A guideline for hydrologically consistent models

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät

der Christian-Albrechts-Universität zu Kiel

vorgelegt von
Dipl.-Geoökol. Matthias Pfannerstill

Kiel, 2015

Erste Gutachterin: Prof. Dr. Nicola Fohrer
Zweite Gutachterin: Prof. Dr. Natascha Oppelt

Tag der mündlichen Prüfung: 03.11.2015
Zum Druck genehmigt: 03.11.2015

gez. Prof. Dr. Wolfgang J. Duschl, Dekan

Contents

Abstract

Hydrological models are established tools to simulate discharge and hydrological processes of watersheds. Scenario simulations and resource management studies are carried out, assuming that hydrological processes are simulated appropriately. However, several studies revealed that satisfying discharge reproduction may be achieved despite of inappropriate simulation of the hydrological processes.

A reason for this might be an inappropriate process representation in the model due to inadequate process equations and parameters. Consequently, model deficiencies need to be identified to improve the representation of dominant hydrological processes of the specific catchment. The ultimate goal of model improvements should be to achieve hydrological consistency, which can be defined as a realistic reproduction of all hydrological processes of the study catchment together with an appropriate discharge simulation.

In this thesis, an example serves to demonstrate how appropriate discharge and process reproduction can be achieved by modifying a hydrological model. It is shown, how the model modification leads to improved discharge simulations for a lowland catchment due to the improved ability of the model to capture the groundwater processes of the study catchment. The deficient groundwater model component was modified by implementing a more complex groundwater process representation with a fast and a slow reacting aquifer. The evaluation of the new model version revealed improved model performance for the low flow phases due to delayed groundwater contribution to the discharge at low flow events and due to fast groundwater contribution at high discharge.

To achieve satisfying simulation of all discharge phases, a newly developed multi-metric framework was applied to evaluate the modified model. The advantage of this new multi-metric framework is a fair-balanced performance evaluation for high and low discharge phases simultaneously. Additionally, a new evaluation criterion for very low flows was developed. This criterion was integrated into the multi-metric framework to evaluate very low flows together with all remaining discharge phases.

Finally, diagnostic information about simulated hydrological processes was obtained by the application of a temporal parameter sensitivity analysis. For this, temporal parameter sensitivities were calculated to derive simulated processes of the modified model. The modified model was then verified by comparing simulated processes with observations and known processes of the study catchment. In this way, proper process simulation according to the observed processes of the real-world was confirmed. Due to the verified process reproduction and the improved model performance for all discharge phases, the ability of the model to capture the dominant processes of the study catchment was proved.

As a synthesis, the provided methods that were exemplarily used to improve a hydrological model were interpreted into a more general context. This synthesis lead to the main achievement of this thesis: The individual steps of model deficiency detection, improvement, evaluation, and verification were joined to a structured guideline. It is hypothesised that this guideline is applicable to any hydrological model. Consequently, other hydrological models may benefit from this structured guideline to improve their hydrological consistency.

Zusammenfassung

Hydrologische Modelle sind etablierte Hilfsmittel um Abflüsse und hydrologische Prozesse von Einzugsgebieten zu simulieren. Die Berechnung von Szenarien und Ressourcenplanungen werden unter der Annahme vorgenommen, dass die hydrologischen Prozesse realistätsnah wiedergegeben werden. Zahlreiche Studien belegen allerdings, dass eine zufriedenstellende Abflussnachbildung trotz unzureichender Abbildung der hydrologischen Prozesse erreicht werden kann.

Ein Grund für die unzureichende Abbildung hydrologischer Prozesse im Modell kann die fehlerhafte Implementierung von Prozessgleichungen und deren Parametern sein. Diese Modelldefizite müssen identifiziert werden, um eine verbesserte Abbildung der dominanten Prozesse des untersuchten Einzugsgebiets zu gewährleisten. Das übergeordnete Ziel derartiger Modellverbesserungen sollte in diesem Zusammenhang die hydrologische Konsistenz sein, welche durch eine gemeinsame, realistische Abbildung der hydrologischen Prozesse und des Abflusses des untersuchten Einzugsgebiets definiert ist.

Diese Dissertation zeigt beispielhaft auf, wie eine zufriedenstellende Abbildung von Abflüssen und hydrologischen Prozessen durch die Modifizierung eines hydrologischen Modells erfolgen kann. Es wird aufgezeigt, wie die Modifizierung des hydrologischen Modells zu einer verbesserten Abbildung von Abflüssen für ein Tieflandeinzugsgebiet führt. Die verbesserte Abflussabbildung ist darauf zurückzuführen, dass das modifizierte Modell die Grundwasserprozesse des Einzugsgebiets besser nachbilden kann. Dafür wurde die fehlerhafte Grundwasserkomponente des hydrologischen Models verändert, indem die komplexen Grundwasserprozesse des Einzugsgebiets durch einen schnell und einen langsam reagierenden Grundwasserspeicher repräsentiert werden. Die Auswertung der neuen Modellversion zeigte eine Verbesserung der Modellgüte im Niedrigabfluss auf. Diese verbesserte Modellgüte wurde durch verzögerten Grundwasserabfluss aus dem langsamen Aquifer bei Niedrigabfluss erzielt und durch Grundwasserfluss aus dem schnell reagierenden Aquifer bei hohen Abflüssen erreicht.

Um die zufriedenstellende Abbildung aller Abflussphasen sicherzustellen, wurde das modifizierte Modell mit einem neu entwickelten Auswertungsansatz mit mehreren Gütemaßen untersucht. Der Vorteil dieses neuen Auswertungsansatzes liegt darin, dass die gleichzeitige Auswertung von Hoch- und Niedrigabflüssen bei gleicher Gewichtung vorgenommen werden kann. Zusätzlich wurde ein weiteres Gütemaß entwickelt, mit dem gezielt extreme Niedrigabflüsse ausgewertet werden können. Dieses Gütemaß wurde in den neuen Auswertungsansatz integriert, um eine zufriedenstellende Modellgüte in den extremen Niedrigwasserphasen gemeinsam mit alle weiteren Abflussphasen zu erreichen.

Abschließend wurden durch die Anwendung einer temporalen Sensitivitätsanalyse diagnostische Informationen über die simulierten hydrologischen Prozesse bestimmt. Dafür wurden temporale Parametersensitivitäten für das modifizierte Modell berechnet, um daraus simulierte Prozesse abzuleiten. Das modifizierte Modell wurde schließlich verifiziert, indem die simulierten Prozesse des veränderten Modells mit Beobachtungen und bekannten Prozessen aus dem Einzugsgebiet verglichen wurden. Dadurch wurde die adäquate Prozess-

simulation im Vergleich zu beobachteten Prozessen aus der Realität überprüft. Auf Grund der verifizierten Prozesswiedergabe und der verbesserten Modellgüte für alle Abflussphasen wurde die Fähigkeit des modifizierten Modells bestätigt, die dominanten Prozesse des Einzugsgebiets abbilden zu können.
Als Synthese dieser Arbeit wurden die Methoden dieser Arbeit, welche beispielhaft zur Erreichung der hydrologischen Konsistenz genutzt wurden, in einen generellen Kontext gesetzt. Diese Synthese führt zum übergeordneten Ergebnis dieser Dissertation: Die einzelnen Schritte, bestehend aus Identifizierung des Modelldefizits, Modellverbesserung, Auswertung der Modellgüte sowie Verifizierung, wurden zu einer strukturierten Anleitung zusammengefasst. Es wird angenommen, dass diese Anleitung auf jedes hydrologische Modell angewendet werden kann. Somit können andere hydrologische Modelle von dieser strukturierten Anleitung profitieren, um die hydrologische Konsistenz zu steigern.

Acknowledgement

It was a really exciting and challenging task for me to prepare and finish this thesis. I wish to express my gratitude to all those who accompanied me on this way and to all who contributed to the completion of this dissertation.

First and foremost, I wish to thank Prof. Dr. Nicola Fohrer for supervision and for laying the foundation of this thesis by taking me up into her working group. Although our first contact was my short mail with some attached pieces of paper asking for thesis opportunities, you took the time to support me by preparing the important proposal for a doctoral scholarship. From the very beginning of my time at Kiel you have always given me freedom to develop and realise new ideas and you supported me to become an individual and creative scientist.

I also thank Prof. Dr. Natascha Oppelt for agreeing to serve as the second reviewer of this thesis.

I sincerely thank Dr. Björn Guse who led me on the track of model diagnostics and who encouraged me to always question things critically (whether modelling issues or strategic decisions about paper problems). My papers would never have been published without your fruitful discussions and suggestions of how to deal with complex problems. Your readiness of beeing reachable at anytime (holidays, weekends, sunday crime thriller time) was a piece of luck for me and took my work forward. I really appreciate our way of working together!

Many thanks are owed to my (former) colleagues from the University of Kiel and the Department of Hydrology and Water Resources Management for creating such a pleasant work climate and for giving support for several problems in the last years. I had so many helping discussions with you about coding, proofreading, paper problems but also a great time aside from work.

I sincerely thank the German Federal Environmental Foundation (DBU) for giving me financial and ideational support to prepare this thesis. Funding of the Heinz-Wüstenberg-Stiftung allowed me to have a very educational stay at the SWAT developers of the Grassland, Soil and Water Research Laboratory of the USDA-ARS in Temple.

Last but not least I need to express my greatest appreciation to my friends and my family. My friends showed understanding for my limited free time and only little opportunities for visits. I am thankful to my parents for supporting me during my studies of geoecology and during preparation of this thesis. You always had time for encouraging telephone calls.

And of course, I am extremely grateful to my wife Ann-Christin: It was not always easy for you but you were always there for me and you motivated me to keep on going. Many holidays and many weekends of togetherness were cancelled due to urgent work for my thesis. You

always showed understanding and acceptance for giving priority to prepare this thesis. It is remarkable, how often you had to proofread so many paragraphs of discipline-specific texts wihout being despaired. Thank you so much for everything!

1 Introduction

1.1 Motivation

Hydrological models are applied to investigate several aspects of water resource management, comprising practical and scientific water-related questions (Borah and Bera, 2003). The simulation of water quantity and the identification of water flowpaths are used to analyse processes and changes in the hydrological system (e.g. Tallaksen et al., 1997; Niehoff et al., 2002; Bronstert et al., 2007; Hunter et al., 2007; Laaha and Blöschl, 2007; Volk et al., 2009; Thielen et al., 2009; Huang et al., 2013).
According to Beven (1989) simulation models aim to predict the behaviour of the real world. However, to achieve reliable and satisfying predictions the hydrological processes of the real world need to be integrated into the model appropriately. As a consequence, a precise model development lays the foundation for proper model results. For this, different development steps need to be considered and are described in the following (Fig. 1).
Based on the studies of Gupta et al. (2008) and Reusser et al. (2009), the first step of model developments is the analysis of observed and measured hydrological processes of the real world. For instance, hydrological processes that are essential to explain discharge generation for the study catchment are determined. In this way, knowledge about dominant processes within the catchment can be considered for the model development (Fig. 1a).
Recent progress in understanding of hydrological processes increased the demand to incorporate this knowledge into hydrological models (e.g. Fenicia et al.; Tetzlaff et al.; Hrachowitz et al., 2013; McMillan et al., 2013), which is especially important for the second step of model development (Fig. 1b). The detailed knowledge about hydrological processes within the catchment is transferred to a model concept that describes the individual processes preferably with physical process equations. To give an example, the exchange between the river and a groundwater aquifer may be described as a result of hydraulic gradients. This exchange may then be represented with a gradient-driven process equation. However, the degree of detail for the hydrological process representation with equations is limited since physical descriptions of processes and feedbacks require intensive investigations of the whole catchment. Furthermore, sufficient computational capacity to resolve model results in an acceptable time is often limited. As a consequence hydrological processes and their dynamics are simplified for the integration into hydrological models (Clark et al., 2008; McMillan et al., 2013), but the simplified assumptions still aim to reproduce the most important hydrological processes. Thus, it is assumed that single model equations are still able to capture the dominant processes as well as the interactions between hydrological processes (Gupta et al., 2012).
However, the formulation of single model equations to depict individual processes is just one step towards the final hydrological model. As shown by Yilmaz et al. (2008) and Reusser et al. (2009), the single model equations are hierarchically organised and joined into a model structure to describe processes with process equations and parameters (Fig. 1c). According to the purpose of the model, individual processes are emphasised and other individual processes are neglected (Fenicia et al., 2011). For example, a hydrological model that is developed

to be suitable to groundwater dominated catchments needs a strong focus on groundwater process representation with an appropriate combination of storages. This basic suitability for the study catchment can be achieved by proper model structure development.
Going one step further in the model development, the following parameterisation of the model (Fig. 1d) is another crucial step. Parameters are estimated and selected to describe temporal process dynamics (e.g. van Werkhoven et al., 2009; Pokhrel et al., 2012) and temporal hydrological patterns (e.g. discharge). By defining appropriate parameter values, specific hydrological processes may be emphasised if the model structure is capable to allow this parameter based emphasis (Yilmaz et al., 2008; Pokhrel et al., 2012).
Finally, the parameterised model can be applied to a catchment to simulate its hydrological system and processes (Fig. 1e). Preferably, the developed model should be able to reproduce observed and measured hydrological processes. In general, discharge is the mostly used criterion in hydrology to decide whether the model was properly applied and whether adequate model behaviour has been achieved (e.g. Madsen, 2000; Krause et al., 2005; Dawson et al., 2007; Pokhrel et al., 2012). Especially for the decision about adequate model behaviour, model parameters are particularly relevant. During the parameterisation, specific parameters are used to emphasise the dominance of individual processes. For example, groundwater parameters are expected to be highly relevant to describe the discharge for groundwater dominated catchments. Consequently, the behaviour of the model should be a result of interactions between temporally dynamic processes, which are expected to be represented by specific groundwater model parameters.
Due to differing relevances of the parameters to describe the behaviour of the model and the simulated discharge, parameter sensitivity analyses are a valuable method to explain simulation results. By determining the temporal sensitivity of parameters on the discharge, information about the model's behaviour is provided (e.g. Wagener et al., 2003; Herbst et al., 2009; Reusser et al., 2009; van Werkhoven et al., 2009; Garambois et al., 2013; Herman et al., 2013; Guse et al., 2014). In addition to proper discharge simulation, this information of parameter sensitivity is highly relevant to ensure that the simplified assumptions of the model structure are in accordance with the real world.

1.1.1 Diagnostic model analyses

Diagnostic model analyses help to determine if the reproduction of the governing hydrological processes have been achieved by the model and if they have been achieved for the right reason (e.g. Kirchner, 2006; Clark et al., 2008; Euser et al., 2013; Hrachowitz et al., 2014). For this, comparisons between observed and simulated hydrological data (Krause et al., 2005; Gupta et al., 2008; Bennett et al., 2013) are used to provide a first impression about the model's ability to reproduce different aspects of the hydrological system. At this stage of analysis, various characteristics of the provided data is related to expected hydrological behaviour that is based on expert-knowledge (Herbst et al., 2009; Hrachowitz et al., 2014).
Diagnostic model analyses are able to provide two core results, comprising model performance and hydrological process simulation. These core results are needed for further interpretation

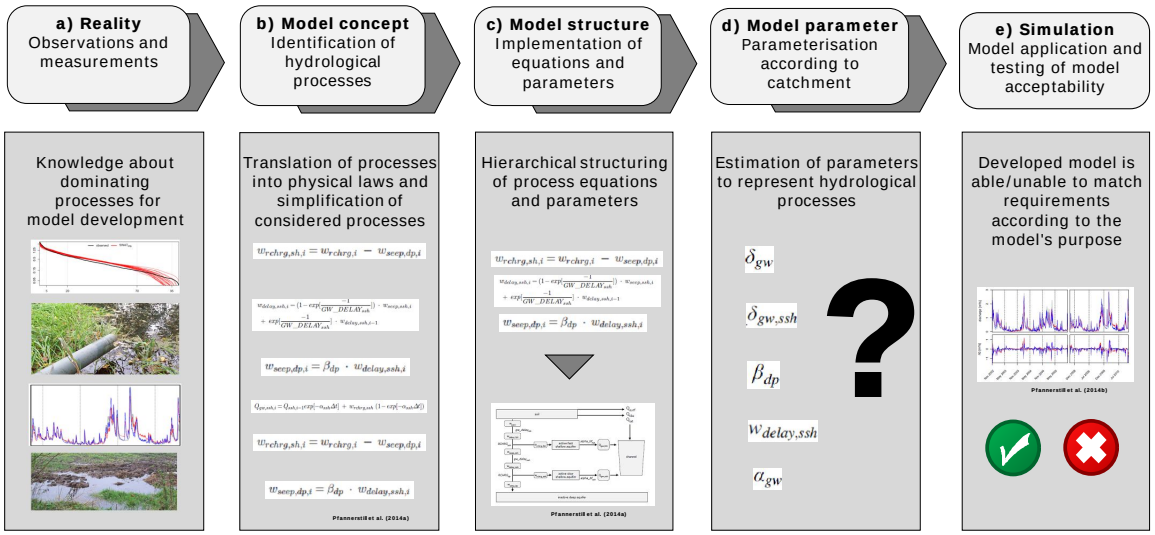

Figure 1: Steps for the development of hydrological models, derived from Gupta et al. (2008) and Reusser et al. (2009).

of the model's behaviour. The match between simulated and oberserved hydrological data can be distinguished with performance metrics and provides basic diagnostic information as the first core result of diagnostic model analyses. According to Gupta et al. (1998) and Gupta et al. (2008), diagnostic model evaluations with multiple performance metrics allow the determination of the models ability to represent dominant hydrological processes and typical patterns. Knowledge of the model's ability to simulate hydrological processes is the second core result that is obtained from diagnostic model analyses. This identified ability of the model is then related to the model structure and its parameters (Yilmaz et al., 2008). In this context the "principle of hydrologic consistency" was introduced by Martinez and Gupta (2011) to promote the necessity of linking model performance with model parameters to represent hydrological processes that are under investigation. Beside of linking model performance with model parameters to test for hydrological consistency, this initial concept was further developed (e.g. Euser et al., 2013; Hrachowitz et al., 2014). In these studies, expert-knowledge and hydrological characteristics of the catchment were utilised to decide about realism of the applied model structures. However, all studies have in common that they aim to analyse the reproduction of discharge and discharge patterns together with hydrological process reproduction.

The detailed analysis about the representation of characteristic discharge events and discharge patterns leads to comprehensive knowledge of the model behaviour. For this, the amount of information made available from hydrological output data has to be increased (Wagener et al., 2003; Merz et al., 2011). It is investigated if all relevant discharge phases of the hydrograph and the hydrological processes can be reproduced with the existing hydrological model (e.g. Gupta et al., 1998; Yilmaz et al., 2008; Clark et al., 2011; Pechlivanidis et al., 2014).

1.1.2 Model evaluation

The classical model evaluation aims to quantify the agreement between simulated and observed data for calibration and validation purposes. Several methods for the calibration rely on this agreement to identify parameter sets for satisfying discharge reproduction (e.g. Beven and Binley, 1992; Duan et al., 1993; Vrugt et al., 2003a,b; Vrugt and Ter Braak, 2011). Comparing the available datasets leads to information how the model performs for the purpose of model application (Bennett et al., 2013). There are several evaluation criteria that are commonly used to provide a quantitative assessment of the agreement between simulated and observed hydrological characteristics (e.g. Madsen, 2000; Krause et al., 2005; Dawson et al., 2007). As shown for the discharge evaluation with the Nash-Sutcliffe efficiency (NSE, Nash and Sutcliffe, 1970), there is a strong focus on peak flows by neglecting possible underestimations during low flow (Krause et al., 2005). The same advantages and disadvantages were found for the coefficient of determination (r^2) due to high sensitivity to high extreme values (Legates and McCabe, 1999) such as peak flow and a reduced sensitivity to systematic model errors during low flow (Krause et al., 2005). As summarised by Moriasi et al. (2007b), additional commonly used evaluation criteria are the mean absolute error (MAE), the mean square error (MSE), and the root mean square error (RMSE). These evaluation criteria strongly react to all peak errors (Reusser et al., 2009). To reduce this strong focus on peak flow, there is the possibility to make use of logarithmic forms of the NSE for example (Krause et al., 2005). Another popular evaluation criterion is the percent bias (PBIAS), which has to be applied carefully since positive and negative errors may balance each other out (Bennett et al., 2013).

Based on the summary of commonly used evaluation criteria and according to Krause et al. (2005), there is no common performance metric to account for all discharge phases. In this context, evaluation criteria are the generic term to describe this agreement. More specifically, performance metrics use a single overall statistic to aggregate model errors over a large range of hydrologic behaviours (Boyle et al., 2000). However, this aggregration leads to implications since the overall performance metric hides information about how different model components perform (Wagener et al., 2003). Using only one single overall and aggregated performance metric prevents the analysis of single discharge events. Consequently, the decision about model performance is limited since the performance of a model may differ in extreme conditions like during floods or low flow as opposed to average conditions (Orth et al., 2015).

Dawson et al. (2007) recommend that hydrological modellers should report performance for several criteria to provide information about the fitting between simulated and oberserved data. Going one step further, a combination of different evaluation criteria with multi-objective evaluation (e.g. for high and low flow) seems to be more beneficial (Krause et al., 2005). However, the assessment of multiple objectives is challenging due to trade-offs between the different objectives since it is often impossible to optimise all objectives simultaneously as shown by Madsen (2000) and by Zhang et al. (2011b).

The aim of optimising multiple objectives is directly linked to the simulation of hydrological

processes of the model. Diagnostic model analyses make use of performance metrics to analyse if the degree of realistic representation of the real world has been achieved sufficiently (Gupta et al., 2008). Multiple performance metrics are combined for the diagnostic model evaluation (Gupta et al., 1998; Boyle et al., 2000; Krause et al., 2005) to capture different phases of the hydrograph during the evaluation process. Considering model failures in different phases of the hydrograph, Reusser et al. (2009) and Reusser et al. (2011) used TIme series of GRouped ERrors (TIGER) to extract information about temporal patterns of dominant error types (e.g. peak simulation) for diagnostic model analyses. Beside of statistical error metrics it is also benefitial to include discharge-based metrics that capture hydrological functions (van Werkhoven et al., 2009).

In this context, signatures are defined as discharge response characteristics that provide insight into the hydrologic function of catchments (Yilmaz et al., 2008; Sawicz et al., 2011). For this, Yilmaz et al. (2008) developed a diagnostic approach to detect signature patterns in the watershed for which the functionality can be detected and somehow quantified from the data (e.g. discharge). According to this approach, four primary functions of a watershed system incorporating overall water balance, vertically redistribution of excess rainfall between the faster and slower runoff components, redistribution of runoff in time, and redistribution of moisture were defined to finally relate these to their incorporation into the model (Yilmaz et al., 2008). As shown by Shafii and Tolson (2015), these signatures can be used as evaluation criteria in multi-objective model calibrations. Similarly Boyle et al. (2000), Bekele and Nicklow (2007), and Zhang et al. (2011b) used several objectives to improve the calibration of models but without explicitly considering signatures. The advantage of applying signatures during model calibration is the detection of inadequacies of hydrological process simulation for the catchment despite of good performance metric values (e.g. Hrachowitz et al., 2014).

The flow duration curve (FDC) provides a cumulative distribution that depicts the relationship between the discharge magnitude and its frequency (Vogel and Fennessey, 1994; Smakhtin, 2001). Pokhrel et al. (2012) demonstrated that signatures which are derived from the FDC can be used to select parameters for a more realistic representation of the hydrologic behaviour of the watershed. The approaches of Yilmaz et al. (2008) and Pokhrel et al. (2012) highlight the applicability of signatures as additional evaluation criteria since specific hydrological characteristics of the catchment can be taken into account (Yilmaz et al., 2008; Sawicz et al., 2011; Pokhrel et al., 2012).

However, the combination of several performance metrics and signatures leads to a trade-off for the selection of the best model run since improved performance of one criteria may lead to less satisfying performance of another criteria (Gupta et al., 1998; Madsen, 2000). Regarding those conflicts for the achievement of good performance for different criteria, this can be interpreted as a sign of model deficiency (Kollat et al., 2012). This potential model deficiency needs further analyses aiming to detect why the applied model structures are not capable to achieve satisfying model performance and to resolve explanations for the model behaviour.

1.1.3 Temporal parameter sensitivity

Classical sensitivity analyses are helpful to identify parameters which have the highest impact on the model's output during the whole time series (e.g. Saltelli et al., 2000; van Griensven et al., 2006; Nossent et al., 2011; Nossent and Bauwens, 2012). However, the diagnostic information content of classical sensitivity analyses is limited as the provided information is aggregated to the whole time series preventing to detect hydrological patterns and parameter dynamics within the investigated time series. To overcome this limitation, temporal parameter sensitivity analyses are applied. TEmporal Dynamics of PArameter Sensitivity (Sieber and Uhlenbrook, 2005; Reusser et al., 2011, TEDPAS) that are resolved in a high temporal resolution provide further insights into the behaviour of the hydrological model. The sensitivity can be used as diagnostic information by assuming that temporal patterns can be interpreted to the occurence of simulated hydrological processes.
As shown in several studies, temporal parameter sensitivity analyses were found to be a valuable method to diagnose the relevance of parameters for the simulated hydrological output (e.g. Garambois et al., 2013; Herman et al., 2013; Guse et al., 2014). For efficient model calibration, temporal parameter sensitivities are used to reduce the amount of parameters that are needed for calibration (van Werkhoven et al., 2009). Reusser et al. (2011), Massmann and Holzmann (2012), and Massmann et al. (2014) considered the relation between parameters and the model performance. Regarding calibration strategies, the temporal parameter sensitivities provide knowledge about how different parameters can be used to improve specific phases of the hydrograph. With respect to the reproduction of specific hydrological processes, the results of temporal parameter sensitivity analyses can be used to constrain parameter ranges according to the hydrological processes within the study catchment (Pokhrel et al., 2012; Hrachowitz et al., 2014). In this way, process-oriented and constrained parameter calibration prevents calibration related model deficiencies.
For the case that process-oriented and constrained parameter calibration does not lead to satisfying model performance, there is the need to identify model structure deficiencies that are caused by inappropriate process equation and parameter implementation. According to Yilmaz et al. (2008), the selected evaluation criteria must be capable to identify model components that cause model structure related performance problems. As shown by Reusser and Zehe (2011), Herman et al. (2013), and Guse et al. (2014), temporal parameter sensitivities can be linked to performance metrics to identify the deficient model component that is responsible for poor performance in specific discharge phases. In this way, insights into the model structure are provided since the dynamics of temporal parameter sensitivity for a model component can be isolated and related to the hydrological processes within the catchment pointing to the aspects of the model component that needs improvement (Gupta et al., 2008).

