties In Rate Coefficients .2. Applications. Journal Of Chemical Physics 59 (8), 3879–3888.

Schneebeli, M., Sokratov, S. A., 2004. Tomography of temperature gradient metamorphism of snow and associated changes in heat conductivity. Hydrological Processes 18 (18), 3655–3665.

Schulla, J., Jasper, K., 2001. Model Description WaSiM-ETH.

Seibert, J., Beven, K. J., 2009. Gauging the ungauged basin: how many discharge measurements are needed? HYDROLOGY AND EARTH SYSTEM SCIENCES 13 (6), 883–892.
URL http://www.hydrol-earth-syst-sci.net/13/883/2009/hess-13-883-2009.html

Shamir, E., Imam, B., Gupta, H. V., Sorooshian, S., 2005. Application of temporal streamflow descriptors in hydrologic model parameter estimation. Water Resources Research 41 (6), W06021.

Sieber, A., Bremicker, M., 2006. Verifikation von DWD-LM/LME und Meteomedia-EZMOS-Niederschlagsvorhersagen im Hinblick auf die Hochwasservorhersage. Tech. Rep. 43.2, Landesanstalt für Umwelt, Messungen und Naturschutz Baden-Württemberg (LUBW).

Sieber, A., Uhlenbrook, S., Aug. 2005. Sensitivity analyses of a distributed catchment model to verify the model structure. Journal Of Hydrology 310 (1-4), 216–235.

Singh, A. K., 1999. An investigation of the thermal conductivity of snow. Journal Of Glaciology 45 (150), 346–351.

Slater, A. G., Clark, M. P., 2006. Snow data assimilation via an ensemble Kalman filter. Journal Of Hydrometeorology 7 (3), 478–493.

Sobol, I. M., Feb. 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Mathematics and Computers in Simulation 55 (1-3), 271–280.
URL http://linkinghub.elsevier.com/retrieve/pii/S0378475400002706

SRTM, 2002. Shuttle Radar Topography Mission (SRTM) Elevation Data Set. dataset.

Stähli, M., Jonas, T., Gustafsson, D., Aug. 2009. The role of snow interception in winter-time radiation processes of a coniferous sub-alpine forest. Hydrological Processes 23 (17), 2498–2512.
URL http://doi.wiley.com/10.1002/hyp.7180

Stewart, R. E., Jul. 1985. Precipitation types in winter storms. Pure and Applied Geophysics PAGEOPH 123 (4), 597–609.
URL http://www.springerlink.com/content/v372761422171m28

Storck, P., Lettenmaier, D. P., Bolton, S. M., Nov. 2002. Measurement of snow interception and canopy effects on snow accumulation and melt in a mountainous maritime climate, Oregon, United States. Water Resources Research 38, 1223.

URL http://dx.doi.org/10.1029/2002WR001281

Sturm, M., Holmgren, J., Konig, M., Morris, K., 1997. The thermal conductivity of seasonal snow. Journal Of Glaciology 43 (143), 26–41.

Sturm, M., Perovich, D. K., Holmgren, J., 2002. Thermal conductivity and heat transfer through the snow on the ice of the Beaufort Sea. Journal Of Geophysical Research-Oceans 107 (C21), 8043.

Tang, Y., Reed, P., van Werkhoven, K., Wagener, T., 2007a. Advancing the identification and evaluation of distributed rainfall-runoff models using global sensitivity analysis. Water Resources Research 43, W06415.

Tang, Y., Reed, P., Wagener, T., van Werkhoven, K., 2007b. Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation. Hydrology And Earth System Sciences 11 (2), 793–817.

Thiemann, M., Trosset, M., Gupta, H., Sorooshian, S., 2001. Bayesian recursive parameter estimation for hydrologic models. Water Resources Research 37 (10), 2521.
URL http://www.agu.org/pubs/crossref/2001/2000WR900405.shtml

Todini, E., 2007. Hydrological catchment modelling: past, present and future. Hydrology And Earth System Sciences 11 (1), 468–482.

TU-Dresden, 2010. {Technischen Universität Dresden} Precipitation data for 4 climate stations around the {W}eisseritz catchment. {D}ata.

van de Giesen, N., Andreini, M., Selker, J., 2009a. Trans-African Hydro-Meteorological Observatory. In: Geophysical Research Abstracts. Vol. 11. pp. EGU2009–13512.

van de Giesen, N., Degen, C., Hut, R., 2009b. Affordable Acoustic Disdrometer: Design, Calibration, Tests. AGU Fall Meeting Abstracts, F903+.

van den Boogaart, K. G., Jachner, S., Petzoldt, T., 2007. qualV: Qualitative Validation Methods.