1.1.4 Modification of model structures for improved performance

In the case of model failures, diagnostic model analyses provide information about failing model components (e.g. Gupta et al., 1998; Wagener et al., 2003; Yilmaz et al., 2008; Pechli-

vanidis et al., 2014). The aim of model modifications is to improve the model's hydrological process simulation (Gupta et al., 2008). As shown in studies of Fenicia et al. (2008) and Reusser et al. (2009) models can be assessed to improve the understanding of the hydrological processes in the study catchment. This knowledge provides understanding of the relation between the model structure and the catchment (Guse et al., 2014) and should be used to improve the model performance by integrating expert-knowledge about processes and results of model structure diagnostics (Hrachowitz et al., 2013).
For model development, the processes within the catchment have to be transferred into model equations and model parameters. It has to be ensured that the new model structure suits to the catchment (Euser et al., 2013; Hrachowitz et al., 2014), and that the main processes are captured while maintaining minimum levels of complexity (Fenicia et al., 2008; Zhang et al., 2011a). Regarding the problem of needed degree of complexity, an increased complexity may lead to over-parameterisation but a too simple model structure may suffer from an incomplete representation of relevant processes (Orth et al., 2015).
Examples of model structure improvements to better capture relevant processes of the real world can numerously be found. Perrin et al. (2003) present the modification of a simple four-parameter model with improved model performance for simulated low flows. A conceptual hydrological model of low complexity was modified by Koren et al. (2014) by integrating concepts of another model to better capture the effects of freezing and thawing soil on the runoff generation process in a physically-based way. The simulation of effective rainfall routing using multiple tanks instead of a linear reservoir was suggested by McMillan et al. (2011) to improve the performance in the low flow tail. A systematical stepwise update of the model structure with progressive incorporation of new hypotheses about the catchment behaviour was proposed by Fenicia et al. (2008). They found out that the improved model performane benefits from improved simulation of vertical storage distribution and storage variations in the horizontal dimension.
All summarised examples of model structure modification and adaptation have in common that the main aim is to improve the model performance in reproducing discharge or the representation of the discharge components. However, the decision if the model modification leads to a better representation of the hydrological processes is an important task after model modifications. The final decision is again supported by diagnostic model analyses, comprising a good model performance and a verification of the model's ability to simulate the relevant hydrological processes compared to modelers' expectations and catchment knowledge.
The confirmation of improved model structures is shown in a first step by good model performance, represented by good match between observed and simulated data. As mentioned in chapter 1.1.2, appropriate evaluation criteria have to be selected to determine the degree of improvement for the aspects of the hydrological system that were under investigation. However, the application of evaluation criteria does not guarantee that the model simulates all hydrological processes in a reasonable way. A model may be inadequate despite of good performance due to insufficient reproduction of hydrological signatures (Hrachowitz et al., 2014). Consequently, further diagnostic model analyses are needed to analyse if the modified model leads to realistic simulations of the whole hydrological system and if the process dy-

namics are captured by the improved model structure.
TEDPAS provides temporal parameter sensitivities for individual model components, which can be used to analyse the relevance of modified model components with respect to other model components. Since individual parameter sensitivities of model components can be intepreted to a simulated occurrence of hydrological processes, it is diagnosed if all simulated hydrological processes show the expected temporal parameter sensitivity. For this, expert-knowledge about the observed hydrological processes and measurements within the catchment are utilised to verify if the modified model component was adequately improved. In this context, verification is defined as the confirmation of adequate temporal parameter sensitivity and adequate process simulation with observed processes of the catchment. The observations-based expectations are formulated with hypotheses that are in this context not a formal statistical test but a qualitative utilisation of expert knowledge (Clark et al., 2011). As a consequence, knowledge about the catchment is a prerequisite. Euser et al. (2013) and Hrachowitz et al. (2014) exemplarily demonstrated frameworks to diagnose the hydrological consistency of applied models. These examples highlight the necessity of methods to diagnose if the improved model performance is related to the modifications of the model component and if all hydrological processes are simulated reasonably.
Interpreting this demand in a more general way, diagnostic model analyses can be incorporated into the modification and development of hydrological models to ensure hydrological consistency. However, up to now there is no established way of how these diagnostic model analyses have to be integrated in modification and development procedures. This limitation will be taken up in the following to firstly summarise preliminary work with model diagnostics for a hydrological model. Finally, these results are used to motivate the research questions of this thesis.

1.2 Preliminary work

The derived necessity of diagnostic methods in the field of hydrological modeling is the initial point to motivate this thesis. The introduction provided an overview of the state of the art in diagnostic model analyses to detect and improve model structure deficiencies. With respect to this overview, an example of a diagnostic model analysis is used to motivate the research questions of this thesis. In this context, the preliminary work can be understood as a summary of previous studies. Firstly the hydrological model that was under investigation is briefly described and previous studies about the poor model performance of low flow discharge phases is summarised. Based on the studies that revealed poor model performance for the low flow discharge, the demand to overcome structural deficiencies is derived.

1.2.1 The Soil and Water Assessment Tool

The Soil and Water Assessment Tool (SWAT, Arnold et al., 1998) is a semi-distributed eco-hydrological model that has been applied to several questions about integrated watershed-modeling (Arnold and Fohrer, 2005; Krysanova and Arnold, 2008; Kiesel et al., 2010; Strauch et al., 2013). The impacts of land management on the landscape can be assessed on a daily

time step by considering the water cycle together with nutrient and pesticide fluxes (Neitsch et al., 2011).

To assess these impacts, SWAT incorporates a spatial representation of catchment information. The subbasins are spatially defined for the watershed but the hydrological resoponse units (HRU) within these subbasins are lumped to reduce computational efforts. According to several studies about model calibration and paramter identification (e.g. Lenhart et al., 2002; White and Chaubey, 2005; van Griensven et al., 2006; Cibin et al., 2010), the SWAT model is very complex since a high number of parameters is integrated into the different model components.

The SWAT model and the similar SWIM model (Krysanova et al., 1998) were subject to a number of improvements to enhance the simulation of several aspects comprising plant growth or nutrient balance (e.g. Hesse et al., 2012; Strauch and Volk, 2013; Hesse et al., 2013). With respect to water balance simulation, Moriasi et al. (2007a) and Moriasi et al. (2012) highlight the necessity of realistic reproduction of tile drainage systems for runoff generation, which was incorporated into SWAT as a tool to design cost-effective and environment-friendly tile drain water management systems. A model integration framework was developed by Guzman et al. (2015) to improve modelling of surface and groundwater interactions. Adaptations to simulate different reservoir management options with SWIM were shown by Koch et al. (2013a). To reproduce correctly the runoff generation in a low mountain region, Eckhardt et al. (2002) found a strong underestimation of interflow due to failed partition between interflow and groundwater recharge. By adapting the soil water system to the hydrological functions of the study catchment, a more reasonable representation of the runoff components was achieved.

Regarding the adequate simulation of runoff components, Luo et al. (2012) modified the model structure of SWAT aiming to improve the baseflow simulation with two river contributing groundwater storages. The adequate simulation of low flow phases was also subject to other studies (Wu and Johnston, 2007; Zhang et al., 2011b; Conradt et al., 2012; Koch et al., 2013b), which evaluated the model performance mainly by common performance metrics such as NSE and percent bias (PBIAS, Moriasi et al., 2007b). All studies about the baseflow simulation have in common, that they assume a model structure deficit for the groundwater component of SWAT due to poor model performance in discharge phases. Consequently, the groundwater model component of SWAT needs a detailed analysis to identify the reasons for poor model performance.

1.2.2 Structural deficiency of the SWAT groundwater component

The given examples of poor model performance for the SWAT model in low flow phases of Wu and Johnston (2007), Zhang et al. (2011b), Luo et al. (2012), and Koch et al. (2013b) were related to lowland catchments with specific hydrological characteristics as represented by groundwater dominated discharge. The consideration of these specific hydrological characteristics is highly relevant for successful model application as it has to be ensured that the model structure suits to the catchment (Euser et al., 2013; Hrachowitz et al., 2014). Guse

et al. (2014) identified poor model performance for several performance metrics (Tab. 1) due to groundwater model component failure for the Treene catchment, which is a typical lowland catchment. The multi-objective evaluation revealed a poor model performance especially in low flow phases. As demonstrated in Table 1, this poor performance becomes apparent due to the use of several performance metrics which are related to different parts of the hydrograph. A detailed description of the different performance metrics and their relevance with respect to different discharge phases is summarised in Guse et al. (2014).

Table 1: Qualitative summary of model performance for the SWAT model in the Treene catchment, differentiated by discharge phase after Guse et al. (2014). The model performance is depicted qualitatively by the classes of satisfying (+), moderate (0) and poor (-) for the measures of peak difference (PDIFF), root mean square error (RMSE), mean relative error (MRE), Nash-Sutcliffe efficiency (NSE), longest common sequence (LCS), and scaled mean square error (SMSE) (cf. Dawson et al., 2007; Reusser et al., 2009; Guse et al., 2014).

	PDIFF	RMSE	MRE	NSE	LCS	SMSE
peaks	-	-	0	+	+	-
recession	0	+	0	+	+	+
low flow	-	-	-	-	-	+

A diagnostic interpretation of this poor model performance was done in Guse et al. (2014) by combining TEDPAS with model evaluation results. The diagnostic information about model deficiencies were related to the groundwater model component and its incoporated parameters. They concluded that groundwater processes of the studied lowland catchment are not sufficiently represented in the model structure of SWAT and that the groundwater model component has the highest potential for the improvement of model performance in groundwater dominated discharge phases.

1.3 Research questions and objectives

Considering the low flow simulation of the SWAT model in lowlands, the structural deficiency of the groundwater model component was found the be the main reason for poor model performance. Based on this conclusion, this thesis aims to show in a first step how the diagnostic information of model failure can be used to derive a new concept for the groundwater model structure of SWAT. By using expert-knowledge, the focus of a model modification is the derivation of a new concept for the representation of groundwater processes. For this, theoretical concepts of groundwater processes, groundwater concepts of other hydrological models, and groundwater process knowledge that is based on field studies were used to derive possible model structure modifications for SWAT. By selecting the intensively studied Kielstau subcatchment of the Treene lowland catchment, it is ensured that characteristic hydrological processes of lowlands and knowledge about these processes are considered for the model modification. Since Guse et al. (2014) provided diagnostic information about the groundwater component failure for the Treene catchment, the first research question focuses on the utilisation of this information (Chapter 2):

- **How can deficient model structures be improved with information of diagnostic model analyses?**

This research question is answered exemplarily for the modification of the SWAT model groundwater component which aims to improve the groundwater dominated low flow phases. The decision if modifications of the model lead to the desired improvement, appropriate evaluation criteria for the investigated hydrological processes have to be considered. Several performance metrics are available to evaluate model results with respect to different discharge phases. Signatures are especially helpful to assess the hydrological functions of the catchment but also typical characteristics of the hydrograph. Yilmaz et al. (2008) and Pokhrel et al. (2012) emphasise that FDC segments are recommendable to analyse the simulation of defined discharge magnitudes. To evaluate low flow discharge magnitudes, it has to be ensured that the different segments of the FDC capture the discharge phases which are under investigation. However, up to now there is no established FDC segment to assess extreme low flow conditions. Especially in lowlands, these conditions are controlled mainly by groundwater. Consequently an additional segment in the low flow segment would support the identification of the model's inability or ability to simulate accurate low flow discharge. In this thesis, a specific low flow segment of the FDC is derived to confirm improvements of the SWAT groundwater component. Based on these requirements, the second question that needs to be resolved is formulated (Chapter 3):

- **How can modified model structures be evaluated by considering all relevant discharge phases?**

The evaluation of results upon model modifications with performance metrics provides a first impression whether the modifcations have lead to improved model performance. However,

the evaluation with performance metrics does not consider if all hydrological processes are accurately reproduced. The investigation of certain aspects of the hydrological system requires expensive measuring techniques that are often not affordable. This is in particular valid for the catchment scale. At this scale, it is generally impossible to obtain representative data of all relevant hydrological processes. The consideration of expert knowledge helps to overcome this limitation. It supports the decision if the improvement of the model structure leads to improved suitability for the study catchment. Qualitative hypotheses and rules of thumb about hydrological processes within the catchment have to be used for the comparison with modeled processes. In this way, a proper representation by the model structure is verified. Based on this assumption, the third research question is raised (Chapter 4):

- **How can modified model structures be verified with diagnostic model analyses and expert knowledge?**

Finally, the formulated research questions are intepreted in a more general context. The steps for improving the SWAT groundwater component are used to derive a general procedure for hydrologically consistent model improvements. The core idea of this procedure is to make available a guideline that incorporates the steps of model deficiency detection, model improvement and following model evaluation and verification. In this thesis, the different steps are exemplarily described for the improvement of the SWAT model groundwater component. However, these steps can be applied to any other hydrological model so that the main research question of this thesis is answered by transferring the individual results into a more general context (Chapter 5):

- **How can hydrologically consistent models be achieved?**

In this way, a general guideline to achieve hydrologically consistent models is the synthesis of this thesis. The steps to derive this guideline are described within this thesis according to Figure 2 by providing an overview about the content of the individual chapters.

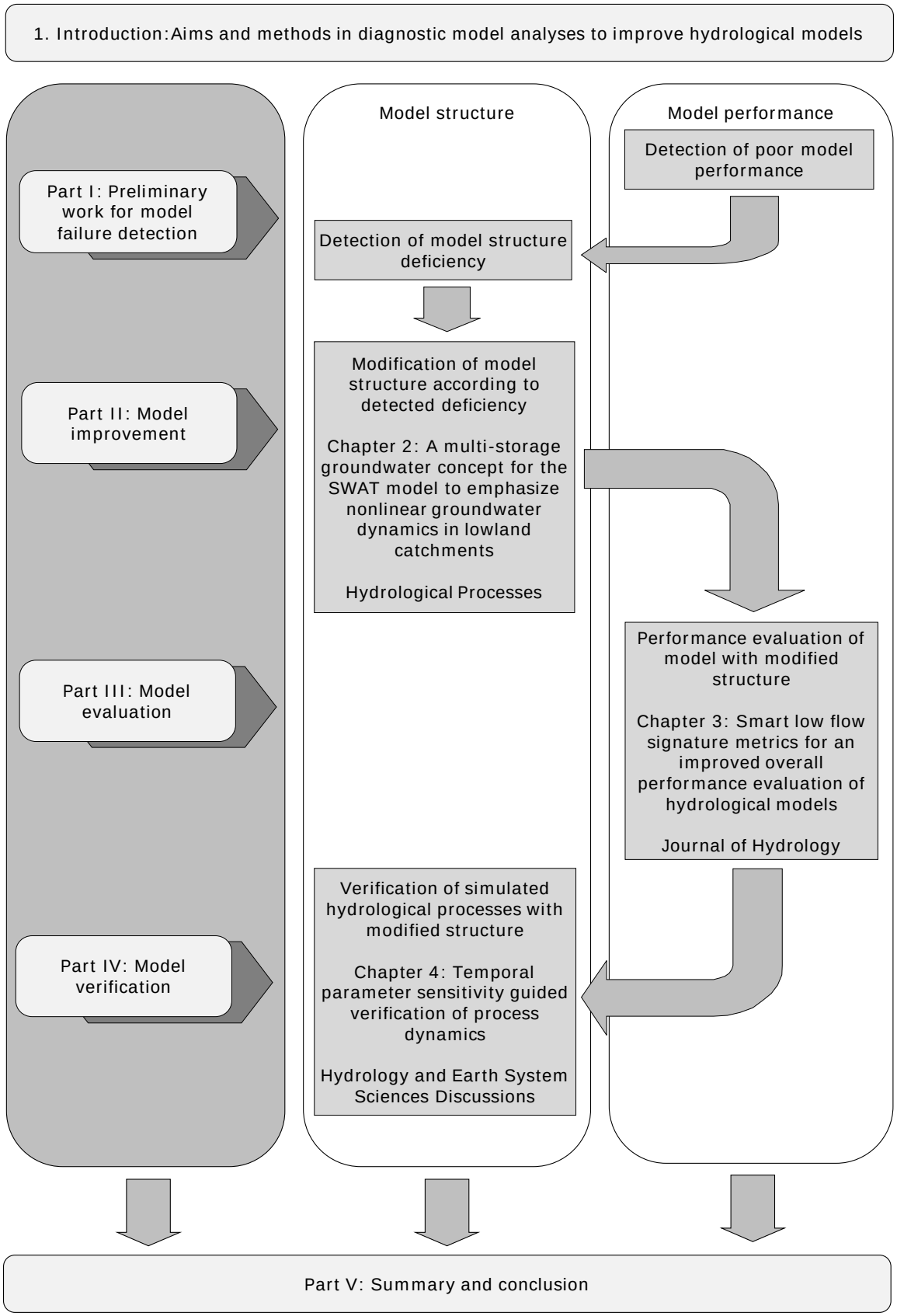

Figure 2: Structure of this thesis with summarised content of the individual chapters.

2 A multi-storage groundwater concept for the SWAT model to emphasize nonlinear groundwater dynamics in lowland catchments

Pfannerstill, M., Guse, B., and Fohrer, N.: A multi-storage groundwater concept for the SWAT model to emphasize nonlinear groundwater dynamics in lowland catchments, Hydrol. Process.,28, 5599-5621, doi:10.1002/hyp.10062, 2014.

Received: 21 May 2013 - Accepted: 5 September 2013

Abstract

Hydrological models are useful tools to analyze present and future conditions of water quantity and quality. The integrated modeling of water and nutrients needs an adequate representation of discharge components. In common with many lowlands, the groundwater contribution to the discharge in the North German lowlands is a key factor for a reasonable representation of the water balance especially in low flow periods. Several studies revealed that the SWAT model with its world-wide application may cause poor model performance for low flow periods. This paper deals with the extension of the groundwater module of the SWAT model to enhance low flow representation. The current two-storage concept of SWAT was further developed to a three-storage-concept. Therefore the groundwater module was modified by splitting the active groundwater storage, which contributes to the discharge, into a fast and a slow contributing aquifer. The results of this study show that the groundwater module with three storages leads to good prediction of the overall discharge especially for the recession limbs and the low flow periods. The improved performance is reflected in the signature measures for the mid segment (PBIAS: $-3{,}6\,\%$ vs. $14{,}0\,\%$) and the low segment (PBIAS: $-3{,}6\,\%$ vs. $-80{,}5\,\%$) of the flow duration curve. In comparison to the original groundwater module, the three-storage groundwater module is more process-based than the original version owing to the introduction of a fast and a slow groundwater flow component. As two of the three groundwater storages can be activated or deactivated for each subbasin, spatial heterogeneity of the landscape is taken into account. As a result the proposed three storage groundwater concept can be transferred to other catchments.

2.1 Introduction

Hydrological models are used to address questions of water quality and water quantity in practice and science (Borah and Bera, 2003). One important purpose is the modeling of water and nutrient fluxes to investigate agricultural influences on discharge and nutrient dynamics (e.g., David *et al.*, 2009; Schilling and Wolter, 2009). Furthermore, hydrological models allow to evaluate scenarios of climate and land use change with their effects on the hydrology of catchments (Huang *et al.*, 2013; Volk *et al.*, 2009; Bronstert *et al.*, 2007; Niehoff *et al.*, 2002). With this wide application of hydrological models, the requirements for an integrated modeling of water and nutrients are increasing steadily.

The environment acts as a system with strong nonlinearity and feedbacks. Due to this complexity, rainfall-runoff models try to simplify and approximate processes (De Vos *et al.*, 2010). Depending on the focus of the model and specific characteristics of the catchment, certain processes get emphasized or neglected. Referring to the model structure, a complex process description can help to improve the model performance. However, detailed descriptions of all processes are impossible due to raising model complexity and the lack of appropriate data. As a consequence, there is a trade-off between complexity in process representation and uncertainty introduced while parameterizing all these processes.

As is typical for lowlands, groundwater interaction is an important characteristic in many catchments of the German lowlands. Shallow groundwater tables result in a groundwater flow, which is often the major contributor to streamflow (Gonzales *et al.*, 2009; Schmalz *et al.*, 2008a; Hattermann *et al.*, 2004; Wittenberg, 2003). This contribution of the groundwater storage leads to baseflow within the stream (Eckhardt, 2008) with a high fraction in dry seasons. As a consequence, knowledge about baseflow is required for the assessment of water quality and low flow conditions (Eckhardt, 2008).

Describing complex interactions between groundwater storages, fluxes and baseflow is challenging. Hydrological models often simply combine an estimated groundwater storage volume with a linear regression coefficient to determine groundwater flow (Nathan and McMahon, 1990). However, such a linear description of groundwater flow ignores the nonlinear behaviour of groundwater dynamics. A reason for nonlinearity might be that the groundwater storage includes delayed storages which are closely linked to heterogeneity of the catchments (Samuel *et al.*, 2012). Nathan and McMahon (1990) revealed that baseflow is rather a conceptual convenience than a precise description of the natural processes. These processes may vary seasonally and could result in different recession constants. Wittenberg (2003) and Nathan and McMahon (1990) stated that the value of the reservoir or recession constant is not constant, but increases generally with falling flow until the highest recession constant characterizes the recession characteristics of the catchment.

Kirchner (2009) assumed that many of the processes and rate coefficients which control water flow in the subsurface are nonlinear and strongly dependent on storage. Analyses of streamflow recessions showed that a simple linear reservoir model does not represent the recession curve adequately (Fenicia *et al.*, 2006). To account for the nonlinearity, baseflow could be interpreted as the outflow of two or more parallel linear and nonlinear reservoirs with different response times (Staudinger *et al.*, 2011; Brandes *et al.*, 2005; Wittenberg, 1999). Especially for lowlands, the nonlinear description of groundwater storages is able to improve baseflow and recession periods. This was also shown for a lowland in Northern Germany (Wittenberg, 1999).

Models for a detailed description of the groundwater processes are widely used to analyze the interaction between groundwater and river (Munz *et al.*, 2011; Krause and Bronstert, 2007; Saenger *et al.*, 2005). However, these approaches need detailed and time-consuming field measurements for the process calibration. In the field of integrated catchment modeling for the evaluation of climate and land-use change, conceptual models which are capable to represent all hydrologic phases in combination with agricultural practice seem to be more ad-

equate. The eco-hydrological Soil and Water Assessment Tool (SWAT, Arnold *et al.*, 1998) has a wide range of application (Strauch *et al.*, 2013; Krysanova and Arnold, 2008; Arnold and Fohrer, 2005), and was also used in several studies of the North German lowlands (Guse *et al.*, 2013; Koch *et al.*, 2013; Kiesel *et al.*, 2010; Schmalz *et al.*, 2008b).
There are several investigations with the SWAT model revealing limitations by modeling dry seasons with groundwater dominated discharge. Wu and Johnston (2007) found underestimated baseflow for dry years in a Great Lake watershed and calibrated a SWAT model especially for dry seasons. In the study of Koch *et al.* (2013), baseflow was constantly overestimated during the validation period, although it was slightly underestimated during calibration. Additionally, Watson *et al.* (2003) showed problems of baseflow recession rates, as the recession limb of the predicted discharge was much higher than the observed. Eckhardt (2008) concludes that it seems to be a shortcoming of the model, that dry seasons cannot be reproduced satisfactorily with other discharge events (high flow, mid flow).
In an analysis of the temporal dynamic of parameter sensitivity, Guse *et al.* (2013) detected that groundwater and evaporation parameters are the main reason for poor performance of low flow reproduction by SWAT for a lowland catchment. They propose the modification of the SWAT groundwater module to enhance the process-oriented representation of low-flow periods. One possible modification is the activation of groundwater contribution by the deep aquifer to the channel as suggested by Luo *et al.* (2012).
In our investigations we focussed on the improvement of groundwater process representation in the SWAT model. Based on the aforementioned assumptions and suggestions of Staudinger *et al.* (2011), Kirchner (2009), Fenicia *et al.* (2006) and Wittenberg (1999), we emphasized the nonlinearity of the groundwater module. The main aims are: (i) to implement a more complex groundwater storage approach in the SWAT model and (ii) to test and validate if the modifications are suitable to estimate baseflow, especially during recession and low flow conditions.

2.2 Materials and methods

2.2.1 Study area

The study area of our investigations is the Kielstau catchment, which is located within a lowland area of the federal state Schleswig-Holstein in Northern Germany. The catchment area is about 50 km^2 and is a subbasin of the Treene River. The mean annual precipitation and temperature are 918,9 mm and 8,2° (Station: Gluecksburg-Meierwik, period: 1961 - 1990; DWD, 2012). The discharge at the gauging station Soltfeld is measured since 1985 by LKN (2012). For our investigations, we used the hydrological years from 1997 to 2010.
The landscape was mainly influenced by glacial and periglacial processes of the late Pleistocene (Lundqvist, 1986). The valleys, rolling hills and depressions were formed by subglacial melt water and shearing forces of the ice shields (Riedel and Polenky, 1987; Wahnschaffe and Schucht, 1921). The topography ranges between 27 m and 78 m above mean sea level. The predominant soils in higher regions of the catchment are Haplic Luvisols and Stagnic Luvisols. Along the stream and its tributaries, Sapric Histosols are dominant (WRB; BGR, 1999).

One important hydrological characteristic of the Kielstau catchment is the high fraction of drained agricultural area in the catchment, which is estimated to be 38 % (Fohrer *et al.*, 2007). Schmalz *et al.* (2008a) described the dynamics of the near-surface groundwater at a riparian wetland as a dynamic interaction between groundwater and surface water. The near-surface groundwater is generally controlled by precipitation and, close to the river, also by river water level (Schmalz *et al.*, 2008a). Further information about the catchments and results of investigations can be found in Schmalz and Fohrer (2010) and Fohrer and Schmalz (2012).

2.2.2 The SWAT model

The SWAT model is a semi-distributed, eco-hydrological model (Arnold *et al.*, 1998) which simulates not only the water cycle but also nutrient and pesticide fluxes, soil erosion, plant growth cycles and management practices on a daily time step (Neitsch *et al.*, 2011). Subbasins have a distinct spatial position, while the hydrological response units (HRU) within the subbasins are lumped. SWAT is a conceptual model with empirical and process-oriented components. These different components result in a very complex model with a high number of parameters (Cibin *et al.*, 2010). The modeling processes of SWAT are divided into a land and water phase (Neitsch *et al.*, 2011). The first step is the calculation of the water cycle at the land phase. The water balance is calculated by changes in soil water storages for each day, based on the calculation of the relevant hydrological processes. The precipitation is the main input to the water balance. Evapotranspiration, runoff, soil water percolation and groundwater flow are the most important processes, which influence the water balance equation in each subbasin. After calculating the water balance, the subbasins are connected in the water phase and the water is routed through the subbasins along the river stream. Further details about the SWAT model concept can be found in Neitsch *et al.* (2011).

2.2.2.1 The groundwater module of the SWAT model

The detailed process equations of the groundwater processes are summarized in Neitsch *et al.* (2011). In the following description the main groundwater processes of the original SWAT version are shown, which control the groundwater contribution to the discharge of the channel. In the SWAT model, soil water percolates into groundwater, which is divided into a shallow and a deep aquifer. The shallow aquifer represents an unconfined aquifer that may discharge into the channel. On the other side, the deep aquifer is described as a confined aquifer. As a consequence, the deep aquifer does not contribute to the streamflow within the watershed. Thus, the deep aquifer is considered as inactive, because water is lost from the system within the modeled catchment.

An exponential delay function accounts for the delay between the time water exits the soil profile and enters the aquifers. The recharge is calculated with an integrated delay time (Eq. 1, cf. Neitsch *et al.* 2011):

$$w_{rchrg,i} = (1 - exp[\frac{-1}{\delta_{gw}}]) \cdot w_{seep} + exp[\frac{-1}{\delta_{gw}}] \cdot w_{rchrg,i-1} \quad (1)$$

where $w_{rchrg,i}$ (mm H_2O) is the amount of water entering the aquifer on day i. To represent the delay time of recharge due to geologic formations, the parameter δ_{gw} (days) is used. w_{seep} (mm H_2O) is the percolation of soil water out of the last soil layer on day i. The parameter $w_{rchrg,i-1}$ (mm H_2O) is the amount of water, entering the aquifer on the day before $(i-1)$. A fraction of the daily recharge can be routed to the deep aquifer. By using a partition coefficient, the daily recharge $w_{rchrg,i}$ is diverted to calculate the recharge of the deep aquifer (Eq. 2, cf. Neitsch *et al.* 2011):

$$w_{seep,dp,i} = \beta_{dp} \cdot w_{rchrg,i} \tag{2}$$

where β_{dp} (-) is the aquifer percolation coefficient to calculate the recharge of the deep aquifer $w_{seep,dp,i}$ (mm H_2O). The net amount of water recharging the shallow aquifer $w_{rchrg,sh,i}$ (mm H_2O) is then (Eq. 3, cf. Neitsch *et al.* 2011):

$$w_{rchrg,sh,i} = w_{rchrg,i} - w_{seep,dp,i} \tag{3}$$

In the SWAT model, the shallow aquifer contributes actively to the discharge of the channel mainly as baseflow. The groundwater flow $Q_{gw,i}$ (mm H_2O) is the calculated groundwater contribution to the main channel on day i (Eq. 4, cf. Neitsch *et al.* (2011)), where α_{gw} (1/days) is the baseflow recession constant and Δt the time step (1 day):

$$Q_{gw,i} = Q_{gw,i-1} \cdot exp[-\alpha_{gw} \cdot \Delta t]) + w_{rchrg,sh,i} \cdot (1 - exp[-\alpha_{gw} \cdot \Delta t]) \tag{4}$$

Furthermore, the equation incorporates the recharge of the shallow aquifer ($w_{rchrg,sh,i}$) for day i. A schematic description of these conceptual groundwater processes can be found in Fig. 3. The groundwater processes in SWAT can be calculated on HRU level since every groundwater parameter value can be assigned to each HRU independently.

2.2.2.2 Modifications of the groundwater module

The modifications of the groundwater module are motivated by the aforementioned assumptions and suggestions of Staudinger *et al.* (2011), Kirchner (2009), Fenicia *et al.* (2006) and Wittenberg (1999). The two storage concept of the original groundwater module was extended to emphasize the nonlinearity of the baseflow. To separate the groundwater flow of the shallow aquifer into a fast and a slow component, the aquifer was split into two storages (fast shallow aquifer and slow shallow aquifer). We assumed that a part of the groundwater does not contribute to the channel as the catchment of the Kielstau is relatively small with a size of 50 km^2. To realize this process of inactivated groundwater contribution for small catchments, the assumption of the original SWAT model was taken up. A part of the groundwater recharge is routed to the deep aquifer, which is not connected to the channel and considered as inactive for contribution. With this inactive aquifer, it is possible to account for groundwater flow, which leaves the catchment as proposed in Guse *et al.* (2013).
The extended groundwater concept ($SWAT_{3S}$) was implemented into the SWAT model by adding additional process equations and by modifing existing process equations. The sep-

aration of the shallow aquifer of the original SWAT into a fast shallow aquifer and a slow shallow aquifer needs modifications of the recharge for each aquifer. The amount of water ($w_{rchrg,fsh}$) recharging the fast shallow aquifer is calculated with Eq. 5:

$$w_{rchrg,fsh} = w_{rchrg,i} \; - \; w_{seep,sh,i} \tag{5}$$

To emphasize the nonlinearity, the recharge of the second groundwater storage (slow shallow aquifer) was connected with a time delay function. As a consequence, the seepage from the first to the second aquifer $w_{seep,ssh,i}$ will be delayed independently from the delay function of the fast shallow aquifer seepage with Eq. 6:

$$w_{delay,ssh,i} = (1 - exp[\frac{-1}{\delta_{gw,ssh}}]) \; w_{seep,ssh,i} \; + \; exp[\frac{-1}{\delta_{gw,ssh}}] \; w_{delay,ssh,i-1} \tag{6}$$

where $\delta_{gw,ssh}$ is the time delay of recharge due to geologic formations which may differ from the geologic formations between the soil and the fast shallow aquifer. The parameter $w_{delay,ssh,i-1}$ represents the amount of water, percolating from the fast shallow to the slow shallow aquifer on the day before $(i-1)$. The amount of delayed percolation water between the fast shallow and the slow shallow aquifer is then diverted into the recharge of the slow shallow aquifer and the seepage to the deep aquifer, which does not contribute to the channel. The seepage to the deep aquifer $w_{seep,dp,i}$ is calculated by Eq. 7:

$$w_{seep,dp,i} = \beta_{dp} \; \cdot \; w_{delay,ssh,i} \tag{7}$$

The remaining part of the delayed water between fast shallow and slow shallow aquifer $w_{delay,sh,i}$ may then recharge the slow shallow aquifer $w_{rchrg,shs}$, which is calculated with Eq. 8:

$$w_{rchrg,ssh} = w_{delay,ssh,i} \; - \; w_{seep,dp,i} \tag{8}$$

After the distribution of the percolation and seepage water to the different groundwater storages the groundwaterflow is calculated. The equation for the groundwater flow of the fast shallow aquifer (Eq. 9, $Q_{gw,fsh,i}$) is based on the original equation 4. $\alpha_{gw,fsh}$ is the baseflow recession constant and Δt is the time step:

$$Q_{gw,fsh,i} = Q_{gw,fsh,i-1} \; exp[-\alpha_{gw,fsh}\Delta t] \; + \; w_{rchrg,fsh} \; (1 - exp[-\alpha_{gw,fsh}\Delta t]) \tag{9}$$

Equally to the fast shallow aquifer, the slow shallow aquifer may contribute to the discharge as a second component by Eq. 10:

$$Q_{gw,ssh,i} = Q_{ssh,i-1} exp[-\alpha_{ssh}\Delta t] \; + \; w_{rchrg,ssh} \; (1 - exp[-\alpha_{ssh}\Delta t]) \tag{10}$$

where $Q_{gw,ssh,i}$ is the groundwater contribution of the slow shallow aquifer to the main channel on day i. α_{ssh} is the baseflow recession constant for the slow shallow aquifer and Δt is the time step. Furthermore, the equation incorporates the recharge of the slow shallow aquifer

($w_{rchrg,ssh}$) for day i. A schematic description of these conceptual groundwater processes can be found in Fig. 3.

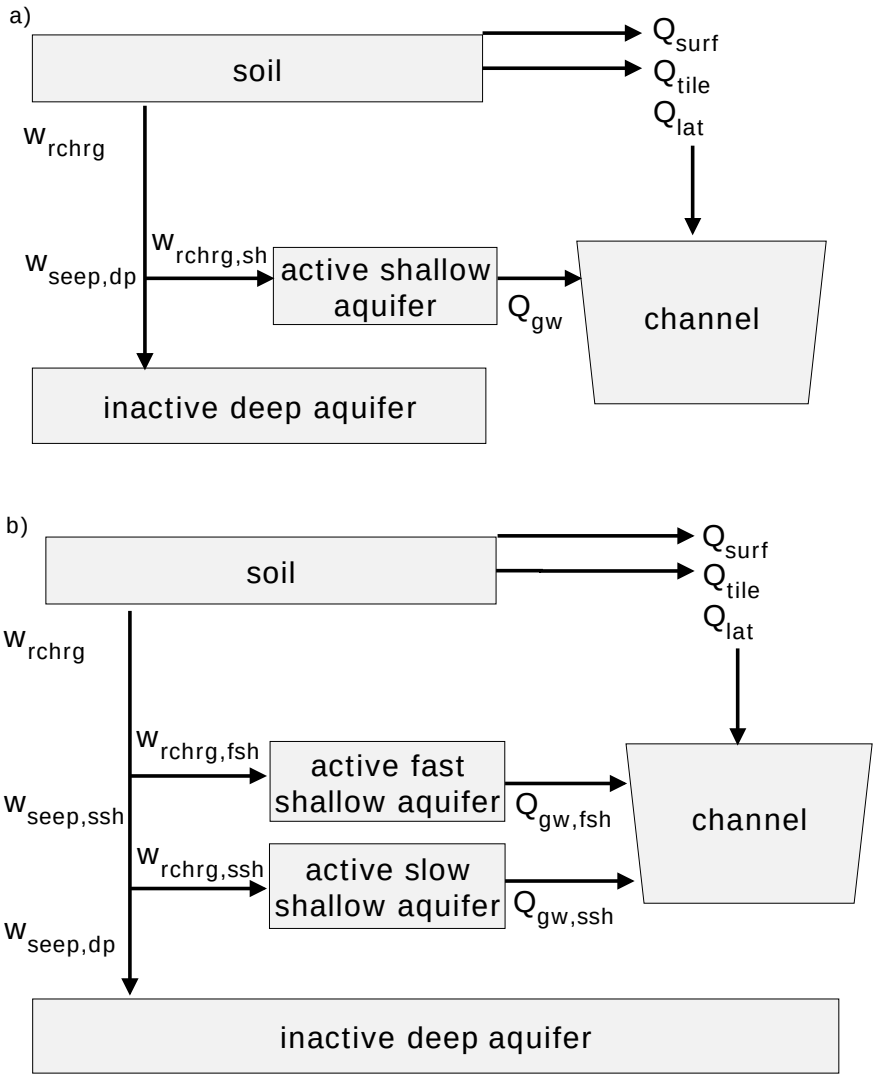

Figure 3: Schematic description of the concepts for groundwater processes in the original two groundwater storage SWAT model (a) and the modified three storage model SWAT$_{3S}$ (b). Q_{surf}, Q_{tile}, Q_{lat} is the surface runoff, the tile drainage flow and the lateral flow into the channel. Q_{gw} denotes the groundwater flow into the channel, which is separated in the modified version (b) into a fast (fsh) and a slow (ssh) flow component. w_{rchrg} describes the amount of water, entering the groundwater. For SWAT$_{orig}$, w_{rchrg} is separated into the amount of water entering the shallow aquifer and the amount of water entering the deep aquifer $w_{seep,dp}$. The new SWAT$_{3S}$ separates the w_{rchrg} into a recharge $w_{rchrg,fsh}$ for the fast shallow aquifer and into the seepage to the slow shallow aquifer ($w_{seep,ssh}$). This seepage is split into the amount of water entering the deep aquifer ($w_{seep,dp}$) and the slow shallow aquifer ($w_{rchrg,ssh}$).

2.2.3 Model setup

We used the ArcSWAT interface (version 2012.10.1.6) to setup the initial SWAT model with its original groundwater module (Rev. 582). Within the watershed delineation, the Kielstau catchment was divided into 36 subbasins with 2214 HRUs. The HRUs were defined by using three slope classes (<2,6 %, 2.6 - 4,6 % and > 4,6 %) to focus on the flat topography of the catchment. To reduce the number of HRUs, slope classes and soil types with a contribution of less than 5 % and land use with a contribution of less than 2 % for each subbasin were defined as threshold to be aggregated during HRU generation. The input for soil data was a soil map with a resolution of 1:200.000 (BGR, 1999) and the slope classes were derived from the digital elevation model (LVermA, 1995) with a raster cell size of 5 m x 5 m. Land use data was generated from a mapping campaign in 2011/2012. Based on this land use mapping, 13 different crop rotations were derived and spatially defined for each subbasin. It was assumed that these crop rotations represent the long-term agricultural practice within the catchment. The soil and crop databases were taken from the SWAT model presented in Fohrer *et al.* (2013), which focused on accurate agricultural representation in the catchment. To define areas with drainage tiles, the estimated spatial distribution of drainage tiles for the catchment of Fohrer *et al.* (2007) was used.
Climate data was obtained from measured data and regional interpolated data. The weather station Gluecksburg-Meierwik in the north of the Kielstau catchment offers a good relation between precipitation and measured discharge. Since this weather station does not offer all needed climate data for the SWAT model, data from the Statistical Regional model (STAR, Orlowsky *et al.* 2008) was used to fill this gap. The STAR model generates a date-to-date mapping for weather stations by which a date from the observational period is assigned a date of the future projection period. Due to the small size of the Kielstau catchment, we identified only one suitable station situated in the nearby of the catchment. The climate data of this station was used for all other weather inputs as wind speed, temperature, solar radiation, and humidity.
To run $SWAT_{3S}$, the groundwater input files had to be reprocessed to add additional input parameters like the groundwater delay-time, the recession constant of the slow shallow aquifer, and the coefficent to separate the seepage of the slow shallow aquifer into seepage to the deep aquifer and the recharge of the slow shallow aquifer. The groundwater parameter values were all defined on HRU level. Due to the lack of information, we considered no spatial heterogeneity of the groundwater parameters. All groundwater parameters were equal for the whole catchment.

2.2.4 Model evaluation

For model evaluation, discharge data was divided into a calibration and a validation data set. A warm-up phase from 1996 to 1999 was used to achieve steady state conditions for the model. For calibration the discharge from 1999 to 2002 and from 2005 to 2008 was selected. For validation we choose the discharge from 2002 to 2005 and from 2008 to 2010. The selection of years for calibration and validation was based on the wetness conditions, which were

derived from the annual precipitation. With this selection, calibration and validation were characterized by equalled distribution of dry, regular, and wet periods. The model evaluation was then divided into the calibration and the validation process. Model evaluation was done with the R-package hydroGOF (Zambrano-Bigiarini, 2012) and a simple reproduction of the flow duration curve (Smakhtin, 2001; Vogel and Fennessey, 1994). As proposed in Moriasi *et al.* (2007), the Nash-Sutcliffe efficiency (NSE, Nash and Sutcliffe, 1970) and the percent bias (PBIAS, e.g., Gupta *et al.*, 1999) were used as performance measures for the modeled discharge. Legates and McCabe (1999) stated that the NSE is characterized by disproportional weighting of high values. Krause *et al.* (2005) found no existing sensitivity for systematic over- or underestimation for the NSE. As a consequence, we used this performance measure only for high flow dynamics and temporal discharge dynamics evaluation. Based on the findings of Boyle *et al.* (2001), Bekele and Nicklow (2007), van Werkhoven *et al.* (2009) and Pokhrel *et al.* (2012), we used two additional measures for calibration. For an adequate representation of all phases of the hydrograph, additional signature measures as proposed in Yilmaz *et al.* (2008) and Pokhrel *et al.* (2012) were used to calibrate the low flow and mid flow periods. These signature measures incorporate the PBIAS of the low flow with the 10 percentile (hereafter PBIASlow) of the flow duration curve (FDC) and a PBIAS of the mid flow with a range of 20 % to 70 % of time flow equaled or exceeded (hereafter PBIASmid).

2.2.4.1 Model calibration

Both model setups (two/three storage concept) were calibrated independently. The calibration parameters (Tab. 2) were chosen after Schmalz and Fohrer (2009), Tattari *et al.* (2009), Bärlund *et al.* (2007), van Griensven *et al.* (2006), and Santhi *et al.* (2001) and based on further projects with SWAT in the Kielstau catchment (Fohrer *et al.*, 2013; Kiesel *et al.*, 2010).

Sets of parameter value variations were generated for calibration using the Latin Hypercube Sampling of the R-package FME (Soetaert and Petzoldt, 2010). In this sampling, the space for each parameter is subdivided into equally-sized segments and one parameter value in each of the segments is drawn randomly. Depending on each parameter, the Latin Hypercube Sampling set represents different types of variation. Certain parameters were varied by simply replacing values as other parameters were varied by multiplication with the value of the Latin Hypercube Sampling set or by adding/subtracting parameter values (Tab. 2). To obtain a realistic simulation of drainage tile simulation, the depth of the impervious layer (DEP_IMP) has to be higher than for the drainage tiles (DDRAIN). This constellation was defined before the calibration. In the case that DDRAIN depth values were larger than the DEP_IMP value due to the multiplication of the Latin Hypercube Sampling, we modified the multiplication value of the Latin Hypercube Sampling for DEP_IMP. The initial value of DEP_IMP was multiplied by the multiplication value of DDRAIN so that DEP_IMP is larger than DDRAIN. Both SWAT model versions were calibrated with 5000 model runs (Tab. 2). The generation of the parameter sets and the replacement of parameter values in the SWAT model input files were executed with the R environment (R Core Team 2013).

Table 2: Selection of parameters for the calibration of the two storage and three storage SWAT model. Additional parameters for the three storage version are marked with * and were not used in the original two storage version of SWAT. The type of variation indicates, if the parameter value was varied by replacing (r), multiplication (m) or addition/subtraction (as).

Parameter name	Abbreviation	Process	Type of variation
Curve number	CN2 / CNURBAN	surface runoff/soil water	as
Surface runoff lag coefficient	SURLAG	surface runoff routing	r
Weighting factor normal flow	MSKCO1	channel routing	r
Weighting factor low flow	MSKCO2	channel routing	r
Weighting factor	MSKX	channel routing	r
Manning value for channel	CH_N	channel flow	r
Available soil capacity	SOL_AWC	soil water	as
Saturated hydraulic conductivity	SOL_K	soil water	m
Depth to impervious layer	DEP_IMP	lateral flow/drainage flow	m
Time to drain to field capacity	TDRAIN	drainage flow	m
Drain tile lag time	GDRAIN	drainage flow	m
Depth to subsurface drain	DDRAIN	drainage flow	m
Effective radius of drains	RE	drainage flow	r
Distance between two drain tiles	SDRAIN	drainage flow	r
Daily drainage coefficient	DRAINCO	drainage flow	r
Multiplication factor for lateral ksat	LATK	drainage flow	r
Soil evaporation compensation	ESCO	evapotranspiration	r
Plant uptake compensation facor	EPCO	evapotranspiration	r
Reach evaporation adjustment factor	EVRCH	evapotranspiration	r
Groundwater delay shallow aquifer	δ_{gw}	groundwater	r
Recession constant shallow aquifer	α_{gw}	groundwater	r
Percolation fraction deep aquifer	β_{dp}	groundwater	r
Recession constant fast shallow aquifer	$\alpha_{gw,fsh}*$	groundwater	r
Percolation fraction slow shallow aquifer	$\beta_{ssh}*$	groundwater	r
Groundwater delay slow shallow aquifer	$\delta_{gw,ssh}*$	groundwater	r
Recession constant slow shallow aquifer	$\alpha_{gw,ssh}*$	groundwater	r
Percolation fraction inactive deep aquifer	$\beta_{dp}*$	groundwater	r

The selection of the best calibration runs were performed stepwise for each model setup. In the first step, all calibration runs with NSE higher than 0.60 and within a PBIAS range of $-25\,\%$ to $25\,\%$ as recommended in Moriasi *et al.* (2007) were selected for a reduced dataset. To select calibration runs with acceptable performance in the mid-flow, the PBIASmid was used in a range from $-25\,\%$ to $25\,\%$ on the aforementioned reduced dataset. This reduced dataset was ordered by increasing absolute values of the PBIASlow to select the best 25 calibration runs. The same selection scheme was applied on the data set of calibration for the SWAT_{3S}. To identify the best calibration run for validation, we analyzed the 25 calibration runs by comparing the aforementioned performance measures and by inspecting the flow duration curve.

To evaluate the overall water balance characteristics, the datasets of the 25 best calibration runs for each model version were analyzed. Therefore, the contribution to the discharge was separated by the surface runoff (Q_{surf}), the lateral flow (Q_{lat}), the tile flow (Q_{tile}), and the groundwater contribution of the fast shallow aquifer (Q_{fsh}, SWAT_{3S}) and the shallow aquifer (Q_{sh}, SWAT_{orig}). In addition, the contribution of the slow shallow aquifer (Q_{ssh}) was extracted from the model outputs of SWAT_{3S}. Furthermore, the loss of groundwater out of the system and the mean annual loss due to evapotranspiration were extracted.

2.2.4.2 Model validation

After selecting the best calibration run, we used the aforementioned performance measures (NSE, PBIAS, PBIASmid, PBIASlow) for validation of the discharge from 2002 to 2005

and from 2008 to 2010 for $SWAT_{orig}$ and $SWAT_{3S}$. To analyze the discharge component distribution in detail, the discharge component separation was done for the best calibration run of the two storage and the three storage version. Furthermore the groundwater levels were analyzed to show differences in groundwater dynamics.

2.3 Results

In the first section, we present the results of general performances and characteristics of each model version, which are based on the calibration results. In the second section, we show the results of direct comparison for the best model run of each model version based on validation.

2.3.1 General performance and model characteristics

The 25 best calibration runs were used to describe the tendency of discharge contribution for the different discharge components (Fig. 4a, Fig. 4b) and the evapotranspiration (Fig. 4c).

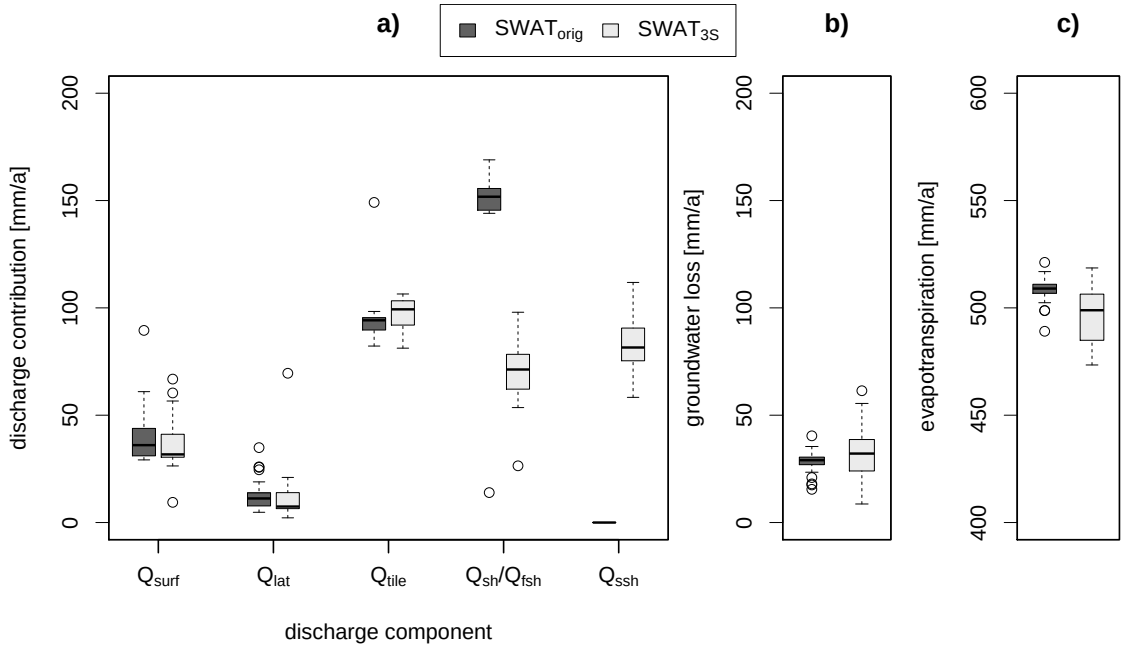

Figure 4: (a) Comparison of components to the discharge contribution for the original SWAT version and $SWAT_{3S}$ for the best 25 model runs. Components are surface runoff (Q_{surf}), the lateral flow (Q_{lat}), the tile flow (Q_{tile}) and the groundwater contribution of the shallow aquifer (Q_{sh})/fast shallow aquifer (Q_{fsh}). The contribution of the slow shallow aquifer (Q_{ssh}) is only available for the three storage version as the two storage version does not allow contribution of the deep aquifer. Groundwater loss due to inactivated contribution by the aquifer is shown in (b). (c) shows evapotranspiration for both model versions.