Van Griensven, A., Meixner, T., Grunwald, S., Bishop, T., Diluzio, M., Srinivasan, R., 2006. A global sensitivity analysis tool for the parameters of multi-variable catchment models. Journal of Hydrology 324 (1-4), 10–23.
URL http://linkinghub.elsevier.com/retrieve/pii/S0022169405004488

van Werkhoven, K., Wagener, T., Reed, P., Tang, Y., 2009. Sensitivity-guided reduction of parametric dimensionality for multi-objective calibration of watershed models. Advances in Water Resources 32 (8), 1154–1169.
URL http://www.sciencedirect.com/science/article/B6VCF-4VVGKP4-1/2/1de75f731fc9db8ceb89067b5b35abaf

Vrugt, J. A., Gupta, H. V., Bastidas, L. A., Bouten, W., Sorooshian, S., Aug. 2003. Effective and efficient algorithm for multiobjective optimization of hydrologic models. Water Resources Research 39 (8), 1214.

Wagener, T., Freer, J., Zehe, E., Beven, K. J., Gupta, H., Bardossy, A., 2006. Towards an uncertainty framework for predictions in ungauged basins: The Uncertainty Working Group. In: Sivapalan, M., Wagener, T., Uhlenbrook, S., Liang, X., Lakshmi, V., Kumar, P., Zehe, E., Tachikawa, Y. (Eds.), Predictions in ungauged basins: promise and progress, IAHS Publication no. 303. IAHS Press, p. 454.
URL http://books.google.com/books?hl=en&lr=&id=1wUjy_2M_34C&oi=fnd&pg=PA454&dq=Towards+an+

uncertainty+framework+for+predictions+in+ungauged+basins:+the+uncertainty+working+group.+In&ots=ZHJUQliO0l&sig=sA4w1TShMWIksIvkeWiGKcfpRfA

Wagener, T., McIntyre, N. R., Lees, M. J., Wheater, H. S., Gupta, H. V., 2003. Towards reduced uncertainty in conceptual rainfall-runoff modelling: dynamic identifiability analysis. Hydrological Processes 17 (2), 455–476.
URL http://dx.doi.org/10.1002/hyp.1135

WASY, 2006. Schätzung dominanter {A}bflussprozesse mit {WBS} {FLAB} (Assessment of dominant runoff processes with {WBS FLAB}). Tech. rep., WASY Gesellschaft für wasserwirtschaftliche Planung und Systemforschung mbH and Internationales Hochschulinstitut Zittau.

Watson, F. G. R., Anderson, T. N., Newman, W. B., Alexander, S. E., Garrott, R. A., Sep. 2006a. Optimal sampling schemes for estimating mean snow water equivalents in stratified heterogeneous landscapes. Journal Of Hydrology 328 (3-4), 432–452.

Watson, F. G. R., Newman, W. B., Coughlan, J. C., Garrott, R. A., 2006b. Testing a distributed snowpack simulation model against spatial observations. Journal of Hydrology 328 (3-4), 453–466.
URL http://www.sciencedirect.com/science/article/B6V6C-4JD0JD7-1/2/a733e024e55b39eb9b03771db6c51715

Weihs, C., Ligges, U., Luebke, K., Raabe, N., 2005. klaR Analyzing German Business Cycles. In: Baier, D., Decker, R., Schmidt-Thieme, L. (Eds.), Data Analysis and Decision Support. Springer-Verlag, Berlin, pp. 335–343.

Wilks, D., Sep. 2006. Comparison of ensemble-MOS methods in the Lorenz '96 setting. Meteorological Applications 13 (3), 243–256.

Williams, M. W., Cline, D., Hartman, M., Bardsley, T., 1999. Data for snowmelt model development, calibration, and verification at an alpine site, Colorado Front Range. Water Resources Research 35 (10), 3205–3209.

Winkler, R. D., Moore, R. D., 2006. Variability in snow accumulation patterns within forest stands on the interior plateau of British Columbia, Canada. Hydrological Processes 20 (17), 3683–3695.
URL http://doi.wiley.com/10.1002/hyp.6382

Winkler, R. D., Spittlehouse, D. L., Golding, D. L., 2005. Measured differences in snow accumulation and melt among clearcut, juvenile, and mature forests in southern British Columbia. Hydrological Processes 19 (1), 51–62.
URL http://dx.doi.org/10.1002/hyp.5757

Winsemius, H. C., Schaefli, B., Montanari, A., Savenije, H. H. G., 2009. On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information. Water Resources Research 45, —-.

Winstral, A., Elder, K. J., Davis, R. E., 2002. Spatial snow modeling of wind-redistributed snow using terrain-based parameters. Journal Of Hydrometeorology 3 (5), 524–538.

Wittenberg, H., Sivapalan, M., 1999. Watershed groundwater balance estimation using streamflow recession analysis and baseflow separation. Journal Of Hydrology 219 (1-2), 20–33.

Xie, X. L., Beni, G., 1991. A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (8), 841–847.

Xu, N., 2002. A survey of sensor network applications. IEEE Communications Magazine 40 (8), 102–114.
URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.131.9647&rep=rep1&type=pdf

Yan, J., 2004. som: Self-Organizing Map.

Yapo, P. O., Gupta, H. V., Sorooshian, S., 1998. Multi-objective global optimization for hydrologic models. Journal Of Hydrology 204 (1-4), 83–97.