The comparison of the discharge components for discharge contribution reveals that the surface runoff, the lateral flow, and the tile drainage flow is similar (Fig. 4a). The average groundwater contribution for the two storage version is about 155 mm/a. To compare the overall groundwater contribution of $SWAT_{3S}$, the contribution of the fast shallow aquifer (71 mm/a) and the slow shallow aquifer (82 mm/a) were summed up. This comparison shows that both

versions simulate nearly the same groundwater contribution (152 mm/a vs. 153 mm/a). Additionally, the inactivated aquifers are compared for each groundwater module (Fig. 4b). The annual loss of groundwater, which does not occur at the discharge gauge at the outlet of the catchment, differs for the two groundwater modules. $SWAT_{3S}$ predicts a mean groundwater loss to the deep aquifer of 32 mm/a and the original SWAT version a mean loss of 29 mm/a. Referring to the evapotranspiration, the two versions have only small differences in the mean evapotranspiration. As already observed in the groundwater components, variation is higher for the three storage version. In the mean, $SWAT_{3S}$ tends to have less evapotranspiration.

After the selection of the 25 best calibration runs for each model version, the datasets were analyzed for the tendency of measure performance (Fig. 5). In the mean, $SWAT_{3S}$ and $SWAT_{orig}$ produce the same NSE performance (0.65 vs. 0.66). Referring to the PBIAS and the PBIASmid, both model versions tend to overestimate the overall discharge and the mid flow. The most important difference between the two versions becomes obvious in the low flow period. The mean PBIASlow of $SWAT_{3S}$ (−1,2 %) is near to optimum value. In contrast, the best two PBIASlow value of $SWAT_{orig}$ are at -2.8 and 7,7 %. However, these values are at the outer range for $SWAT_{orig}$ since the mean PBIASlow is at 65,9 %. As a consequence most of the model runs predict remarkably less discharge in this range of the flow duration curve.

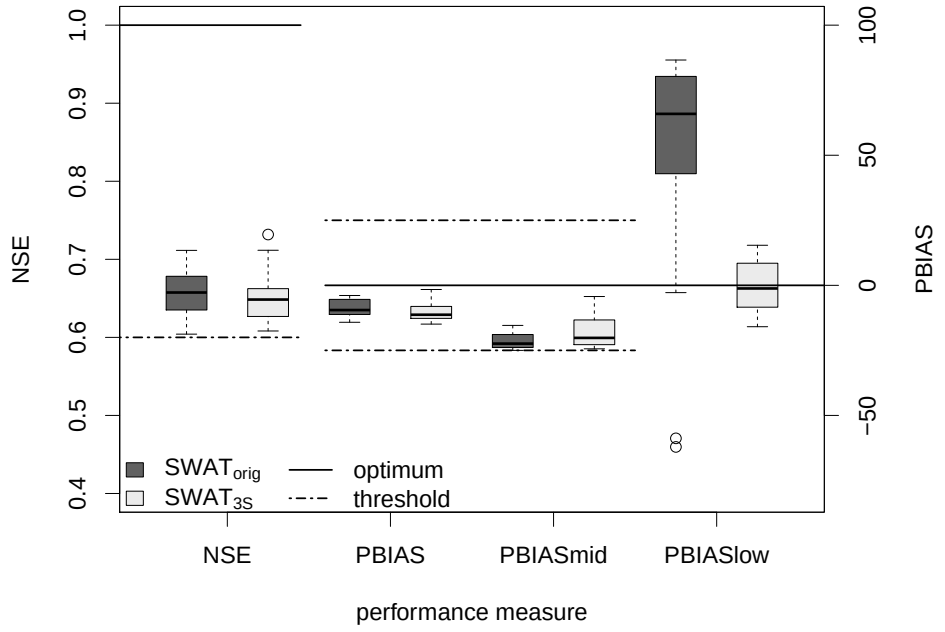

Figure 5: Comparison of performance measures for the best 25 calibration runs of the original SWAT model and $SWAT_{3S}$. For the NSE, the dotted lines indicate the threshold value for the selection of best calibration runs and the continuous line the optimum value. For the PBIAS and PBIASmid, the dotted lines indicate the threshold value for good model performance between −25 % and 25 %. The continuous line marks the optimum value for all PBIAS measures.

To identify differences in parameter value distribution for each model version, the total range and the range for the 25 best calibrations for all calibration parameters were summarized

(Tab. 3). Largest differences were found for the groundwater parameters. Referring to the ranges of the best 25 model runs, $SWAT_{3S}$ and $SWAT_{orig}$ show nearly the same parameter range values except for SOL_AWC and ESCO. $SWAT_{3S}$ uses ESCO values up to 0.95 whereas maximum values of $SWAT_{orig}$ are at 0.84. For the available soil water capacity, both model version use the same maximum increase of 0.1. Nonetheless the calibrated parameter range is smaller for $SWAT_{orig}$ than for $SWAT_{3S}$ (-0.01 vs. 0.1 vs. -0.05 vs. 0.1).

Table 3: Parameter variation values of the whole Latin Hypercube Sampling and for the 25 best calibration runs of the two storage and three storage SWAT model. Additional parameters for the three storage version are marked with * and were not used in the original two storage version of SWAT. Calibrated parameter values are shown for the best model run of each SWAT version.

Parameter	Total range	$SWAT_{orig}$		$SWAT_{3S}$	
		best range	calibrated value	best range	calibrated value
CN2	-10 - 10	-9.60 - 8.50	- 2.4	-9.87 - 9.46	- 5.2
CNURBAN	-15 - 0	-15 - - 0.7	- 13.6	-14.93 - -0.2	- 11.4
SURLAG	0.2 - 1.2	0.20 - 1.20	0.58	0.21 - 1.14	0.85
MSKCO1	0.1 - 0.5	0.10 - 0.50	0.12	0.10 - 0.50	0.43
MSKCO2	0.5 - 3.5	0.53 - 3.24	1.58	0.78 - 3.38	1.0
MSKX	0.1 - 0.4	0.12 - 0.39	0.28	0.12 - 0.38	0.14
CH_N	0.014 - 0.035	0.015 - 0.034	0.03	0.018 - 0.035	0.033
SOL_AWC	-0.07 - 0.1	- 0.01 - 0.1	+ 0.10	-0.05 - 0.1	+ 0.02
SOL_K	0.5 - 2	× 0.5 - 2.0	× 0.60	× 0.6 - 2.0	× 1.6
DEP_IMP	0.7 - 1.3	× 0.84 - 1.29	× 1.10	× 0.73 - 1.22	× 1.06
TDRAIN	0.5 - 1.5	× 0.52 - 1.34	× 0.78	× 0.62 - 1.50	× 1.14
GDRAIN	0.5 - 1.5	× 0.50 - 1.45	× 0.91	× 0.50 - 1.50	× 1.46
DDRAIN	0.75 - 1.25	× 0.79 - 1.24	× 1.12	× 0.78 - 1.24	× 1.24
RE	10 - 50	10 - 50	40.75	10 - 50	29.2
SDRAIN	15000 - 45000	16214 - 43373	22923	15915 - 43900	29914
DRAINCO	5 - 20	5.2 - 17.9	11.00	5.8 - 19.2	13.5
LATK	1 - 4	1.46 - 3.97	1.70	1.08 - 3.78	1.09
ESCO	0.7 - 1.0	0.70 - 0.84	0.73	0.70 - 0.95	0.73
EPCO	0.7 - 1.0	0.72 - 1.00	0.91	0.71 - 1.00	0.80
EVRCH	0.5 - 1.0	0.50 - 1.00	0.71	0.53 - 0.98	0.72
δ_{gw}	1 - 30	1 - 30	1.2		
α_{gw}	0.01 - 0.1	0.01 - 1.00	0.01		
β_{dp}	0 - 0.2	0.08 - 0.2	0.10		
$\delta_{gw,fsh}$	1 - 30			1 - 24	12
$\alpha_{gw,fsh}$*	0.2 - 1.0			0.2 - 1.0	0.58
β_{ssh}*	0.2 - 0.7			0.5 0.7	0.64
$\delta_{gw,ssh}$*	5 - 50			7 - 47	11
$\alpha_{gw,ssh}$*	0.001 - 0.04			0.002 - 0.026	0.004
β_{dp} *	0.05 - 0.5			0.08 - 0.49	0.38

To evaluate the general performance, we selected the best 25 calibration runs of $SWAT_{3S}$ and $SWAT_{orig}$. The FDCs of the selected runs were compared with the observed discharge data (Fig. 6). The results of this comparison reveal that $SWAT_{3S}$ and $SWAT_{orig}$ performed similar in the FDC segment from 0 % to 50 %. In the segment from 50 % to 70 %, both model versions tend to overestimate the discharge. In constrast to the original SWAT model, $SWAT_{3S}$ produces less overestimation, especially with increasing flow exceedance probability (Fig. 6b). $SWAT_{orig}$ produces an systematic overestimation at 50 % since no model run predicts discharge at the observed discharge level (Fig. 6a). At a flow exceedance probability with more than 70 %, $SWAT_{orig}$ mainly underpredicts the discharge. This can be seen especially after the 95 % exceedance probability, where $SWAT_{orig}$ tends to predict days with no discharge. In contrast, $SWAT_{3S}$ predicts discharge almost until 100 % of flow exceedance probability.

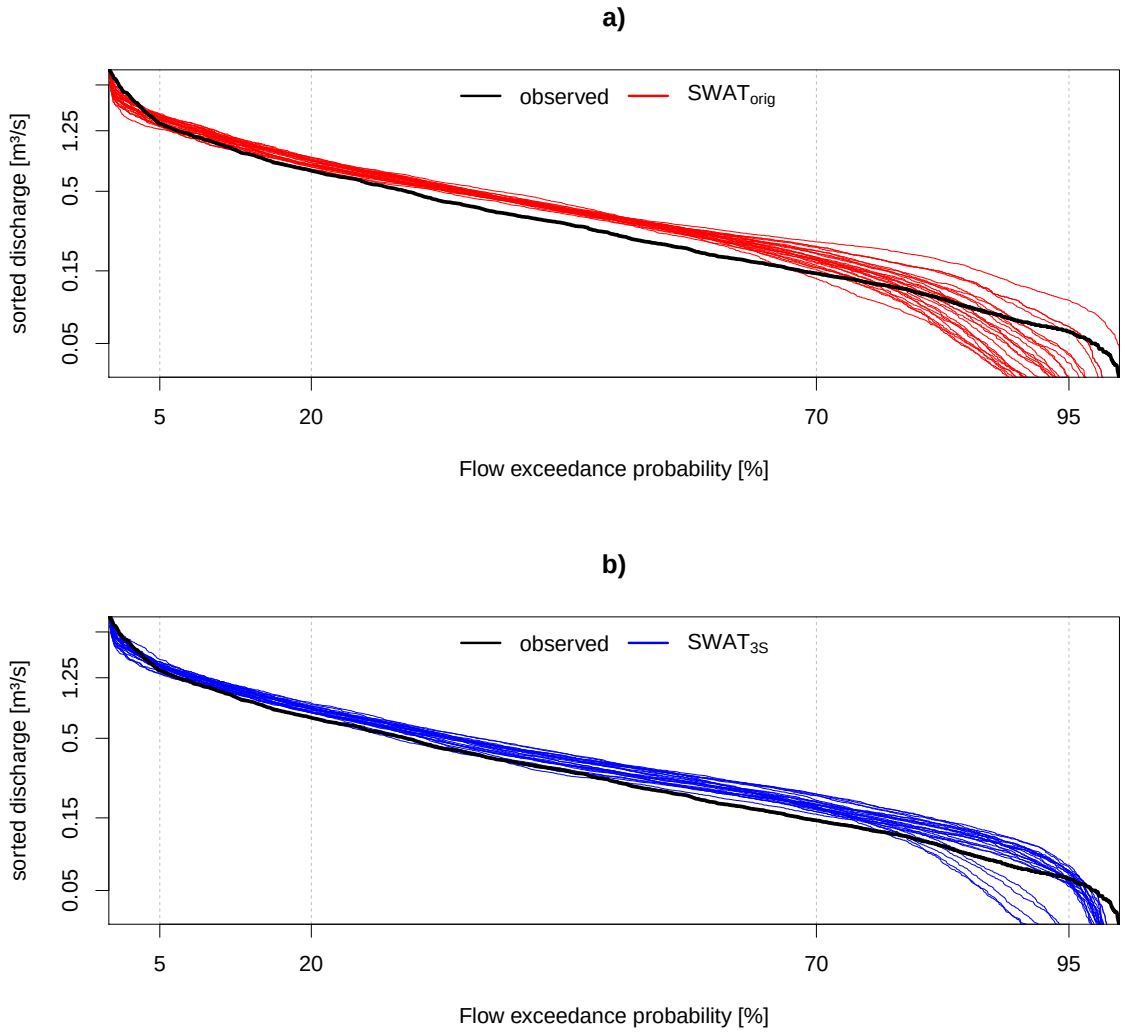

Figure 6: Flow duration curve of the observed discharge (black line) for the calibration period (1999 to 2002; 2005 to 2008). a) discharge of the best 25 calibration runs of the two storage version (red line). b) discharge of the best 25 calibration runs of the three storage version (blue line).

2.3.2 Comparison of the best model runs

For the evaluation of single model performance, we compared the best model run of each SWAT version for the validation period. The discharge dynamics are shown in hydrographs for the different model versions (Fig. 7). The original model tends to predict dry periods with days of no discharge (August 2003/2009 and July 2004). The recession limbs of the original SWAT model are overestimated in time periods of November 2002 till April 2003 and December 2009 till May 2010. SWAT$_{3S}$ underestimates the recession in December 2004 till May 2005 slightly. The visual comparison shows that SWAT$_{3S}$ is superior over the original SWAT model in reproducing late recession limbs with following baseflow. Additionally the extended groundwater module predicts higher baseflow at dry periods than the two storage version. The performance measures were used to analyse the adequacy of the models for the different phases of the hydrograph (Tab. 4).

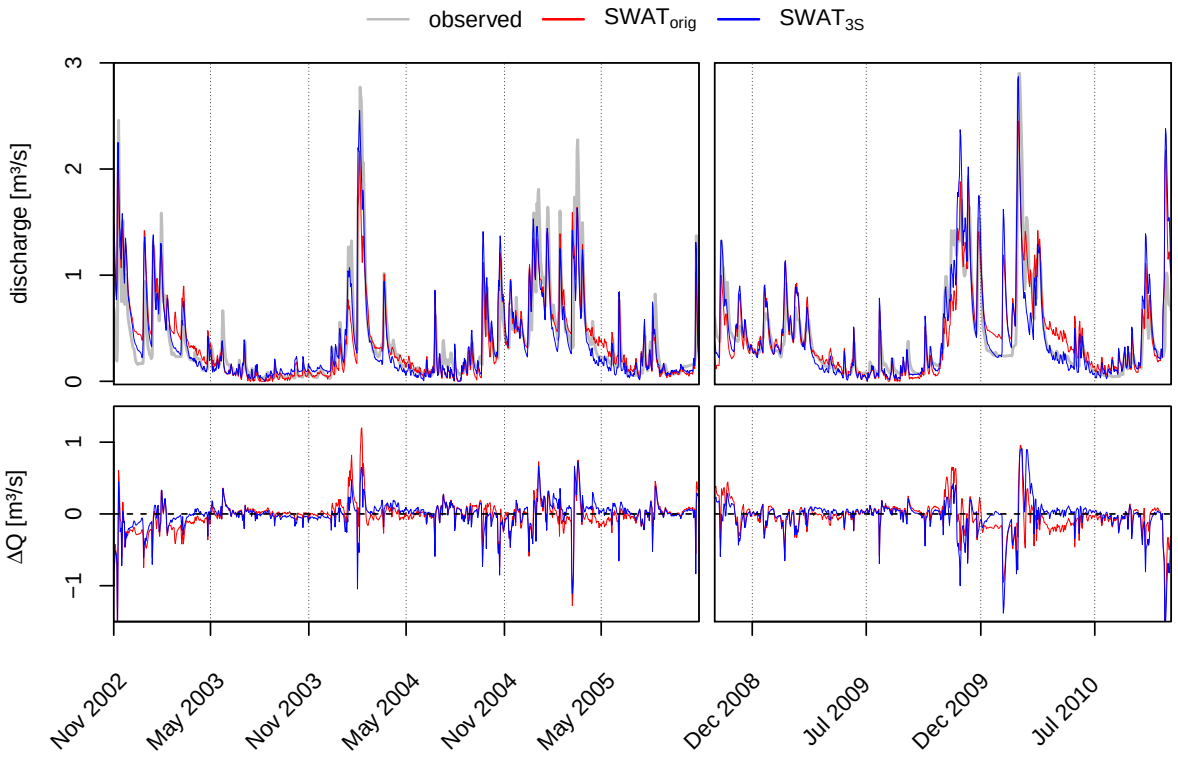

Figure 7: Observed and modelled discharge together with residuals between observed and modeled discharge for the validation period 2003-2005 and 2009-2010. The grey line indicates the observed discharge. The red line is the modeled discharge of the original model (SWAT$_{orig}$) and the blue line is the model output of the modified model (SWAT$_{3S}$).

The results of the validation reveal that both SWAT versions perform similar in predicting the discharge dynamics (NSE SWAT$_{3S}$ 0.72 vs. NSE SWAT$_{orig}$ 0.73). The PBIAS is near to the optimal measure value for both model versions (SWAT$_{3S}$: −4,4 % and SWAT$_{orig}$: −6,3 %). Referring to the prediction of the mid segment of the flow duration curve, SWAT$_{3S}$ produces smaller overestimationts than SWAT$_{orig}$ (−2,4 % vs. −15,9 %). Comparing the low flow segment of the FDC, differences become obvious. The original SWAT version predicts less discharge than observed (46,8 %). The modified version predicts a difference of only 14,8 %.

Table 4: Performance and signature measure values for the calibration and validation of the two storage and three storage SWAT version

	calibration				validation			
version	NSE	PBIAS	PBIAS mid	PBIAS low	NSE	PBIAS	PBIAS mid	PBIAS low
SWAT$_{orig}$	0.66	9.0	23.8	2.8	0.73	-6.3	-15.9	46.8
SWAT$_{3S}$	0.67	-5.4	-11.1	-3.6	0.72	-4.4	-2.4	14.8

The comparison of the discharge components for discharge contribution reveals that the surface runoff and the lateral flow is similar (Fig. 8). Tile flow is reduced for $SWAT_{3S}$ since the contribution of the fast shallow aquifer is higher at periods with high discharge than for $SWAT_{orig}$. Remarkably is the difference in groundwater contribution during the recession phase between the wet and the dry period. While $SWAT_{3S}$ predicts a continuous flow from the slow shallow aquifer into the stream and a fast decreasing amount of water entering the channel from the fast shallow aquifer, $SWAT_{orig}$ describes a slow decreasing contribution of the shallow aquifer. Furthermore, the slow shallow aquifer describes a delayed answer of the wet period since the highest contribution takes place between wet and dry period.

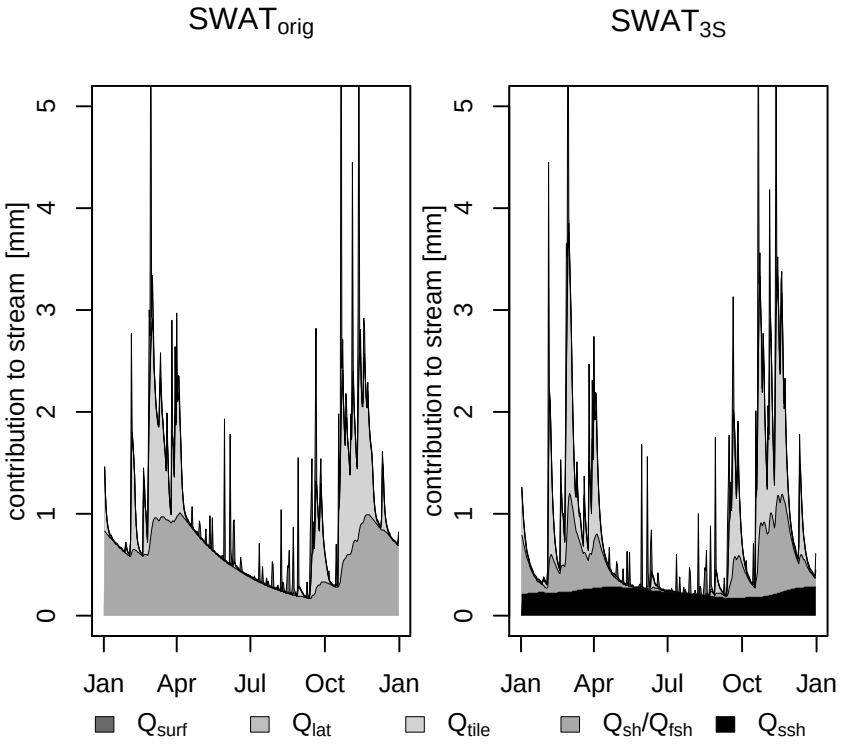

Figure 8: Comparison of components to the discharge contribution in 2010 for the two and the three storage version. Components are surface runoff (Q_{surf}), the lateral flow (Q_{lat}), the tile flow (Q_{tile}), and the groundwater contribution of the shallow/fast shallow aquifer ($Q_{sh}Q_{fsh}$). The contribution of the slow shallow aquifer (Q_{ssh}) is only available for the three storage version as the two storage version does not allow contribution of the second aquifer.

The final values for groundwater parameters after calibration are summarized in Tab. 3. The calibrated groundwater parameter values for the different groundwater modules are differing widely. The recession constant α_{gw} of the two storage version is much smaller (0.01) than the $\alpha_{gw,fsh}$ of the three storage version (0.58) but larger than the $\alpha_{gw,ssh}$ of the three storage version (0.004). The same pattern can be found for the partitioning coefficient of the deep aquifer β_{dp}, which is smaller for the original version than for the three storage version (0.10 and 0.38). The delay time of the original SWAT version is higher (1.2) compared to the delay times of the shallow aquifer ($\delta_{gw,fsh}$: 12.1, $\delta_{gw,ssh}$: 11.4). Differences can be also seen for SOL_K ($SWAT_{3S}$: $\times$ 1.6, $SWAT_{orig}$: $\times$ 0.6), CN2 ($SWAT_{3S}$: -5.2, $SWAT_{orig}$: -2.4) and for

SOL_AWC ($SWAT_{3S}$: 0.02, $SWAT_{orig}$: 0.1). However, the most important differences can be found for the groundwater parameters.

Comparing the best calibration runs for the groundwater level dynamics, the same tendency as for the 25 best model runs is revealed. $SWAT_{3S}$ predicts a groundwater loss to the deep aquifer of 61 mm and the original SWAT version a loss of 22 mm (Fig. 9). For the three storage concept, the fast shallow aquifer shows only little dynamics, whereas the slow shallow aquifer water levels can be described as an amplitude with its valley in the late summer period. In contrast, $SWAT_{orig}$ shows high variability in groundwater levels for the shallow aquifer. The loss to the deep aquifer is highest in winter and spring. In comparison to the three storage concept, the shallow aquifer water levels show much larger variability. The dynamic of water entering the deep aquifer is similar to the three storage version.

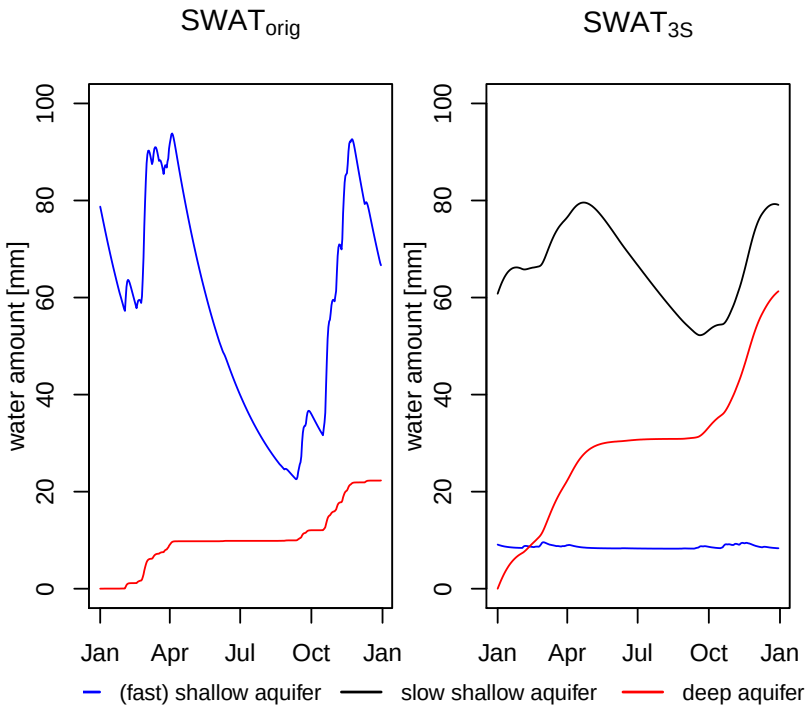

Figure 9: Comparison of water levels for the three storage version (left side) and the two storage version (right side). The red line (loss aquifer) indicates the cumulative loss of groundwater, which is inactive for contribution to the channel. The black line indicates the water level of the slow shallow aquifer for $SWAT_{3S}$. The blue line shows the water levels of the fast shallow aquifer ($SWAT_{3S}$) and the shallow aquifer ($SWAT_{orig}$).

2.4 Discussion

In the following, we discuss the results of the general performance and characteristics of the different SWAT versions in the context of the representation of groundwater processes. Firstly, we compare the groundwater parameters between both model versions, which are the dominant control parameters for the timing of the groundwater contributions. Based on the groundwater parameter dominance, we raised the role of the other model parameters and the importance of multiple performance measures to detect overall performance of model calibrations and validations. Finally, we discuss the modular structure of the new groundwater module.

The results show limitations of the original SWAT version in recession and low flow periods as already described by Guse *et al.* (2013), Koch *et al.* (2013), Eckhardt (2008), Wu and Johnston (2007), and Watson *et al.* (2003). The distributions of the flow duration curves (Fig. 6) reveal that this seems to be a model structure deficit since most of the model runs result in underestimation of the low flow segment. The single reservoir of SWAT_{orig} is unable to describe the fast flow and the slow flow component of the groundwater contribution exactly with the same parameter set. As major improvement, the calibration of the new SWAT_{3S} model resulted in a large number of good model runs for the low flow segment with adequate representation of the mid flow segment. As a consequence, a groundwater structure with two contributing storages seems to be better suited for lowland catchments. The fast flow component controls the recession phase mainly with the fast answer of the fast shallow aquifer. Additionally, the contributing slow shallow aquifer controls the baseflow with its delayed answer of groundwater recharge. A second control option for the baseflow is the loss to the deep aquifer which is especially important for small catchments.

The detailed comparison of the best model runs for both models revisits the topic of model structure and behaviour. The groundwater parameter setting of the single reservoir solution of SWAT_{orig} differs remarkably from the concept of two contributing reservoirs. The time delay for the fast shallow groundwater recharge of SWAT_{3S} is higher than the time delay for the shallow aquifer of the original SWAT version. The delay time implies a delay between the time, water percolates out of the soil and reaches the aquifer. Because of that, the amount of groundwater which is available for groundwater contribution to the stream is limited by the groundwater delay. As this delay time of the original SWAT is very short, a large amount of water is available for the contribution to the stream.

The baseflow recession coefficient is the second dominant parameter, which controls the groundwater contribution. For SWAT_{orig}, the parameter value of α_{gw} is between the $\alpha_{gw,fsh}$ and $\alpha_{gw,ssh}$ values for the fast and the slow aquifer of SWAT_{3S}. In connection with the short delay time for groundwater recharge, the small α_{gw} limits the groundwater flow into the channel. This shows the interacting relationship between the delay time of groundwater recharge and the baseflow recession constant, which can be also found for SWAT_{3S}. The groundwater delay time for the fast shallow aquifer is higher than for the original SWAT but $\alpha_{gw,fsh}$ induces a fast groundwater contribution. With the second recession constant $\alpha_{gw,ssh}$, the slow shallow aquifer describes a long lasting groundwater contribution.

Referring to model behaviour of SWAT$_{orig}$, it becomes clear, why it is challenging to reproduce different recession and low flow phases with only one contributing groundwater storage. Due to the small delay time, the recharge of the storage is very high, but the contribution has to be limited to store water in the aquifer for low flow periods by long lasting recession. This long lasting recession results in overestimation of the discharge.
As the modified version uses an additional delay for the slow shallow aquifer, the characteristic of groundwater contribution differs from the original SWAT. Less than half of the percolating water recharges the fast shallow aquifer of SWAT$_{3S}$, which contributes quickly into the channel. The remaining percolation water recharges mainly the slow shallow aquifer as only a small part of the recharge is diverted to the deep aquifer, which is not connected to the channel. In contrast to the high dynamics in the fast shallow aquifer, the slow shallow aquifer stores the seepage water and flows very slowly into the channel. The parameter setting of delay time and recession constant results in a smoothed amplitude for the groundwater levels in the slow shallow aquifer and a flattened contribution curve. The groundwater levels of the fast shallow aquifer are more or less constant at a low level. This constant water level can be explained by the high baseflow recession constant, which passes most of the recharge to the direct contribution into the channel. As both groundwater storages show the expected behavior, the parameter settings of SWAT$_{3S}$ are concise with the expected role of the newly introduced parameters. Furthermore, this model behaviour accounts for the shallow water tables of our lowland catchment, where boundaries between groundwater and fully saturated soils can be defined hardly.
Referring to the timing of contribution, the fast shallow aquifer contribution is highest in wet periods. During the dry periods, the slow shallow aquifer is the main source for baseflow. In comparison, the one storage solution shows at the transition between wet and dry condition a long stretched groundwater contribution, which decreases very slow. In general, this description may be realistic but in this case, the decrease of groundwater contribution is too small and causes the overestimation of the discharge. This problem seems to be weakened by the two storage approach since the decrease of groundwater contribution is much higher. Referring to the proposed groundwater representation by Fenicia *et al.* (2006), Staudinger *et al.* (2011), and Brandes *et al.* (2005), the usage of two contributing groundwater storages and one storage for groundwater loss results in an emphasized nonlinear description of the groundwater flow.
Comparing the parameters which are not directly connected with the groundwater processes, differences for the calibrated values became obvious. Differences were found especially for the soil parameter SOL_K and SOL_AWC, but also for the surface run-off CN2. For SWAT$_{3S}$, the conductivity of saturated soils was increased and for SWAT$_{orig}$ they were reduced. Referring to the available soil water capacity SOL_AWC, SWAT$_{orig}$ shows much higher calibrated parameter values. For SWAT$_{3S}$, the reduction of CN2 was two times higher than for SWAT$_{orig}$. Despite of the different calibrated parameter values, there is no clear tendency for differences in model behaviour due to different model structures. The combination of the soil parameters for the SWAT$_{orig}$ suggests that the soil acts as a delay storage to compensate for the missing delay storage of the groundwater. However, this cannot be confirmed for the

general behaviour since the parameter ranges for the best 25 model runs are similar for both model versions for SOL_K. Furthermore, the range differences for parameter SOL_AWC were found at the lower part of the parameter space. The fact that the most obvious parameter differences were found for the groundwater module of the SWAT version indicates that our modifications mainly affect groundwater processes. This is in accordance with the findings of Guse *et al.* (2013) who reported that groundwater parameters are the dominant parameters during recession and low flow phases. Nonetheless the new groundwater structure of $SWAT_{3S}$ needs to be tested in other catchments to identify possible structure deficits.
Concerning the calibration process and the selection of best model runs, the importance of the proposed combination of multiple performance measures (Pokhrel *et al.*, 2012; van Werkhoven *et al.*, 2009; Bekele and Nicklow, 2007; Boyle *et al.*, 2001) was shown. Although both model versions achieved good performance for the NSE and PBIAS, many calibrations revealed poor low flow prediction. This poor performance in low flow periods was identified with the additional PBIASlow measure. Referring to recession phases, problems occurred in identifying poor model performance. Although all selected performance measures were in an acceptable range, inadequate discharge reproduction occurred between the mid flow and the low flow segment of the FDC. Further research should address combined performance measures which allow identifying good calibration runs for all phases of the hydrograph.
As already mentioned in the methods section, the groundwater module has a modular structure since the number of active contributing aquifers can be selected modularly on HRU or subbasin level. Consequently, the representation of the characteristic groundwater processes can be adapted to different catchments. Due to the extended groundwater concept, flexibility in groundwater modeling is enhanced. Referring to spatial heterogeneity, groundwater aquifers can be activated and deactivated for each subcatchment which might lead to a better representation of large catchments.
Despite of the promising results for the simulation of recession and baseflow phases with the modified version, the new concept has to be verified by field measurements. It has to be kept in mind that representation of the groundwater processes are simplified and aggregated into a conceptual model. Field measurements should clarify if the extended groundwater concept is an acceptable simplification of the groundwater processes.

2.5 Conclusion

In this paper we investigated if nonlinearities of groundwater processes can be emphasized in the current version of the SWAT model. Since nonlinearities of groundwater processes cannot be reproduced adequately by only one active groundwater storage, we implemented one additional storage including additional time delay functions. With this, it was possible to reach an enhanced representation of the low flow periods. The shallow aquifer was separated into a fast and into a slow flow component. The results show that the extended groundwater module leads to a more process-oriented groundwater modeling as it is shown in the calibration of recession limbs and subsequent low flow phases. Calibrations lead to parameters for the fast shallow aquifer, which describe fast groundwater recharge and a fast response on groundwater

recharge. Referring to the slow shallow aquifer, the calibrated parameters describe a slow groundwater flow and recharge. The calibration of the recession limb and the low flow in the original version was a trade-off between both discharge phases. As a consequence, the calibrated parameters for the only contributing aquifer describe a fast groundwater recharge process with a slow groundwater flow into the channel.
With the process-oriented extension of the groundwater module in $SWAT_{3S}$, very good results of signature and performance measures were achieved. In contrast, the original SWAT version performed less adequate, especially in times of low flow. The results showed that the prediction of low flow and recession has to be verified with appropriate performance measures.

2.6 Acknowledgements

We thank the Government-Owned Company for Coastal Protection, National Parks and Ocean Protection (LKN-SH) of Schleswig-Holstein for the discharge data, the land survey office of Schleswig-Holstein for providing the digital elevation model and the river net, the German Weather Service (DWD) for the climate data and the Potsdam Institute for Climate Impact Research (PIK) for the STAR data. This project has been carried out with financial support of a scholarship for the first author by the German Environmental Foundation (DBU). This work was partially supported by the German Federal Ministry for Education and Research (BMBF) via the project IMPACT (grant number 02WM1136). We would like to thank the community of the open source software R, which was used for calibration of the SWAT model and following analysis. Also, we are very grateful to Michael Strauch and the anonymous reviewer for their contructive and thoughtful suggestions.

2.7 References

Arnold J G, Fohrer N. 2005. Swat2000: Current capabilities and research opportunities in applied watershed modelling. *Hydrological Processes* 19: 563–572.

Arnold J G, Srinivasan R, Muttiah R S, Williams J R. 1998. Large area hydrologic modeling and assessment part I: Model development. *Journal of the American Water Resources Association 34* 1: 73–89.

Bärlund I, Kirkkala T, Malve O, Kämäri J. 2007. Assessing SWAT model performance in the evaluation of management actions for the implementation of the Water Framework Directive in a Finnish catchment. *Environmental Modelling & Software* 22: 719–724.

Bekele E G, Nicklow J W. 2007. Multi-objective automatic calibration of SWAT using NSGA-II. *Journal of Hydrology* 341: 165–176.

BGR. 1999. Bundesanstalt fuer Geowisschenschaften und Rohstoffe - Bodenuebersichtskarte im Maßstab 1:200.000. Verbreitung der Bodengesellschaften.

Borah D K, Bera M. 2003. Watershed-scale hydrologic and nonpoint-source pollution models: Review of mathematical bases. *Trans. ASAE 46* 6: 1553–1566.

Boyle D P, Gupta H V, Sorooshian S, Koren V, Zhang Z, Smith M. 2001. Toward improved streamflow forecasts: Value of semidistributed modeling. *Water Resources Research* 37: 2749–2759.

Brandes D, Hoffmann J G, Mangarillo J T. 2005. Base flow recession rates, low flows, and hydrologic features of small watersheds in Pennsylvania, USA. *JAWRA Journal of the American Water Resources Association* 41: 1177–1186.

Bronstert A, Kolokotronis V, Schwandt D, Straub H. 2007. Comparison and evaluation of regional climate scenarios for hydrological impact analysis: General scheme and application example. *International Journal of Climatology* 27: 1579–1594.

Cibin R, Sudheer K P, Chaubey I. 2010. Sensitivity and identifiability of stream flow generation parameters of the SWAT model. *Hydrological Processes* 24: 1133–1148.

David M B, Grosso S J D, Hu X, Marshall E P, McIsaac G F, Parton W J, Tonitto C, Youssef M A. 2009. Modeling denitrification in a tile-drained, corn and soybean agroecosystem of Illinois, USA. *Biogeochemistry* 93: 7–30.

De Vos N J, Rientjes T H M, Gupta H V. 2010. Diagnostic evaluation of conceptual rainfall–runoff models using temporal clustering. *Hydrological Processes* 24: 2840–2850.

DWD. 2012. Weather and climate data from the German Weather Service of the station Flensburg (1961-1990). *Online climate data* .

Eckhardt K. 2008. A comparison of baseflow indices, which were calculated with seven different baseflow separation methods. *Journal of Hydrology* 352: 168–173.

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: 139–150.

Fohrer N, Dietrich A, Kolychalow O, Ulrich U. 2013. Assessment of the environmental fate of the herbicides flufenacet and metazachlor with the SWAT model. *Journal of Environmental Quality* 42: 1–11. doi:10.2134/jeq2011.0382.

Fohrer N, Schmalz B. 2012. Das UNESCO Oekohydrologie-Referenzprojekt Kielstau-Einzugsgebiet - Nachhaltiges Wasserressourcenmanagement und Ausbildung im laendlichen Raum. *Hydrologie und Wasserbewirtschaftung* 4: 160–168.

Fohrer N, Schmalz B, Tavares F, Golon J. 2007. Modelling the landscape water balance of mesoscale lowland catchments considering agricultural drainage systems. *Hydrologie und Wasserbewirtschaftung* 51: 164–169.

Gonzales A L, Nonner J, Heijkers J, Uhlenbrook S. 2009. Comparison of different base flow separation methods in a lowland catchment. *Hydrology and Earth System Sciences* 13: 2055.

Gupta H V, Sorooshian S, Yapo P O. 1999. Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration. *Journal of Hydrologic Engineering* 4: 135–143.

Guse B, Reusser D E, Fohrer N. 2013. How to improve the representation of hydrological processes in SWAT for a lowland catchment - Temporal analysis of parameter sensitivity and model performance. *Hydrological Processes* in press.

Hattermann F, Krysanova V, Wechsung F, Wattenbach M. 2004. Integrating groundwater dynamics in regional hydrological modelling. *Environmental Modelling & Software* 19: 1039–1051.

Huang S, Krysanova V, Hattermann F F. 2013. Projection of low flow conditions in Germany under climate change by combining three RCMs and a regional hydrological model. *Acta Geophysica* 61: 151–193.

Kiesel J, Fohrer N, Schmalz B, White M J. 2010. Incorporating landscape depressions and tile drainages of a northern German lowland catchment into a semi-distributed model. *Hydrological Processes* 24: 1472–1486.

Kirchner J W. 2009. Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward. *Water Resources Research* 45: W02429.

Koch S, Bauwe A, Lennartz B. 2013. Application of the SWAT model for a tile-drained lowland catchment in North-Eastern Germany on subbasin scale. *Water Resources Management* 27: 791–805.

Krause P, Boyle D P, Bäse F. 2005. Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences* 5: 89–97.

Krause S, Bronstert A. 2007. The impact of groundwater–surface water interactions on the water balance of a mesoscale lowland river catchment in northeastern Germany. *Hydrological Processes* 21: 169–184.

Krysanova V, Arnold J G. 2008. Advances in ecohydrological modelling with SWAT: A review. *Hydrological Sciences Journal* 53: 939–947.

Legates D R, McCabe G J. 1999. Evaluating the use of goodness-of-fit measures in hydrologic and hydroclimatic model validation. *Water Resources Research* 35: 233–241.

LKN. 2012. Tagesmittelwerte der Messstelle Soltfeld: Wasserstand und Abfluss. *Landesbetrieb fÃ¼r KÃ¼stenschutz, Nationalpark und Meeresschutz Schleswig-Holstein; Hydrologie, Mess- und Beobachtungsdienst* .

Lundqvist J. 1986. Late Weichselian glaciation and deglaciation in Scandinavia. *Quaternary Science Reviews* 5: 269–292.

Luo Y, Arnold J, Allen P, Chen X. 2012. Baseflow simulation using SWAT model in an inland river basin in Tianshan Mountains, Northwest China. *Hydrology and Earth System Sciences* 16: 1259.

LVermA. 1995. Landesvermessungsamt Schleswig-Holstein Digitales Geländemodell fuer Schleswig-Holstein. Quelle: TK25. Gitterweite 25 m x 25 m und TK50 Gitterweite 50 m x 50 m sowie ATKIS-DGM2-1 m x 1 m Gitterweite und DGM 5 m x 5 m Gitterweite, abgeleitet aus LiDAR-Daten.

Moriasi D N, Arnold J G, Liew M W V, Bingner R L, Harmel R D, Veith T L. 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE* 50: 885–900.

Munz M, Krause S, Tecklenburg C, Binley A. 2011. Reducing monitoring gaps at the aquifer–river interface by modelling groundwater–surface water exchange flow patterns. *Hydrological Processes* 25: 3547–3562.

Nash J E, Sutcliffe J V. 1970. River flow forecasting through conceptual models: Part 1 a discussion of principles. *Journal of Hydrology* 10: 282–290.

Nathan R J, McMahon T A. 1990. Evaluation of automated techniques for base flow and recession analyses. *Water Resources Research* 26: 1465–1473.

Neitsch S L, Arnold J G, Kiniry J R, Williams J R. 2011. SWAT theoretical documentation version 2009. *Grassland, Soil and Water Research Laboratory, Agricultural Research Service. Blackland Research Center, Texas Agricultural Experiment Station* .

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: 80–93.

Orlowsky B, Gerstengarbe F W, Werner P C. 2008. A resampling scheme for regional climate simulations and its performance compared to a dynamical RCM. *Theoretical and Applied Climatology* 92: 209–223.

Pokhrel P, Yilmaz K K, Gupta H V. 2012. Multiple-criteria calibration of a distributed watershed model using spatial regularization and response signatures. *Journal of Hydrology* 418: 49–60.

R Core Team. 2013. R: A language and environment for statistical computing. Vienna, Austria: R foundation for statistical computing; 2013.

Riedel W, Polenky R. 1987. Forschungsprojekt des Institutes fuer Regionale Forschung und Information im Deutschen Grenzverein e.V. in Zusammenarbeit mit der Zentralstelle fuer Landeskunde des Schleswig-Holsteinischen Heimatbundes. *Umweltatlas fuer den Landesteil Schleswig* .

Saenger N, Kitanidis P K, Street R L. 2005. A numerical study of surface-subsurface exchange processes at a riffle-pool pair in the Lahn River, Germany. *Water Resources Research* 41: W12424.

Samuel J, Coulibaly P, Metcalfe R A. 2012. Identification of rainfall–runoff model for improved baseflow estimation in ungauged basins. *Hydrological Processes* 26: 356–366.

Santhi C, Arnold J G, Williams J R, Dugas W A, Srinivasan R, Hauck L M. 2001. Validation of the SWAT model on a large river basin with point and nonpoint sources. *JAWRA Journal of the American Water Resources Association* 37: 1169–1188.

Schilling K E, Wolter C F. 2009. Modeling nitrate-nitrogen load reduction strategies for the Des Moines River, Iowa using SWAT. *Environmental Management* 44: 671–682.

Schmalz B, Fohrer N. 2009. Comparing model sensitivities of different landscapes using the ecohydrological SWAT model. *Advances in Geosciences* 21: 91–98.

Schmalz B, Fohrer N. 2010. Ecohydrological research in the German lowland catchment Kielstau. *Status and Perspectives of Hydrology in Small Basins (Proceedings of the Workshop held at Goslar-Hahnenklee, Germany, 30 March - 2 April 2009), IAHS Publ.* 336.

Schmalz B, Springer P, Fohrer N. 2008a. Interactions between near-surface groundwater and surface water in a drained riparian wetland. In *Proceedings of International Union of Geodesy and Geophysics XXIV General Assemble" A New Focus on Integrated Analysis of Groundwater/Surface Water Systems", Perugia, Italy, 11-13 July 2007.*, 21–29. IAHS Press.

Schmalz B, Tavares F, Fohrer N. 2008b. Modelling hydrological processes in mesoscale lowland river basins with SWAT: capabilities and challenges. *Hydrological Sciences Journal* 53: 989–1000.

Smakhtin V U. 2001. Low flow hydrology: A review. *Journal of Hydrology* 240: 147–186.

Soetaert K, Petzoldt T. 2010. Inverse modelling, sensitivity and Monte Carlo analysis in R using package FME. *Journal of Statistical Software* 33: 1–28.

Staudinger M, Stahl K, Seibert J, Clark M P, Tallaksen L M. 2011. Comparison of hydrological model structures based on recession and low flow simulations. *Hydrology and Earth System Sciences* 15: 3447.

Strauch M, Lima J E F, Volk M, Lorz C, Makeschin F. 2013. The impact of Best Management Practices on simulated streamflow and sediment load in a Central Brazilian catchment. *Journal of Environmental Management* 127: 24–36.

Tattari S, Koskiaho J, Barlund I, Jaakkola E. 2009. Testing a river basin model with sensitivity analysis and autocalibration for an agricultural catchment in SW Finland. *Agricultural and Food Science* 18: 3–4.

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: 10–23.

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: 1154–1169.

Vogel R M, Fennessey N M. 1994. Flow-duration curves. I: New interpretation and confidence intervals. *Journal of Water Resources Planning and Management* 120: 485–504.

Volk M, Liersch S, Schmidt G. 2009. Towards the implementation of the European Water Framework Directive? Lessons learned from water quality simulations in an agricultural watershed. *Land Use Policy* 26: 580–588.

Wahnschaffe F, Schucht F. 1921. *Geologie und Oberflächengestaltung des norddeutschen Flachlandes.* Engelhorn.

Watson B M, Selvalingam S, Ghafouri M. 2003. Evaluation of SWAT for modelling the water balance of the Woady Yaloak River catchment, Victoria. *MODSIM 2003 : International Congress on Modelling and Simulation, Jupiters Hotel and Casino, 14-17 July 2003 : integrative modelling of biophysical, social and economic systems for resource management solutions : proceedings* 2003: 01–01.

Wittenberg H. 1999. Baseflow recession and recharge as nonlinear storage processes. *Hydrological Processes* 13: 715–726.

Wittenberg H. 2003. Effects of season and man-made changes on baseflow and flow recession: case studies. *Hydrological Processes* 17: 2113–2123.

Wu K, Johnston C A. 2007. Hydrologic response to climatic variability in a Great Lakes Watershed: A case study with the SWAT model. *Journal of Hydrology* 337: 187–199.

Yilmaz K K, Gupta H V, Wagener T. 2008. A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. *Water Resources Research* 44: W09417.

Zambrano-Bigiarini M. 2012. hydrotsm: Time series management, analysis and interpolation for hydrological modelling. *R package version 0.3-3* .

3 Smart low flow signature metrics for an improved overall performance evaluation of hydrological models

Pfannerstill, M., Guse, B., and Fohrer, N.: Smart low flow signature metrics for an improved overall performance evaluation of hydrological models, J. Hydrol., 510, 447-458, doi:10.1016/j.jhydrol.2013.12.044, 2014.

Received: 11 October 2013 - Accepted 26 December 2013

Abstract

Hydrological models have to be calibrated accurately to provide reasonable model results. For a concise model evaluation, the different phases of the hydrograph have to be considered in multi-metric frameworks with appropriate performance metrics. Low and high flows need to be reproduced simultaneously without neglecting the other phases of the hydrograph.
In this paper, we highlight the relevance of very low and low flows with separate performance metrics. We present a multi-metric evaluation framework to identify calibration runs, which represent the different phases of the hydrograph precisely. A stepwise evaluation was done with commonly used statistical performance metrics (Nash-Sutcliffe, percent bias) and signature metrics, which are based on the flow duration curve (FDC). In order to consider a fairly balanced evaluation between high and low flow phases, we divided the flow duration curve into segments of high, medium and low flow phases, and additionally into very high and very low flow phases. The model performance in these segments was evaluated separately with the root mean square error (RMSE).
Our results show that this evaluation method leads to an improved selection of good calibration runs to enhance the overall model performance by the refined segmentation of FDC. By combining performance metrics for high flow conditions with low flow conditions, this study demonstrates the challenge of calibrating a model with a satisfactory performance in high and low phases simultaneously. Consequently, we conclude that an additional performance metric for very low flows should be included in model analyses to improve the overall performance in all phases of the hydrograph.