Zehe, E., Becker, R., Bardossy, A., Plate, E., 2005. Uncertainty of simulated catchment runoff response in the presence of threshold processes: Role of initial soil moisture and precipitation. Journal of Hydrology 315 (1-4), 183–202.

Zehe, E., Blöschl, G., 2004. Predictability of hydrologic response at the plot and catchment scales: Role of initial conditions. Water Resources Research 40 (10), W10202.

Zehe, E., Elsenbeer, H., Lindenmaier, F., Schulz, K., Bloschl, G., 2007. Patterns of predictability in hydrological threshold systems. Water Resources Research 43 (7), W07434.

Zehe, E., Fluhler, H., 2001. Preferential transport of isoproturon at a plot scale and a field scale tile-drained site. Journal Of Hydrology 247 (1-2), 100–115.

Zehe, E., Graeff, T., Morgner, M., Bauer, A., Bronstert, A., Jun. 2010. Plot and field scale soil moisture dynamics and subsurface wetness control on runoff generation in a headwater in the Ore Mountains. Hydrol. Earth Syst. Sci. 14 (6), 873–889.
URL http://www.hydrol-earth-syst-sci.net/14/873/2010/http://www.hydrol-earth-syst-sci.net/14/873/2010/hess-14-873-2010.pdfDO-10.5194/hess-14-873-2010

Zehe, E., Lee, H., Sivapalan, M., 2006. Dynamical process upscaling for deriving catchment scale state variables and constitutive relations for meso-scale process models. Hydrology And Earth System Sciences 10 (6), 981–996.

Zehe, E., Maurer, T., Ihringer, J., Plate, E., 2001. Modeling water flow and mass transport in a loess catchment. Physics and Chemistry of the Earth Part B-Hydrology Oceans and Atmosphere 26 (7-8), 487–507.

Zeileis, A., Grothendieck, G., 2005. zoo: S3 Infrastructure for Regular and Irregular Time Series. Journal of Statistical Software 14 (6), 1–27.
URL http://www.jstatsoft.org/v14/i06/

Acknowledgments

I want to thank my advisor Erwin Zehe for providing the possibility to work in a very suitable environment, for his continuing encouragement and support and the many valuable discussions and his feedback. He had advice for all the situations I encountered during the work on this thesis.

For taking the time to read and review my thesis I want to thank my referees Bruno Merz and Hoshin Gupta.

I'm grateful for the good collaboration and support from all the members of the OPAQUE project. Special thanks goes to the people from the project providing very valuable data and information, namely Uwe Ehret, Angela Sieber, Ulf Winkler, and Alexander Liebert. Special thanks to Axel Bronstert for his great support as Co-leader of the OPAQUE project and head of the Institue of Geoecology and the chair of Hydrology.

I enjoyed a very relaxed and collaborative environment at Institute of Geoecology which was very important to me as a good working environment. Thus, my thanks goes to my fellow PhD students and coworkers Thomas, Jan, Andi, Markus, Daniel, Hauke, Till, Maik, David, Bettina and Boris. The OPAQUE student workers Mareike, Niko and Erick were very supportive with their contribution to the field work and the data management. Many valuable discussions with Theresa Blume, Bettina Schaefli Markus Weiler and Uwe Ehret were of importance for this work.

My Diploma and project students working on parts of the project, Stefan, Sophie and Jenny for their enthusiasm and great work. Christian Rinner also contributed to the snow height estimation algorithm with his Bachelor's thesis at TU München.

Also great thanks to the numerous people involved during all the measurement campaigns. We obtained great support during the installation of the equipment from Peter Eckart, Bernd Böhme and Timo Junkers from the Landestalsperren Verwaltung Sachsen (State office for reservoir management). Prof. Matschulat, Prof. Bernhofer, Kurt Herklotz, Uwe Eichelmann and Annette Riedl were of great support during the search of a useful location for and during the installation of the snow pillow at their field station in Oberbärenburg.

Much thanks to Hoshin Gupta, Maik Heistermeister, Till Franke, David Kneis, and the anonymous reviewers for their comments on earlier versions of some of the chapters, whose comments helped to significantly improve them.

For their love and continuing encouragement throughout this work I want to thank my family Jenny, Lukas, Gertrud, Rudolf, Theresa, Benjamin and Simon. Also, I'm greatful for the support from my friends Peter, Thomas, Kristin, Annette, Britta and Kai and the friends from EPE, orienteering, salsa and rowing.

Thanks to Wouter Buytaert for the inspiring discussions and collaboration related to open source software and hydrology. A major part of the analysis was carried out with the open source statistical software R and contributed packages, so I would like to thank its community as well as Boris Schröder for proposing to use this great tool.

This study has been funded as part of OPAQUE (operational discharge and flooding predictions in head catchments), a project within the BMBF-Förderaktivität "Risikomanagement extremer Hochwasserereignisse" (RIMAX).

Author's declaration

I prepared this dissertation without illegal assistance. The work is original except where indicated by special reference in the text and no part of the dissertation has been submitted for any other degree.

This dissertation has not been presented to any other University for examination, neither in Germany nor in another country.

Dominik Reusser
Potsdam, September 2010