3.1 Introduction

Hydrological models are used in practice and science to assess a wide range of hydrological problems such as climate and land use change or to predict extreme events in terms of flood and low flow events for river management (*Tallaksen et al.*, 1997; *Hunter et al.*, 2007; *Laaha and Blöschl*, 2007; *Thielen et al.*, 2009). Hydrological complexity is reflected in different phases within the discharge time series. A challenge of hydrological models is to adequately represent all phases with the same model parameter set (*Madsen*, 2000). To achieve a satisfying reproduction of the hydrological processes, hydrologic models have to be calibrated to the conditions of the study catchments. Generally, model parameters are calibrated for

specific catchment characteristics to the measured discharge time series. The most suitable parameters are selected with a sensitivity analysis (see *van Griensven et al.*, 2006) or on the user's experience respectively. The next step is the calibration of selected parameters with following evaluation of model results by visual inspection of the hydrograph fitting and the application of performance measures (e.g. *Moriasi et al.*, 2007).
The Nash-Sutcliffe efficiency (NSE, *Nash and Sutcliffe*, 1970), which is often used to evaluate simulation results in hydrology, is sensitive to differences in the observed and simulated means and variances (*Legates and McCabe*, 1999). However, this performance measure is more sensitive to extreme values (*Legates and McCabe*, 1999) and tends to neglect possible deviations in low flow periods as it is not very sensitive to systematic over- and underestimations of the model (*Krause et al.*, 2005). The root mean square error (RMSE) overemphasizes flood peaks and leads to a bad calibration of low flow periods (*Boyle et al.*, 2000; *Madsen*, 2000; *Bekele and Nicklow*, 2007). As a consequence, a better performance for high flows than for low flows may result in an underestimation in long dry periods (*De Vos et al.*, 2010). Furthermore, a good performance in some periods with high flows is able to dominate the global performance and masks the poor performance in other periods like low flow periods (*Zhang et al.*, 2011). Referring to high and low flow calibration, the application of one single criterion tends to measure the difference between the simulated and observed hydrographs by matching one aspect of the hydrograph at the expense of another (*Boyle et al.*, 2000; *Wagener et al.*, 2001). Furthermore, the application of one single performance measure is insufficient to take into account the representation of all relevant processes (*Gupta et al.*, 1998; *Wagener and Gupta*, 2005; *Gupta et al.*, 2008). This was also stated by *Madsen* (2000), who found no overall best performance measures during the calibration process. The reason for this shortcoming is the loss of valuable information by projecting from the high dimension of the data set down to the single dimension of the residual-based summary statistic (*Gupta et al.*, 2008; *Herbst and Casper*, 2008). Since matching of all parts of the hydrograph is favourable, a trade-off between different phases of the hydrograph has to be accepted. This trade-off effect can be minimized in multi-objective approaches with multiple performance measures, whose importance for discharge calibration was revealed in *Boyle et al.* (2000), *Bekele and Nicklow* (2007), *De Vos et al.* (2010), *Zhang et al.* (2011), and *Guse et al.* (2013).
To assess different phases of the hydrograph *van Werkhoven et al.* (2009) and *Zhang et al.* (2012) included statistical and hydrological metrics into the calibration process. They defined statistical metrics for the base and peak flow, hydrological metrics for the midrange flow and long-term water balance. The different parts of the hydrograph reflect different catchment functions (e.g. baseflow recession during dormant season of the vegetation) that can be captured in individual model components through parameter selection informed by careful hydrograph analysis (*Carrillo et al.*, 2011). The importance of different performance metrics was also mentioned in *van Werkhoven et al.* (2009), *Martinez and Gupta* (2011), and *Herman et al.* (2013), who proposed a multiple criteria application for diagnostic model analysis. In a diagnostic analysis on differing watershed behavior during rainfall and dry periods, the hydrograph can be separated into driven, non-driven quick, non-driven slow discharge as defined by *Boyle et al.* (2000). Also *Bekele and Nicklow* (2007) applied specific objective

functions to fit different portions of time series. *Madsen et al.* (2002) separated performance measures for high and low flows, which were only considered in periods above or below a threshold for high or low flows, respectively.
Yilmaz et al. (2008) used the overall water balance, vertical redistribution, temporal and spatial redistribution as signatures measures for major behavioural functions. Signature measures are defined as hydrologic response characteristics that provide insights into the hydrologic function of catchments (*Sawicz et al.*, 2011). *Pokhrel et al.* (2012) stated that several signature measures give a better overall representation of the hydrologic characteristics of the catchment. Both studies used the flow duration curve (FDC) to diagnose model performance for different flow characteristics of the catchment. There are several suggestions for splitting up the FDC into segments, which describe characteristic hydrological processes within the catchment (*Yilmaz et al.*, 2008; *Yokoo and Sivapalan*, 2011; *Cheng et al.*, 2012; *Pokhrel et al.*, 2012).
Dividing the flow duration curve into segments leads to a process-based calibration for the dominant processes within the catchment, which are reflected by the different parts of the hydrograph. However, *van Werkhoven et al.* (2009) see the limitations of the FDC to fully reflect the quality of simulations, since it includes no information on accurate flow timing. In contrast to time series, FDC indicates only that the right distribution of flow levels occurred throughout the record (*van Werkhoven et al.*, 2009). Thus, *van Werkhoven et al.* (2009) proposed a combination of statistical and signature metrics to capture the different parts of the hydrograph as well as their timing.
Dunn (1999) found high uncertainty for low flow predictions without specific consideration of a low flow criterion. Especially in lowlands, distinct low flow periods occur frequently but with high variability in the minimum discharge. In this case, it is uncertain, if the traditional segmentation (low flow: 70 % time flow equalled or exceeded) of the FDC is sufficient to calibrate low flow periods. For an adequate representation of the very low flow periods with respect to very high flows, additional segmentations could be an approach to calibrate a fairly balanced representation of extreme periods. In our investigations we took up these questions and focused on following topics:

- How can all phases of the hydrograph be combined in a multi-metric framework evaluation?
- Does a multi-metric framework detect calibration runs with a reasonable reproduction of all phases of the hydrograph?
- Does the additional segmentation of the FDC into low flow segments lead to an improved reproduction of low flows?

3.2 Materials and methods

3.2.1 Study area

Our investigations were carried out in the Kielstau catchment (50 km^2), which is located within a lowland area of the federal state Schleswig-Holstein in Northern Germany. The to-

pography ranges between 27 m and 78 m above mean sea level with a flat landscape, described by rolling hills and depressions. In the higher regions of the Kielstau catchment, Haplic and Stagnic Luvisols are the dominant soils. Along the stream and its tributaries primarily Sapric Histosols are found (*BGR*, 1999). As a consequence of this flat topography, the groundwater is a specific characteristic of this lowland catchment. *Schmalz et al.* (2008) describe the dynamics of the near-surface groundwater at a riparian wetland as a dynamic interaction between groundwater and surface water. The near-surface groundwater is generally controlled by precipitation and, close to the river, also by river water level (*Schmalz et al.*, 2008). Due to high water levels, a high fraction of approximately 38 % of the agricultural area is drained (*Fohrer et al.*, 2007). Further information about the catchments and results of investigations can be found in *Schmalz and Fohrer* (2009) and *Fohrer and Schmalz* (2012).
The hydrological characteristics are typical for a northern German lowland. The mean annual precipitation and temperature are 918,9 mm and 8,2° (*DWD*, 2012). The annual discharge is characterized by a mean outlet discharge at the gauging station Soltfeld of 0,42 $m^3 s^{-1}$, a mean low flow discharge of 0,05 $m^3 s^{-1}$ and a mean high flow discharge of 2,75 $m^3 s^{-1}$ (*LKN*, 2013). Referring to the seasonality of the discharge, high flow events take place from November to January (*LKN*, 2013). The lowest discharge is observed from June to the late August (*LKN*, 2013). For our study, we used the mean daily discharge of the gauging station Soltfeld from 1999 to 2010.

3.2.2 The SWAT model

For our multi-metric framework development, we used the Soil and Water Assessment Tool (SWAT2012; *Arnold et al.*, 1998). With this semi-distributed, eco-hydrological model, the discharge and the water cycle was simulated on a daily time-step for the Kielstau catchment. SWAT is a process-based conceptual model with abstracted, empirical components. These different components result in a very complex model with a high number of parameters (*Cibin et al.*, 2010). The model concept of SWAT divides the processes into a land and a water phase (*Neitsch et al.*, 2011). The water balance at the land phase is calculated by changes in soil water storages for each day, based on the calculation of the relevant processes. The main water input is the precipitation. To solve the water balance equation, the most important processes such as evaporation, runoff, soil water percolation and groundwater flow are considered. After calculating the water balance, the subbasins are connected in the water phase and the water is routed through the subbasins.
Several studies reveal that the SWAT model has limitations in simulating low flows (*Wu and Johnston*, 2007; *Eckhardt*, 2008; *Guse et al.*, 2013; *Koch et al.*, 2013; *Pfannerstill et al.*, 2013). The reason for poor model performance in low flow periods may be the usage of only one groundwater storage, which contributes to the channel (*Guse et al.*, 2013; *Pfannerstill et al.*, 2013). As we try to consider the performance measures especially for low flow periods, we applied the modified $SWAT_{3S}$ version (*Pfannerstill et al.*, 2013). $SWAT_{3S}$ is characterzied by a multi-storage groundwater concept, which emphasizes nonlinear groundwater dynamics. For this, an additional active groundwater storage was introduced into the groundwater

module of SWAT. The application of this version in the Kielstau catchment leads to substantial improvement of reproduction for the low flow periods by the model (*Pfannerstill et al.*, 2013). Further information about the concept and process equations are reported in *Pfannerstill et al.* (2013).

3.2.3 Model setup and calibration

The SWAT model input files were setup with the ArcSWAT interface (version 2012.10.1.6; *Winchell et al.*, 2010). For the watershed delineation, the Kielstau catchment was divided into 36 subbasins with 2214 HRUs, which were defined by using three slope classes ($< 2{,}6\,\%$, 2.6 - 4,6 % and $> 4{,}6\,\%$). Major input data was soil (1:200.000, *BGR*, 1999), a digital elevation model (5 m x 5 m cell size, *LVermA*, 1995) and the land-use data from 2012. The estimation of spatial distribution for drainage tiles within the catchment from *Fohrer et al.* (2007) was used to define drainage locations in the SWAT setup. The weather station Gluecksburg-Meierwik in the north of the Kielstau catchment was used for precipitation data. Measured climate data were used from the Statistical Regional model (STAR, *Orlowsky et al.*, 2008) such as wind speed, temperature, solar radiation and relative humidity as additional input. Because we applied the groundwater modified SWAT_{3S} version, the groundwater input files had to be reprocessed to add additional input parameters. A detailed description of the groundwater parameters for the additional groundwater process equations can be found in *Pfannerstill et al.* (2013).

The calibration parameters for $SWAT_{3S}$ were chosen after *Santhi et al.* (2001), *White and Chaubey* (2005), *van Griensven et al.* (2006), *Bärlund et al.* (2007), *Schmalz and Fohrer* (2009), *Tattari et al.* (2009) and based on further SWAT projects in the Kielstau catchment (*Kiesel et al.*, 2010; *Fohrer et al.*, 2013). The R-package FME (*Soetaert et al.*, 2010) was used to generate a set of variations for the calibration parameters with the Latin Hypercube Sampling of the R-package FME (*Soetaert et al.*, 2010) for 5000 model runs. The SWAT model input files were rewritten for each calibration run with varied parameter values from the calibration parameter set with the R environment (*R Core Team*, 2013) as already used in *Pfannerstill et al.* (2013) and *Guse et al.* (2013).

The discharge data was divided into a calibration and validation data set. A warm-up phase from 1996 to 1999 was used to achieve steady state conditions for the model. For calibration, the discharge data from 1999 to 2002 and from 2005 to 2008 was selected, and for validation we choose the discharge data from 2002 to 2005 and from 2008 to 2010. The selection of years for calibration and validation was based on the wetness, which was derived from precipitation for every year. With this selection, calibration and validation was characterized by equaled distribution of dry, regular and wet periods.

3.2.4 Model evaluation

3.2.4.1 Commonly used performance metrics

Generally, model simulations are evaluated by performance metrics, which can be divided into statistical metrics and signature metrics (cf. *Yilmaz et al.*, 2008; *van Werkhoven et al.*,

2009). *Moriasi et al.* (2007) cites possible statistical metrics to quantifiy the accuracy of simulated discharge and provided target ranges. Statistical metrics quantify the distance between the observed and simulated flow time series based on assumptions about the statistical characteristics of the model residuals, but do not necessarily indicate how well the hydrological function (e.g. water balance and flow regime) of the system is maintained by the model (*van Werkhoven et al.*, 2009). Tab. 5 and Tab. 6 give an overview for additional possible performance metrics, which were used in studies for model calibration. There are several commonly used statistical metrics to quantify the accuracy of high flow events and their timing (NSE, RMSE, MSE). In order to apply a metric which is less focused on high flows, the RSR is used additionally. The RSR standardizes the RMSE using the standard deviation of the overserved discharge (*Moriasi et al.*, 2007). Due to this scaling, the values of the RSR can be applied to various constituents (*Moriasi et al.*, 2007). Many indices are also used for calibration of low flow events. One possibility is to use the performance metrics for high flow events with a logarithmic transformation of the discharge data to emphasize low flow periods. The mean relative error (MRE) can be used to assess the model performance in the lower magnitudes of datasets (*Dawson et al.*, 2007). Beyond these classical statistical metrics *Yilmaz et al.* (2008) argue that it is also necessary and important to actually formulate quantitative representations in the form of signature metrics to summarize the relevant and useful diagnostic information present in the data. The FDC can be used to calibrate the discharge volume of different flow periods. *Yokoo and Sivapalan* (2011) divided the FDC into fast and slow flow segments which are governed by different components of climate and landscape. Further segmentation of the FDC was done by *Yilmaz et al.* (2008) and *Cheng et al.* (2012) into a fast flow segment, which is controlled by large precipitation events and a mid flow segment, which is controlled by moderate size precipitation events and also related to the intermediate-term primary and secondary base flow relaxation response of the watershed. The slow flow segment considers retention due to catchment storages segment and is controlled by catchment parameters (*Yilmaz et al.*, 2008; *Cheng et al.*, 2012). Also *Pokhrel et al.* (2012) divided the flow duration curve into high flow, mid flow and low flow segments.

Table 5: Application examples of commonly used performance metrics for evaluation of different phases of the hydrograph. Q_i is the observed value, $\widetilde{Q_i}$ is the modelled, $\bar{Q}$ is the mean of the observed data

Performance metric	Equation	Range	Sensitive hydrograph phase	Application examples
Nash-Sutcliffe efficiency	$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}(Q_i - \widetilde{Q}_i)^2}{\sum_{i=1}^{N}(Q_i - \bar{Q})^2}$	$-\infty - 1$	peaks and discharge dynamic	*Krause et al.* (2005); *Bekele and Nicklow* (2007); *Zhang et al.* (2011); *van Werkhoven et al.* (2009); *Yadav et al.* (2007)
root mean square error	$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widetilde{Q}_i - Q_i)^2}$	$0 - \infty$	flood peaks	*Madsen* (2000); *van Werkhoven et al.* (2009); *Bekele and Nicklow* (2007); *Yadav et al.* (2007); *Boyle et al.* (2000)
ratio of root mean square error and standard deviation	$\mathrm{RSR} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Q_i - \widetilde{Q}_i)^2} \Big/ \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Q_i - \bar{Q})^2}$	$0 - \infty$	flood peaks	*Moriasi et al.* (2007)
mean square error	$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(\widetilde{Q}_i - Q_i\right)^2$	$0 - \infty$	high flow	*Pokhrel et al.* (2012)
percent bias in FDC high-segment volume	$\mathrm{BiasFHV} = \frac{\sum_{i=1}^{N}\left(\widetilde{Q}_{high,i} - Q_{high,i}\right)\times 100}{\sum_{i=1}^{N} Q_{high,i}}$	$0 - \infty$	high flow volume	*Pokhrel et al.* (2012); *Yilmaz et al.* (2008)
slope flow duration curve	$\mathrm{SFDCE} = \frac{[log(\widetilde{Q}_{m1}) - log(\widetilde{Q}_{m2})] - [log(Q_{m1}) - log(Q_{m2})]}{[log(Q_{m1}) - log(Q_{m2})]} \times 100$	$-\infty - \infty$	flow from moderate size precipitation events	*Yilmaz et al.* (2008); *Pokhrel et al.* (2012)
log root mean square error	$\mathrm{logRMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(\frac{Q_i}{\widetilde{Q}_i}))^2}$	$0 - \infty$	emphasizing low flows with log of discharge	*Bekele and Nicklow* (2007)
mean square error of log discharge	$\mathrm{MSEL} = \frac{1}{N}\sum_{i=1}^{N}\left(\log 10(\widetilde{Q_i}) - \log 10(Q_i)\right)^2$	$0 - \infty$	low flow	*Pokhrel et al.* (2012)
transformed root mean square error	$\mathrm{TRMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\frac{(1+\widetilde{Q}_i)^{0.3}-1}{0.3} - \frac{(1+Q_i)^{0.3}-1}{0.3})^2}$	$0 - \infty$	low flow	*van Werkhoven et al.* (2009)

Table 6: Application examples of commonly used performance metrics for evaluation of different phases of the hydrograph. Q_i is the observed value, $\widetilde{Q_i}$ is the modelled, $\overline{Q}$ is the mean of the observed data

Performance metric	Equation	Range	Sensitive hydrograph phase	Application examples
percent bias in FDC low-segment volume	$\mathrm{BiasFLV} = \frac{\sum_{i=1}^{N}\left(\widetilde{Q}_{low,i} - Q_{low,i}\right)\times 100}{\sum_{i=1}^{N} Q_{low,i}}$	$-\infty - \infty$	low flow volume	*Pokhrel et al.* (2012); *Yilmaz et al.* (2008)
coefficient of determination	$\mathrm{r}^2 = \frac{\sum_{i=1}^{N}\left(\widetilde{Q}_i - \overline{Q}_i\right)^2}{\sum_{i=1}^{N}\left(Q_i - \overline{Q}_i\right)^2}$	$0 - 1$	discharge dynamics	*Krause et al.* (2005)
percent bias	$\mathrm{PBIAS} = \frac{\sum_{i=1}^{N}\left(\widetilde{Q}_i - Q_i\right)\times 100}{\sum_{i=1}^{N} Q_i}$	$-\infty - \infty$	average tendency of over- and underestimation	*Zhang et al.* (2011); *Gupta et al.* (1999)
runoff coefficient error	$\mathrm{RR} = \frac{\sum_{i=1}^{N}\left(\widetilde{Q}_i - Q_i\right)\times 100}{\sum_{i=1}^{N} Q_i}$	$-\infty - \infty$	overall water balance	*Pokhrel et al.* (2012); *van Werkhoven et al.* (2009); *Yilmaz et al.* (2008)
mean absolute error	$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left\lvert\widetilde{Q}_i - Q_i\right\rvert$	$0 - \infty$	overall discharge	*Lombardi et al.* (2012); *Gupta et al.* (1998)
mean relative error	$\mathrm{MRE} = \frac{1}{N}\sum_{i=1}^{N}\frac{\widetilde{Q}_i - Q_i}{Q_i}$	$-\infty - \infty$	overall discharge	*Dawson et al.* (2007)

3.2.4.2 Additional performance metrics

Since our investigations took place in a catchment with reocurring low flow periods, we focused on the development of additional performance metrics to account for low flow discharge. Although there are many common used performance metrics, there is still a limitation for the previously mentioned low flow metrics: these indices primarily use the whole discharge series for evaluation. For the diagnostic model analysis it is necessary to identify the discharge levels of poor model performance. This identification can be done by using discharge segments of the FDC. In our study, we used the exceedance probability of the FDC for segmentation. The exeedance probability is within the range of 0 to 100 %, where Q_{100} denotes an exceedance probability of 100 % for the discharge.

Yilmaz et al. (2008) mentioned that the metrics of the FDC are designed for calibrating the overall water balance without incorporating the timing of discharge events. They investigated the applicability of the FDC segments, especially for high flow events with a range of flow exceedance probability between 0 and 0,02 %. The low flows are considered within a range of flow exceedance probability between 70 and 100 % with logarithmic discharge volumes without further subdivision.

As a consequence, we designed an additional segmentation of the FDC. *Laaha and Blöschl* (2006) used the flow exceedance probability of 95 % (Q_{95}) for regionalizing low flows of Austria, because the Q_{95} is a low flow characteristic which is widely used due to its relevance for multiple topics of water resources management (*Gustard et al.*, 1992; *Smakhtin*, 2001; *Gustard and Demuth*, 2009). Consequently, the FDC was additionally segmented by the Q_{95} probability. To obtain a fair balance between very high flows and very low flows, the very high flow segment was shifted from the Q_2 to the Q_5. With this shift, the very low and the very high flows are segmented in the flow duration curve in equal ranges. To account for the high flow and low flow periods, two additinal flow ranges were integrated. The high flow range was defined between Q_5 and Q_{20}, and the low flow range between Q_{70} and Q_{95} (Fig. 10a). These 5 FDC metrics (hereafter 5FDC) were used to determine the effect of the newly introduced FDC ranges by comparing the selection of best calibration with 4 FDC metrics (herafter 4FDC) The 4FDC approach uses the low flow segmentation which is based on *Yilmaz et al.* (2008) with a range from Q_{70} to Q_{100} (low flow, Fig. 10b). The performance of the model runs within these FDC segments were analyzed with the RMSE, as this performance measure is due to its quadratic character strongly sensitive to extreme positive and negative error values (Tab. 7, Fig. 10a). Consequently, the RMSE is used to identify poor model performance due to sensitivity for highest under- and overestimation of simulated discharge volume for each FDC segment. The model performance of the different FDC segments for the best calibration runs was compared with the relative metric RSR.

Table 7: Newly defined performance metrics for five segment (5FDC) and four segment (4FDC) model evaluation of different phases of the hydrograph and performance metrics for evaluation of control segments for newly introduced FDC segmentation

Performance metric		5FDC	4FDC	Sensitive hydrograph phase
RMSE_Q5	RMSE in FDC Q_5 very high segment volume	x		very high flow discharge volume
RMSE_Q20	RMSE in FDC between Q_5 and Q_{20} high segment volume	x		high flow discharge volume
RMSE_mid	RMSE in FDC between Q_{20} and Q_{70} mid segment volume	x		mid flow discharge volume
RMSE_Q70	RMSE in FDC between Q_{70} and Q_{95} low segment volume	x		low flow discharge volume
RMSE_Q95	RMSE in FDC Q_{95} low segment volume	x		very low flow discharge volume
RMSE_low	RMSE in FDC between Q_{70} and Q_{100} overall low segment volume		x	overall low flow discharge volume

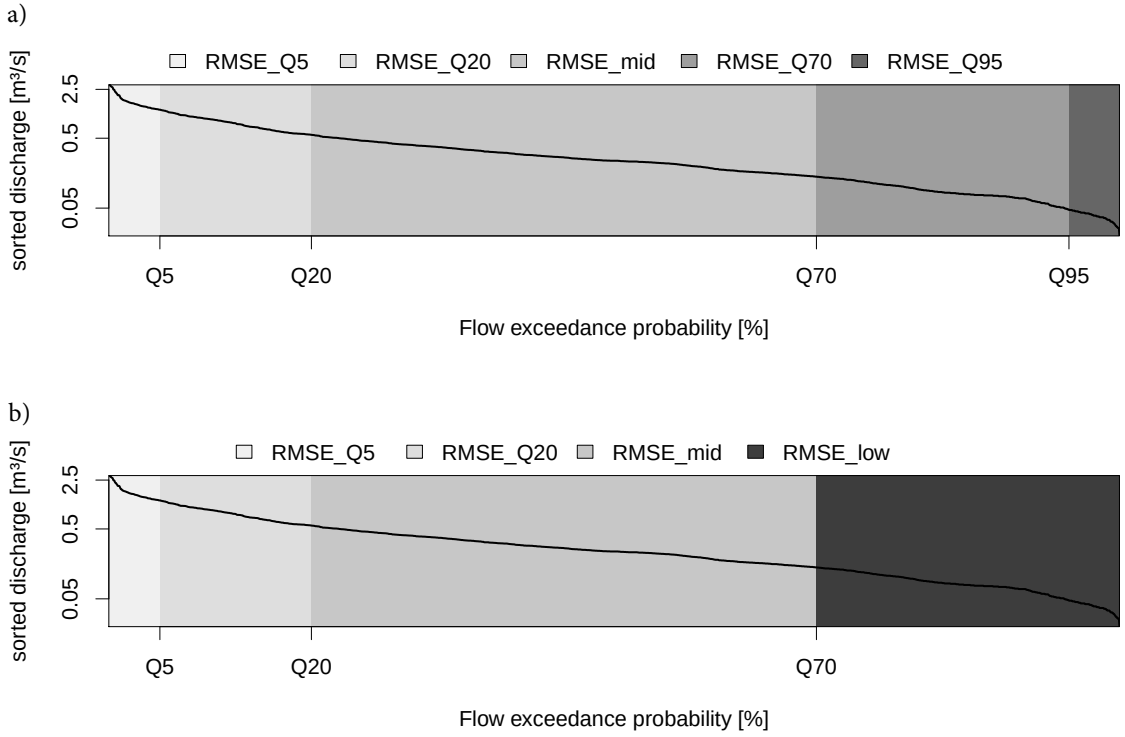

Figure 10: Segments of the FDC for the newly defined performance metrics for evaluation of different phases of the hydrograph (a, 5FDC) and the additional control segment (b, 4FDC).

3.2.5 Multi-metric evaluation for calibration

The model evaluation of the calibration runs was performed with the R-package hydroGOF (*Zambrano-Bigiarini* (2012)) and a simple reproduction of the FDC (*Vogel and Fennessey*, 1994; *Smakhtin*, 2001). Based on the approach of *Pokhrel et al.* (2012), we determined adequate discharge simulation results by a stepwise evaluation with several performance metrics. This evaluation was applied for two different performance metric combinations with different segmentations in the low flow segment of the FDC.

3.2.5.1 Five segment evaluation (5FDC)

In the following, we present exemplarily the proposed methodical concept for the evaluation with five segments of the FDC (hereafter 5FDC) as shown in Fig. 11 and described in detail as follows:

In the first step, a ranking from best performance metric to worst performance metric value was calculated for the NSE, RMSE_Q5, RMSE_Q20, RMSE_mid, RMSE_Q70 and RMSE_Q95 (Fig. 11). The best performance for NSE is 1 and for all RMSE the best performance is 0.

In the second step, a value above the threshold defined by the 1000 best model runs was applied to select the best simulation runs for each performance metric independently. These selections were plotted with the NSE against PBIAS for every performance metric. Afterwards, these selections were intersected with each other to identify the simulation runs with the best

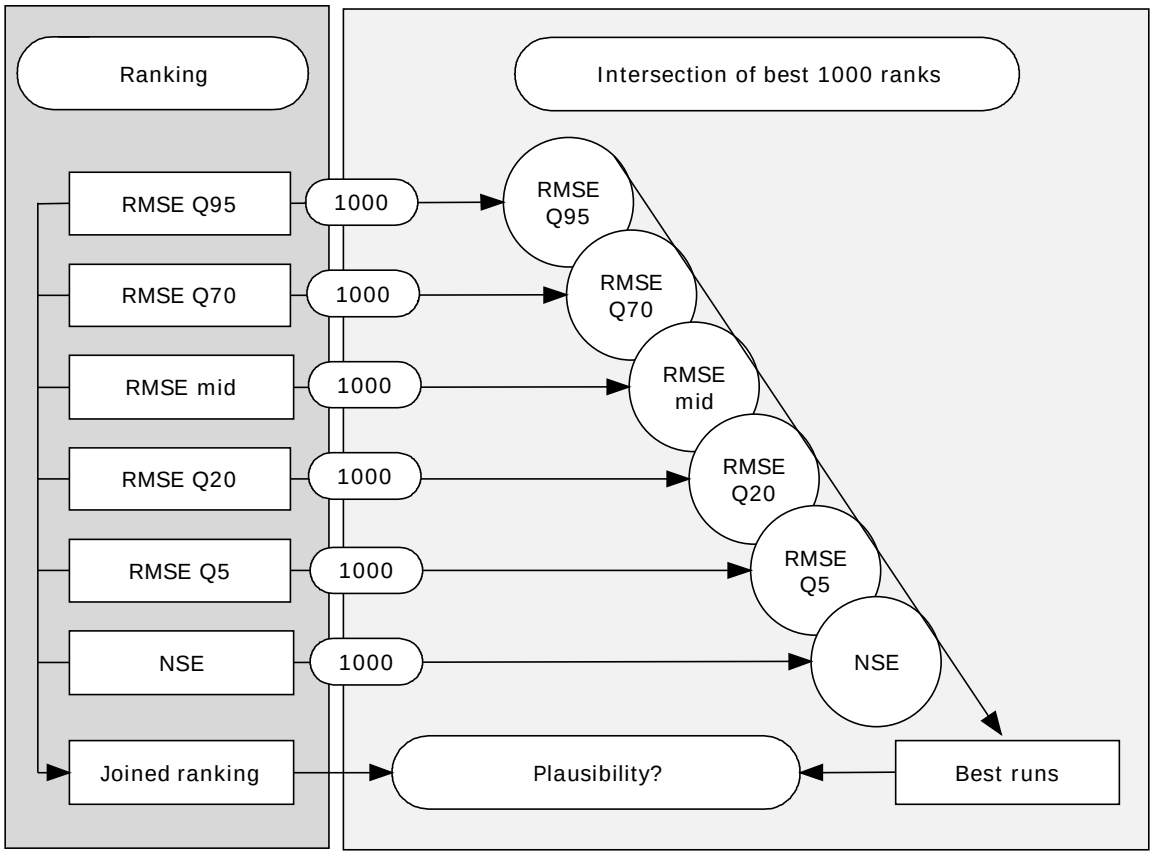

Figure 11: Description of the model evaluation procedure to select best calibration runs. The first column depicts the ranking for each of the seven performance metrics for the whole dataset. For each simulation run, the different ranking values were summed to a joined ranking to perform a plausibility check after the selection procedure. In a second step, simulation runs with the best 1000 ranking for each performance metric were selected. The selected best calibration runs for each performance metrics were intersected with each other to obtain a final selection of the overall best simulation runs. The plausibility check clarifies, if the selected best model runs have the best joined ranking values.

combination, where all performance metrics have a value above the threshold as determined by the 1000 best simulation runs. For every intersection, the simulation runs were plotted with the NSE against PBIAS. These plots were used to analyze the distribution of the model performance for the selected simulation runs. To optimize the visualisation, all calibration runs with NSE lower than 0 were excluded from the data set.

In the third step, we summed up the different ranking values of the performance metrics for each calibration run to obtain a joined ranking. The previously identified best calibration runs were used for a plausibilty check by analyzing the joined ranking values of the selected best simulation runs. For a consistent model evaluation, the intersection of all selections of the different performance metrics should result in a small joined ranking value but also in a combination of optimum values for all performance metrics.

3.2.5.2 Four segment evaluation (4FDC)

To investigate whether the additional segmentation for low flows (RMSE_Q95) improves the reproduction of this low flow period, we repeated the described multi-metric evaluation

with four FDC segments (4FDC). In comparison to the 5FDC approach, we used the definition of the low-flow segment in accordance to *Yilmaz et al.* (2008). For this, the additional segmentation in the low flow segment was neglected. Thus, we repeated the evaluation in the same way as in the 5FDC approach but with the exception that RMSE_low was used instead of RMSE_Q70 and RMSE_Q95. The maximum ranking number of 600 was applied to achieve a selection of best calibration runs for the model evaluation without the additional segmentation of the low flow segment.

3.2.6 Multi-metric evaluation for validation

To evaluate the model performance for the calibration with and without additional very low segments of the validation period, we applied the same performance metrics which were used for the calibration evaluation. The selection of the best model run for validation was based on these performance metrics and the visual inspection of the modelled hydrograph in comparison to the observed discharge. Furthermore, we analyzed the FDC of the best calibration runs visually.

3.3 Results

First we show the results of the model evaluation procedure for the newly introduced metrics of the FDC segments (5FDC). In the first step, the multi-metric evaluation for calibration revealed the best 1000 calibration runs for each performance metric. In the second step, these selected calibration runs were intersected to identify a small group of best calibration runs. Thus, this group of runs includes only calibration runs which are among the best 1000 runs for all considered performance metrics. Secondly, we present the results which were achieved from the application of the 4FDC model evaluation procedure by neglecting the very low segment of the FDC.

3.3.1 Calibration

3.3.1.1 Performance metric based selection

The results of the 5FDC performance metric selection are shown in Fig. 12. For each performance metric, the best calibration runs are identified by the ranking threshold of 1000 (Fig. 12, second column). In Fig. 12 (second column), 5000 model runs are shown as relationship between NSE and PBIAS. Positive values of the PBIAS reflect an underestimation. All runs which are among the 1000 best model runs are highlighted in black. Since this filter was applied for each performanc metric independently, there are different selections of identified best 1000 calibration runs (Fig. 12, third column).

Applying the RMSE_Q95, the selected simulation runs tend to lower PBIAS less than $-5\,\%$ and a NSE to between 0.03 and 0.7. This selection is similar to the RMSE_Q70, which differs slightly due to rejection of simulation runs with high NSE combined with low PBIAS values. The RMSE_mid selection rejects all simulation runs with PBIAS lower than $-20\,\%$ by allowing the whole range of NSE. When applying the RMSE_Q20, the selection is focused on runs with a high NSE and a small PBIAS. This tendency is also reflected with the

RMSE_Q5, which selects simulation runs with high NSE values but with a sharp frontier at an NSE of 0.4 together with simulation runs less than 0 %. The NSE enhances this tendency, as a frontier at an NSE of 0.6 is also present. In contrast to the RMSE_Q5 metric, the NSE allows also simulation runs with a positive PBIAS.

3.3.1.2 Stepwise intersection of best 1000 calibration runs

After the application of the best 1000 runs threshold on each performance metric, the single selections were intersected with each other stepwise (Fig. 12, second column). The intersection of the RMSE_Q95 with the RMSE_Q70 identified calibration runs within a PBIAS range of 0 and −40 % and a range of NSE between 0.25 and 0.7. The amount of acceptable performing model runs was remarkably reduced from 1000 to 266. It is remarkable that only 266 runs are included in the best simulations runs after the intersection of the two low flow metrics (RMSE_Q95, RMSE_Q70). The further intersection with RMSE_mid and RMSE_Q20 reduced the density of simulation runs to 57. With this selection, simulation runs with high NSE values and and negative PBIAS values were strongly rejected. The impact of the RMSE_mid and RMSE_Q20 on the reduced selection was nearly equal (RMSE_mid: n=105, RMSE_Q20: n=57). Intersecting the selection with the RMSE_Q5 resulted in the further reduction due to rejection of simulation runs (n=20) in which NSE are mainly below a value of 0.6. Consequently, the major difference between the last selection and the intersections with RMSE_Q5 is the removal of runs with poor performance in relation to the NSE as expected. This threshold was also important for the last intersection with the NSE selection, since all model runs below a NSE of 0.6 were fully rejected (Fig. 12). However, this additional intersection with the NSE does not lead to a strong reduction of n (20 to 11) but to a cluster of favourable simulation runs with similar NSE and PBIAS.

3.3.1.3 Joined ranking

The joined ranking, which is the sum of the different ranking values of the performance metrics for each calibration run, was used as a plausibility check for intersection results. The first intersection with RMSE_Q70 reduced the selection of simulation runs remarkably (ranking: 1 - 3366). The application of RMSE_mid reduced the range of the joined ranking values again (1 - 1853). The further application of RMSE_Q20 had less effect on joined ranking values (1-1418). The most important influence on the joined ranking was revealed by the RMSE_Q5 and NSE, as the ranking values range were reduced remarkably (RMSE_Q5: 1 - 138, NSE: 1 - 28).

3.3.1.4 Best calibration runs

The stepwise intersection of the best 1000 selections resulted in a small group of best calibration runs (Tab. 8). For plausibility check of the described selection, the joined ranking values are listed with the performance metrics. The joined rankings are within a range of 1 to 28. The NSE values are very similar and the PBIAS, are between -8.7 and 4,4 %. The variations of the RMSE within each segment are small. Highest RSR values were found in

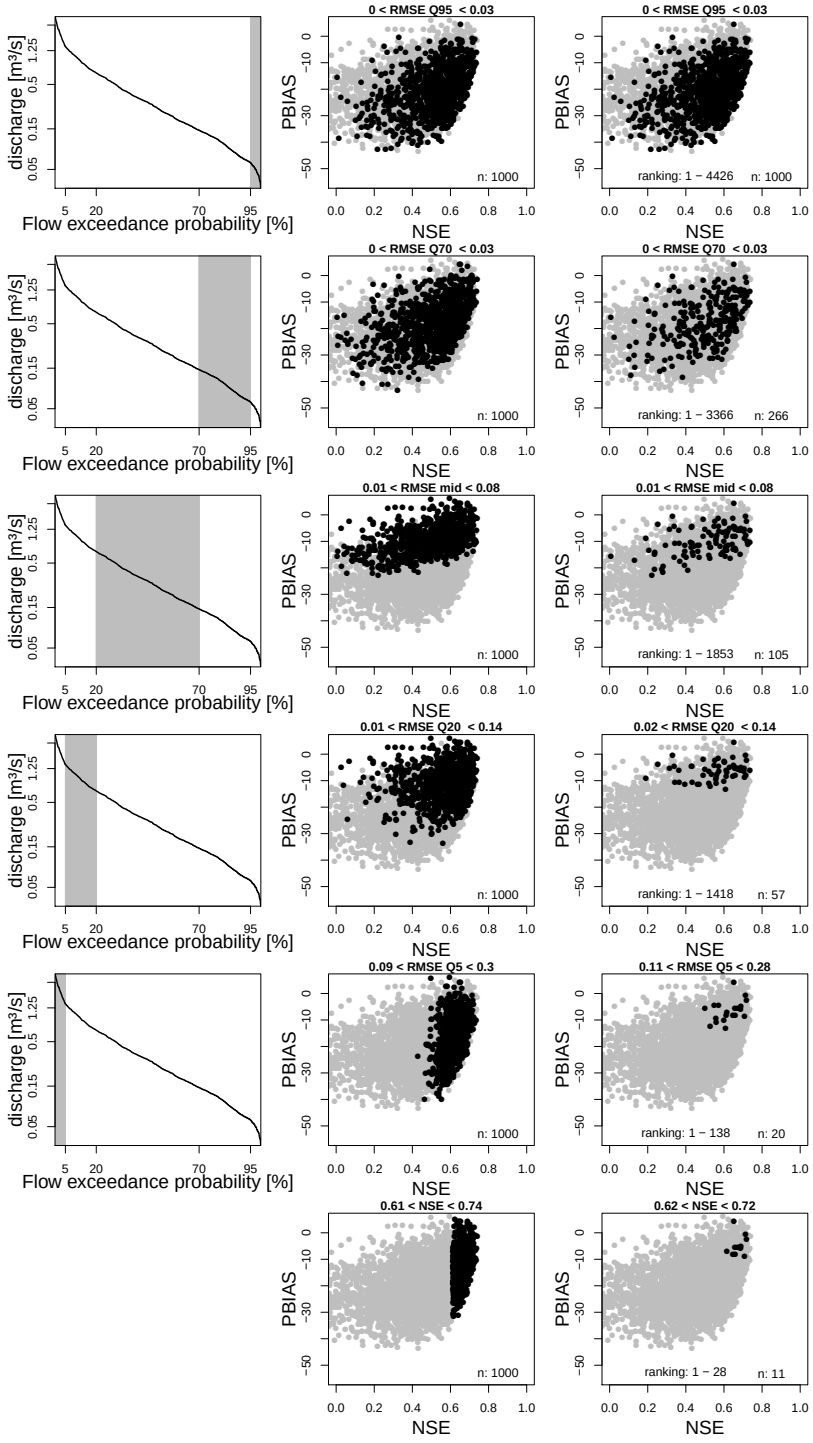

Figure 12: Stepwise evaluation of discharge calibration results. The first column shows the relevant discharge phases of each performance metric of the RMSE_Q5, RMSE_Q20, RMSE_mid, RMSE_Q70 and RMSE_Q95. The second column shows the best 1000 calibration runs for each performance metric (black points). The gray points indicate the complete dataset of 5000 simulation runs. The third column shows the reduction of best model runs by the amount (n) of remaining model runs (black) for each intersection with the RMSE_Q5, RMSE_Q20, RMSE_mid, RMSE_Q70 and RMSE_Q95 and the NSE together with the range of the joined ranking values for the selected model runs.

Table 8: Final selection of best calibration runs after applying 5FDC performance metric based selection of best 1000 calibration runs with subsequent stepwise intersection of best 1000 selections. The maximum value for NSE and the smallest PBIAS, RMSE of all best calibration runs are marked in gray. The RSR values of each FDC segment are integrated for comparison of segment performance

calibration	joined	NSE	PBIAS	RMSE					RSR				
run	ranking			Q5	Q20	mid	Q70	Q95	Q5	Q20	mid	Q70	Q95
140	28	0.71	-8.7	0.245	0.105	0.066	0.018	0.032	0.510	0.514	0.432	0.679	4.122
147	5	0.65	4.4	0.228	0.023	0.024	0.004	0.030	0.474	0.110	0.158	0.163	3.791
166	1	0.71	-0.6	0.207	0.075	0.031	0.006	0.034	0.430	0.365	0.204	0.215	4.322
211	7	0.72	-2.4	0.248	0.065	0.032	0.011	0.033	0.515	0.315	0.213	0.440	4.257
283	2	0.69	-5.9	0.216	0.142	0.038	0.011	0.031	0.450	0.694	0.249	0.423	3.992
1925	12	0.65	-8.2	0.108	0.080	0.043	0.023	0.020	0.225	0.389	0.284	0.869	2.524
2030	3	0.62	-7.0	0.159	0.109	0.039	0.011	0.018	0.331	0.530	0.257	0.409	2.294
2571	20	0.67	-5.4	0.272	0.048	0.036	0.023	0.017	0.564	0.233	0.239	0.890	2.225
2791	13	0.69	-5.0	0.240	0.061	0.042	0.017	0.030	0.499	0.296	0.276	0.635	3.842
2964	8	0.66	-5.4	0.255	0.090	0.031	0.008	0.026	0.530	0.440	0.207	0.320	3.284
4378	25	0.66	-8.1	0.238	0.108	0.044	0.022	0.026	0.495	0.528	0.291	0.831	3.374

the RSR_Q95 segment. For the RSR_Q5 and the RSR_Q20, the metric values are similar. The lowest values were identified in the RSR_mid.

For visualizing the general performance of the best calibration runs, the simulated discharges of Tab. 8 were compared with the observed discharge (Fig. 13). The overlay of the selected calibration runs resulted in small discharge band, because all of the simulated discharges are similar. All simulation runs tend to overpredict high peak events of the hydrograph from 1999 to 2002. Generally, the low flow and recession events are predicted satisfactorily except for recession in May 2001. Furthermore, Fig. 13 shows the distribution of phases, where the different performance metrics of the flow duration curve are relevant, as bands with different shades of gray. Comparing the very high (RMSE_Q5) and the very low phases (RMSE_Q95) periods, the very low flow periods occurred less often, but the duration was much longer than for the very high flow events. Comparing the low (RMSE_Q70) and the high flow events (RMSE_Q20), the same tendency was found. High flow phases occured more often than the low flow phases, but the duration was longer for low flow events. Furthermore, the five patterns of the FDC segments emphasize wet and dry years within the calibration period. The Q95 occurs in summer and autumn. Long periods of Q95 are observed in 2000 and 2008, while the Q95 was not reached from September 2006 to June 2008. The pattern of the Q95 cluster clearly shows that the Q95 segment considers the low flows occuring after long dry phases, as shown in summer 2000. Remarkably, there are phases in which the low flow does not fall below the very low flow threshold (September 2006 to June 2008), but the high flow exceeds the very high flow threshold. The opposite can be seen for April 2000 to November 2001, where very high flow occurs, but high flow does not exceed the very high flow threshold.

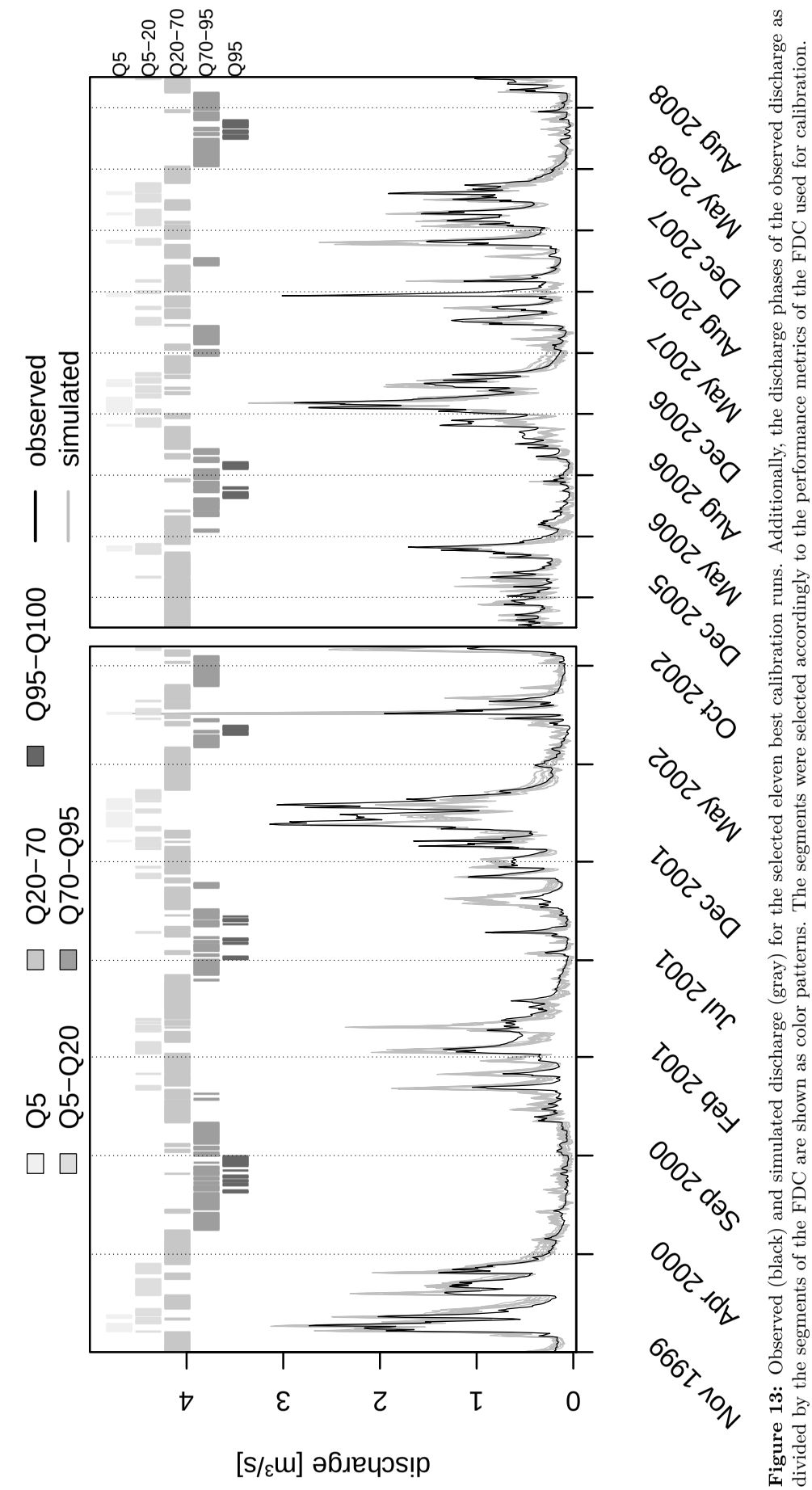

Figure 13: Observed (black) and simulated discharge (gray) for the selected eleven best calibration runs. Additionally, the discharge phases of the observed discharge as divided by the segments of the FDC are shown as color patterns. The segments were selected accordingly to the performance metrics of the FDC used for calibration.

The discharge volume distribution of the selected calibration runs was used to compare the overall volume based performance (Fig. 14a). Referring to the extreme events, the very high flow discharge is well presented by all selected simulation runs. The very low flow discharge shows an underestimation, with the best model run reaching only 98 % of flow exceedence probability (Fig. 14b). In the high and low flow phases, the discharge distribution tends to overpredict. The mid flow segment of the flow duration curve shows an underestimation in the part of 50 % flow exceedance probability.

3.3.1.5 Comparison of 5FDC and 4FDC segment evaluation

The stepwise intersection without additional segmentation in the low flow segment of the FDC (4FDC) revealed a selection of twelve best simulation runs. These simulation runs were compared with the twelve best simulation runs from the model evalutation framework with the additional RMSE_Q95 low flow metric (5FDC, Fig. 14a,c). To highlight the difference in the very low flow segment Q95, we illustrated the best simulation runs, which were identified with the RMSE_low (Fig. 14b,d).
The overall performance for both FDC ensembles is very similar, except for the Q95 segement of the FDC. This can be seen especially in the direct comparison of Fig. 14b and Fig. 14d. The stepwise intersection without the RMSE_Q95 metric revealed mainly simulation runs with a discharge simulation ending close at the Q95 point of the FDC. In contrast, the model evaluation with the RMSE_Q95 metric selected simualtion runs, which predict discharge until the Q100 probability.

3.3.2 Validation of 5FDC and 4FDC

For the validation of the 5FDC segmentation, we selected calibration run 2571. As we focus on the integration of low flow metrics to improve overall model performance, we selected the calibration run with the smallest Q95. Furthermore, our selection is based on visual agreement between oberserved and modelled discharge for the hydrograph and the FDC. Referring to low flow performance of the model, the hydrograph for the validation periods reveals underestimations from Februar 2004 until July 2004, as well as for May 2005 (Fig. 15). In contrast, the model overestimates the discharge in September till November 2003. The general dynamic of the discharge is well reproduced.

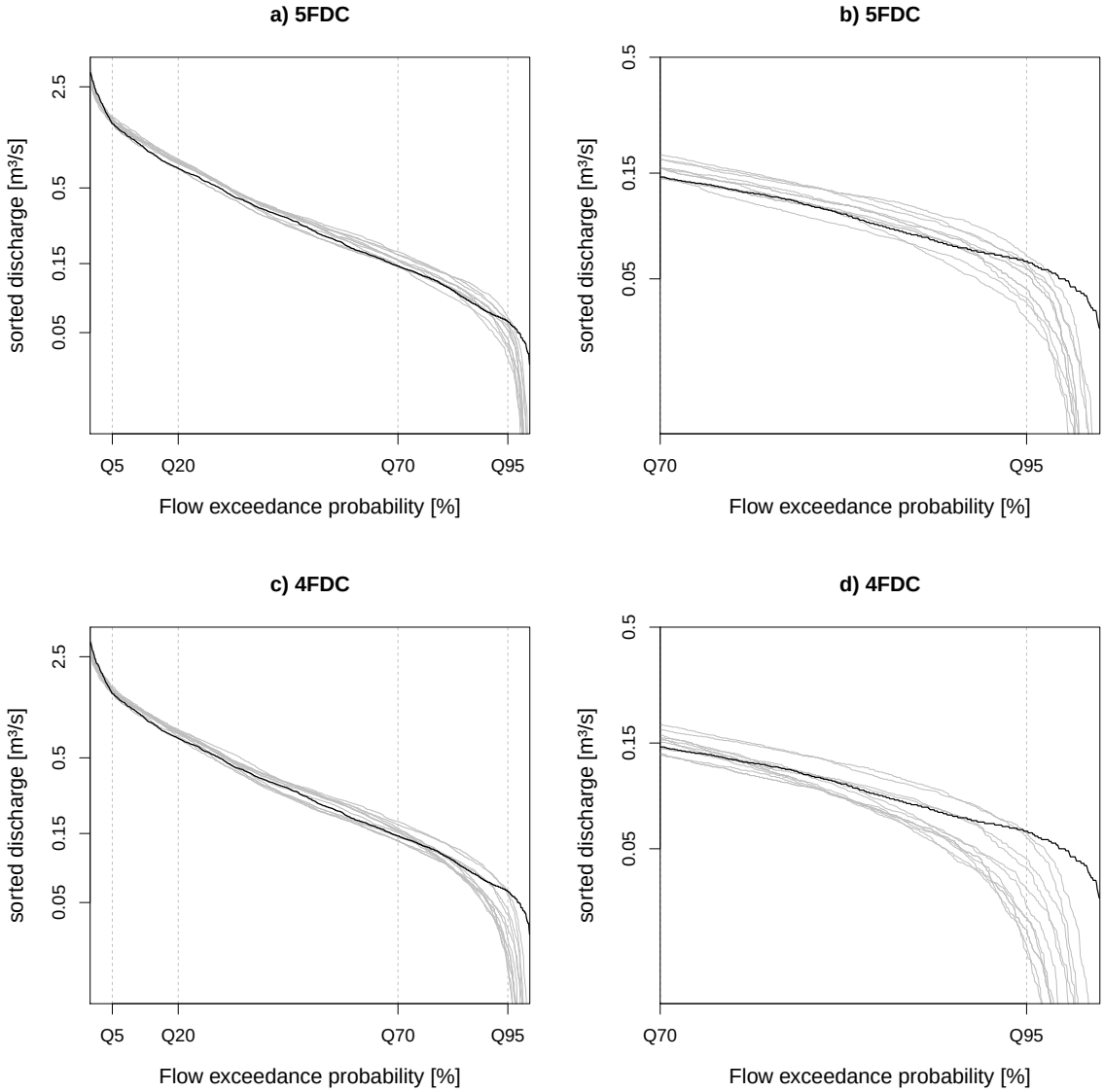

Figure 14: Flow duration curve for the observed discharge (black) and the selected best calibration runs (gray) of the evaluation framework with additional 5FDC segmentation (a) and without additional low flow segment of 4FDC (c). The low flow segment of the FDC is shown for the framework with additional segmentation in the low flow phase (b) and without application (d).

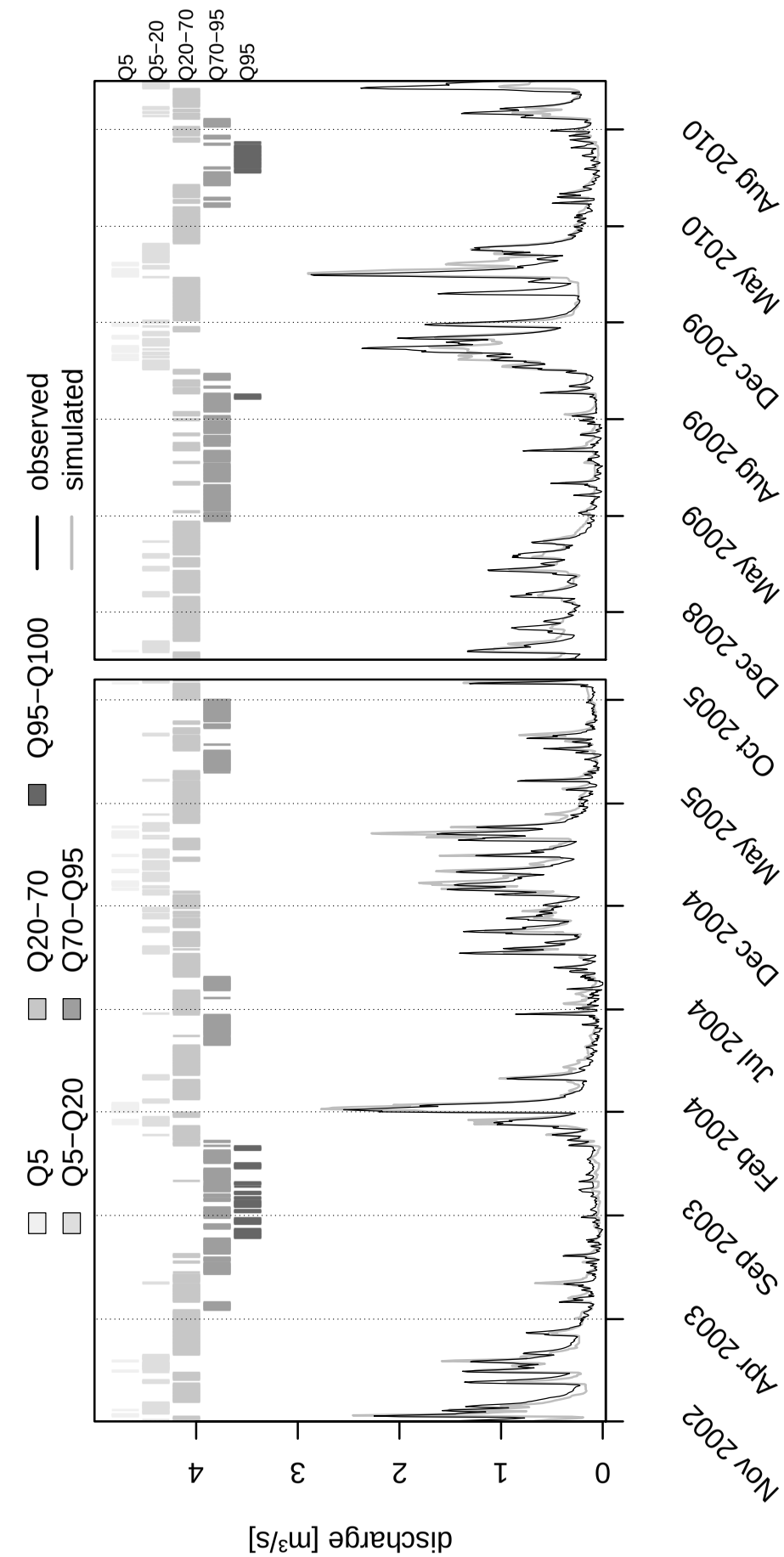

Figure 15: Observed (gray) and simulated discharge (black) for the validation of model run no. 2571. Additionally, the discharge phases of the observed discharge as divided by the segments of the FDC are shown as color patterns. The segments were selected according to the performance metrics of the FDC used for calibration.

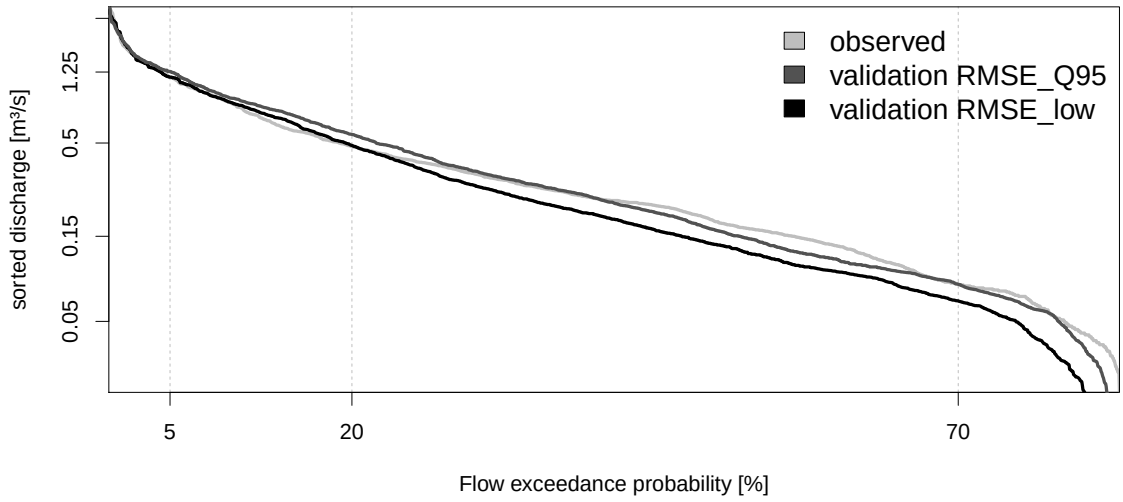

Figure 16: Flow duration curve for the observed discharge (gray) and the validation run with the RMSE_Q95 (dark gray) together with the validation run with the RMSE_low (black).

The reproduction of the discharge volume was also analyzed by the FDC (Fig. 16). The FDC shows that the model performs markedly well in the very high flow segment. In the high flow segment, the discharge volume is underestimated. The mid flow segment shows an overestimation for the flows of high exceedance probability and an underestimation for the flows of low exceedance probability. The low and the very low flow segments show an underestimation, especially at 98 % of flow exceedance probability. In comparison, the application of the RMSE_low metric revealed poor model performance especially in the mid and low flow phases (Fig. 16). The validation of the best simlation run with the RMSE_low evaluation (4FDC) underestimates the discharge with increasing exceedance probability of the FDC.

The model performance for the validation period is also reflected in the performance metrics (Tab. 9). For the evaluation of the simulation run, which was selected by the 5FDC framework with additional FDC segmentation, the NSE and PBIAS indicate good model performance. In comparison to the calibration periods, the model performed better for all FDC metrics. The performance of the simulation run, which was evaluated by the 4FDC method with the RMSE_low metric, shows better performance for the NSE and the high flow segments of the FDC. To identify the performance in the low flow segment, we additionally applicated the RMSE_Q95. It becomes apparent, that the model performed poorer in low flow phases than the model which was identified with the RMSE_Q95 metric. Furthermore, the validation run of the 5FDC procedure revealed better performance for the RMSE_low segment than the validation run of the 4FDC procedure.

Table 9: Validation results with performance metric values of the multi-metric framework for the best calibration run of the 5FDC and 4FDC evaluation procedure.

Calibration run	Evaluation procedure	NSE	PBIAS	RMSE Q5	RMSE Q20	RMSE mid	RMSE Q70	RMSE Q95	RMSE low
2571	5FDC	0.72	4.4	0.090	0.092	0.037	0.011	0.013	0.011
147	4FDC	0.75	6.0	0.069	0.036	0.044	0.024	0.026	0.025

3.4 Discussion

Based on our stated main research objectives, the discussion focuses firstly on the combined multi-metric evaluation framework. We secondly refer to the different performance metrics to quantifiy the overall model performance and finally discuss the performance of certain discharge phases of the hydrograph, especially in low flow phases.

In our study, we showed the application of an evaluation framework, which integrates several performance metrics to achieve a satisfying reproduction of the overall discharge. The joined ranking method revealed the plausibility of the multi-metric evaluation. Theoretically, the joined ranking allows the combination of a poor performed flow segment with a satisfying flow segment due to the independence of all performance metrics. Nonetheless, the multi-metric based identification of good simulation results are all characterized by a preferable joined ranking index. The coincidence in the selected model runs of the ranking based multi-metric evaluation framework with the joined ranking supports our methodical approach.

Referring to the threshold value to select the best model runs for each performance metric, the value of 1000 was found to be preferable for our study. However, the application of the evaluation framework without the additional low flow segmentation needed a reduction of this threshold to identify an manageable set of simulation runs for further analysis. The threshold for selecting best model runs of each performance metric depends on the number of applicated performance metrics and the number of valuable calibration runs.

Furthermore, the stepwise intersection revealed that all applied performance metrics were relevant for the calibration process, since every performance metric is characterized by specific distribution patterns and a specific rejection of simulation runs with poor performance as shown in Fig 12. We conclude, that the developed multi-metric framework is a helpful tool to identify model runs with satisfying performance in all phases of the hydrograph.

Referring to the applied performance metrics, every metric influenced the final selection of the best calibration runs. As mentioned in *Boyle et al.* (2000), *Krause et al.* (2005), *Bekele and Nicklow* (2007), *Moriasi et al.* (2007),*De Vos et al.* (2010) and *Zhang et al.* (2011), the importance of multi-objective discharge calibration with multiple performance metrics is a key to accounting for different discharge events. Our investigations support these findings, as the independent ranking of several performance metrics, followed by a threshold selection, showed characteristic distribution patterns by combining the NSE with PBIAS. Furthermore, every metric showed an characteristic impact on the final calibration run selection. The very high flow metric (RMSE_Q5) selected only simulation runs with high NSE values together with high discharge overestimation as reflected in negative PBIAS values. The high flow segment (RMSE_Q20) showed the same tendency, but allowed slight

discharge underestimation without high overestimation. The mid flow segment (RMSE_mid) emphasized simulation runs with less discharge overestimation. In contrast, the very low (RMSE_Q95) and low flow metric (RMSE_Q70) allowed calibration runs with large overall volume overestimation (PBIAS of more than −40 %), but rejected model runs with less overestimation. The focus of the different performance metrics reveals the importance of the multi-metric evaluation, which is also shown by *Madsen* (2000) and *van Werkhoven et al.* (2009). The application of only one or two performance metrics to find the best model run would likely result in emphasizing single discharge periods by neglecting other important discharge periods.

The necessity of several performance metrics was especially shown by the application of the overall low flow metric (RMSE_Q70) within the 4FDC evaluation, which selected simulation runs with high underestimation in the Q95 segment of the FDC. As a consequence, the application of the additional FDC segments as signature metrics as shown in the 5FDC approach is a helpful strategy to achieve a good model performance for the whole discharge period including verly low flows. As it requires minimal effort for a more accurate result, we propose this method to be integrated into model performance analysis. Our analysis clearly shows which periods are emphasized by the different performance measures. Furthermore, the demand for the use of separate performance measures for extreme high and low flows (Q5, Q95) is highlighted.

Owing to the characteristic focus of the different performance metrics, the selection of best model runs were consequently different. The stepwise intersection of these metric based ranking groups resulted in a decreasing number of possible good model runs. There is no solution with the highest possible NSE and the lowest PBIAS, because the intersection attempts to find the best solution by optimizing each performance metric equally. It is a trade-off between the best single performance of a measure and the other remaining performance metrics. The RMSE_Q5 tends toward higher PBIAS values together with high NSE. In contrast the RMSE_Q95 tends toward lower PBIAS values and a wide range of NSE.

Although our performance metrics are all in a satisfying range, this finding is an indicator for a deficit in the model structure. In our case, the very low and the very high metrics could be used as tool for model diagnostic (*Gupta et al.*, 2008; *Yilmaz et al.*, 2008; *Pokhrel et al.*, 2012). Analyzing the NSE and the PBIAS as indicator for model performance, both flow segments show different tendencies in the distribution of satisfying model performances. Similar differences in distribution patterns could be found for the low, mid, and high flow segments. The RMSE_mid rejects simulation runs, which have a PBIAS lower than −30 %. In contrast, the RMSE_Q70 and RMSE_Q20 allow PBIAS values much lower than −30 %, but the RMSE_Q20 rejects several model runs with a NSE less than 0.4.

The visual inspection of the best calibration runs gave additional information about the overall performance in discharge prediction. The calibrated models showed deficits, especially in the simulation of peaks events. While the discharge series are mainly preferable for evaluating the timing performance, the FDC gives the possibility to analyse the deficits in volume prediction. *Yilmaz et al.* (2008) already showed that the use of a close FDC segment for the very high flow events is helpful for the calibration of peak events. With the extended

segmentation of the FDC in the low flow, the discharge volume analysis revealed the largest errors in the very low flow segment. By comparing the 4FDC and the 5FDC evaluation approach, it became apparent that the application of two low flow segments (RMSE_Q70, RMSE_Q95) is preferable to control the model behaviour in low flow phases. This emphasizes the importance for the segmentation of the FDC into two low flow segments. Based on our additional segmentation of the FDC, the very low and low performance metrics should be used to control the baseflow governing process equations for the improvement of low flow simulation. Referring to processed based evaluation of model behaviour (*Yilmaz et al.*, 2008; *Pokhrel et al.*, 2012), we recommend the incorporation of the very low flow segment into the discussion for an improved model structure analysis.

3.5 Conclusion

In this study, we present a multi-metric framework to improve the overall model evaluation. To achieve this, we subdivided the flow duration curve into five different segments. The results showed that the segmentation of very low/high and low/high flow allows analysis of the model performance for every important discharge event precisely. We conclude that the additional segmentation of the flow duration curve into low and very low flows is essential for taking into account long low flow periods events. The identification of calibration runs with a satisfying reproduction of the whole hydrograph was improved due to combined evaluation of performance metrics for low and high flow periods. Furthermore, the combination of performance metrics for very high and very low discharge periods revealed that it is challenging to calibrate a model for a good performance in very high phases together with very low phases. Thus, the control of the model behavior in these special discharge conditions can be improved by analyzing the parameters of the governing process equations with these appropriate performance metrics.
Thus, we recommend this approach for analyses of the model behavior for all phases of the hydrograph including both extreme conditions. The differentiation into very low and low flow indices is recommended as additional low flow indices for common practice of model evaluation to enhance the overall model performance.

3.6 Acknowledgements

We thank the Government-Owned Company for Coastal Protection, National Parks and Ocean Protection (LKN-SH) of Schleswig-Holstein for the discharge data, the land survey office of Schleswig-Holstein for providing the digital elevation model and the river net, the German Weather Service (DWD) for the climate data and the Potsdam Institute for Climate Impact Research (PIK) for the STAR data. This project has been carried out with financial support of a scholarship for the first author by the German Environmental Foundation (DBU). This work was partially supported by the German Federal Ministry for Education and Research (BMBF) via the project IMPACT (grant number 02WM1136). We would like to thank the community of the open source software R, which was used for calibration of the SWAT model and following analysis.

3.7 References

Arnold, J. G., R. Srinivasan, R. S. Muttiah, and J. R. Williams (1998), Large area hydrologic modeling and assessment part I: Model development, *Journal of the American Water Resources Association 34*, *1*, 73–89, doi:10.1111/j.1752-1688.1998.tb05961.x.

Bärlund, I., T. Kirkkala, O. Malve, and J. Kämäri (2007), Assessing SWAT model performance in the evaluation of management actions for the implementation of the water framework directive in a finnish catchment, *Environmental Modelling & Software*, *22*(5), 719–724, doi:10.1016/j.envsoft.2005.12.030.

Bekele, E. G., and J. W. Nicklow (2007), Multi-objective automatic calibration of SWAT using NSGA-II, *Journal of Hydrology*, *341*(3), 165–176, doi:10.1016/j.jhydrol.2007.05.014.

BGR (1999), Bundesanstalt fuer Geowisschenschaften und Rohstoffe - Bodenuebersichtskarte im Maßstab 1:200.000. Verbreitung der Bodengesellschaften.

Boyle, D. P., H. V. Gupta, and S. Sorooshian (2000), Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, *Water Resources Research*, *36*(12), 3663–3674, doi:10.1029/2000WR900207.

Carrillo, G., P. Troch, M. Sivapalan, T. Wagener, C. Harman, and K. Sawicz (2011), Catchment classification: hydrological analysis of catchment behavior through process-based modeling along a climate gradient, *Hydrology and Earth System Sciences*, *15*, 3411–3430, doi:10.5194/hess-15-3411-2011.

Cheng, L., M. Yaeger, A. Viglione, E. Coopersmith, S. Ye, and M. Sivapalan (2012), Exploring the physical controls of regional patterns of flow duration curves–part 1: Insights from statistical analyses, *Hydrology and Earth System Sciences*, *16*, 4435–4446, doi:10.5194/hess-16-4435-2012.

Cibin, R., K. Sudheer, and I. Chaubey (2010), Sensitivity and identifiability of stream flow generation parameters of the SWAT model, *Hydrological Processes*, *24*(9), 1133–1148, doi:10.1002/hyp.7568.

Dawson, C. W., R. J. Abrahart, and L. M. See (2007, HydroTest: A web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts, *Environmental Modelling and Software*, *22*, 1034–1052, doi:10.1016/j.envsoft.2006.06.008.

De Vos, N., T. Rientjes, and H. Gupta (2010), Diagnostic evaluation of conceptual rainfall–runoff models using temporal clustering, *Hydrological Processes*, *24*(20), 2840–2850, doi:10.1002/hyp.7698.

Dunn, S. M. (1999), Imposing constraints on parameter values of a conceptual hydrological model using baseflow response, *Hydrology and Earth System Sciences*, *3*(2), 271–284.

DWD (2012), Weather and climate data from the German Weather Service of the station Flensburg (1961-1990), *Online climate data*.

Eckhardt, K. (2008), A comparison of baseflow indices, which were calculated with seven different baseflow separation methods, *Journal of Hydrology*, *352*(1), 168–173, doi:10.1016/j.jhydrol.2008.01.005.

Fohrer, N., and B. Schmalz (2012), Das UNESCO Ökohydrologie-Referenzprojekt Kielstau-Einzugsgebiet - nachhaltiges Wasserressourcenmanagement und Ausbildung im ländlichen Raum, *Hydrologie und Wasserbewirtschaftung/Hydrology and Water Resources Management-Germany*, *56*(4), 160–168.

Fohrer, N., B. Schmalz, F. Tavares, and J. Golon (2007), Modelling the landscape water balance of mesoscale lowland catchments considering agricultural drainage systems, *Hydrologie und Wasserbewirtschaftung/Hydrology and Water Resources Management-Germany*, *51*(4), 164–169.

Fohrer, N., A. Dietrich, O. Kolychalow, and U. Ulrich (2013), Assessment of the environmental fate of the herbicides flufenacet and metazachlor with the SWAT model, *Journal of Environmental Quality*, *42*, 1–11, doi:10.2134/jeq2011.0382.

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

Gupta, H. V., S. Sorooshian, and P. O. Yapo (1999), Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration, *Journal of Hydrologic Engineering*, *4*(2), 135–143, doi:10.1061/(ASCE)1084-0699(1999)4:2(135).

Gupta, H. V., T. Wagener, and Y. Liu (2008), Reconciling theory with observations: Elements of a diagnostic approach to model evaluation, *Hydrological Processes*, *22*(18), 3802–3813, doi:10.1002/hyp.6989.

Guse, B., D. E. Reusser, and N. Fohrer (2013), How to improve the representation of hydrological processes in swat for a lowland catchment - temporal analysis of parameter sensitivity and model performance, *Hydrological Processes*, doi:10.1002/hyp.9777, in press.

Gustard, A., and S. Demuth (2009), Manual on low-flow estimation and prediction, operational hydrology report no. 50, WMO-no. 1029.

Gustard, A., A. Bullock, and J. Dixon (1992), *Low flow estimation in the United Kingdom*, Institute of Hydrology, IH Report 108.

Herbst, M., and M. Casper (2008), Towards model evaluation and identification using self-organizing maps, *Hydrology and Earth System Sciences*, *12*(2), 657–667, doi:10.5194/hess-12-657-2008.

Herman, J., P. Reed, and T. Wagener (2013), Time-varying sensitivity analysis clarifies the effects of watershed model formulation on model behavior, *Water Resources Research*, *49*(3), 1400–1414, doi:10.1002/wrcr.20124.

Hunter, N. M., P. D. Bates, M. S. Horritt, and M. D. Wilson (2007), Simple spatially-distributed models for predicting flood inundation: A review, *Geomorphology*, *90*(3), 208–225, doi:10.1016/j.geomorph.2006.10.021.

Kiesel, J., N. Fohrer, B. Schmalz, and M. White (2010), Incorporating landscape depressions and tile drainages of a northern german lowland catchment into a semi-distributed model, *Hydrological Processes*, *24*(11), 1472–1486, doi:10.1002/hyp.7607.

Koch, S., A. Bauwe, and B. Lennartz (2013), Application of the SWAT model for a tile-drained lowland catchment in north-eastern Germany on subbasin scale, *Water Resources Management*, *27*(3), 791–805, doi:10.1007/s11269-012-0215-x.

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

Laaha, G., and G. Blöschl (2006), Seasonality indices for regionalizing low flows, *Hydrological processes*, *20*(18), 3851–3878, doi:10.1002/hyp.6161.

Laaha, G., and G. Blöschl (2007), A national low flow estimation procedure for Austria, *Hydrological Sciences Journal*, *52*(4), 625–644, doi:10.1623/hysj.52.4.625.

Legates, D. R., and G. J. McCabe (1999), Evaluating the use of goodness-of-fit measures in hydrologic and hydroclimatic model validation, *Water Resources Research*, *35*, 233–241, doi:10.1029/1998WR900018.

LKN (2013), Messstelle Soltfeld: Wasserstand und Abfluss, *Landesbetrieb fuer Kuestenschutz, Nationalpark und Meeresschutz Schleswig-Holstein; Hydrologie, Mess- und Beobachtungsdienst.*

Lombardi, L., E. Toth, A. Castellarin, A. Montanari, and A. Brath (2012), Calibration of a rainfall–runoff model at regional scale by optimising river discharge statistics: Performance analysis for the average/low flow regime, *Physics and Chemistry of the Earth, Parts A/B/C*, *42*, 77–84, doi:10.1016/j.pce.2011.05.013.

LVermA (1995), Landesvermessungsamt Schleswig-Holstein Digitales Geländemodell fuer Schleswig-Holstein. Quelle: TK25. Gitterweite 25 m x 25 m und TK50 Gitterweite 50 m x 50 m sowie ATKIS-DGM2-1 m x 1 m Gitterweite und DGM 5 m x 5 m Gitterweite, abgeleitet aus LiDAR-Daten.

Madsen, H. (2000), Automatic calibration of a conceptual rainfall–runoff model using multiple objectives, *Journal of Hydrology*, *235*(3), 276–288, doi:10.1016/S0022-1694(00)00279-1.

Madsen, H., G. Wilson, and H. C. Ammentorp (2002), Comparison of different automated strategies for calibration of rainfall-runoff models, *Journal of Hydrology*, *261*(1), 48–59, doi:10.1016/S0022-1694(01)00619-9.

Martinez, G. F., and H. V. Gupta (2011), Hydrologic consistency as a basis for assessing complexity of monthly water balance models for the continental United States, *Water Resources Research*, *47*(12), doi:10.1029/2011WR011229.

Moriasi, D. N., J. G. Arnold, M. W. V. Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith (2007), Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, *Transactions of the ASABE*, *50*, 885–900.

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

Neitsch, S. L., J. G. Arnold, J. R. Kiniry, and J. R. Williams (2011), SWAT theoretical documentation version 2009, *Grassland, Soil and Water Research Laboratory, Agricultural Research Service. Blackland Research Center, Texas Agricultural Experiment Station.*

Orlowsky, B., F.-W. Gerstengarbe, and P. Werner (2008), A resampling scheme for regional climate simulations and its performance compared to a dynamical RCM, *Theoretical and Applied Climatology*, *92*(3), 209–223, doi:10.1007/s00704-007-0352-y.

Pfannerstill, M., B. Guse, and N. Fohrer (2013), A multi-storage groundwater concept for the SWAT model to emphasize nonlinear groundwater dynamics in lowland catchments, *Hydrological Processes*, in press, doi:10.1002/hyp.10062.

Pokhrel, P., K. Yilmaz, and H. Gupta (2012), Multiple-criteria calibration of a distributed watershed model using spatial regularization and response signatures, *Journal of Hydrology*, *418–419*, 49-60, doi:10.1016/j.jhydrol.2008.12.004.

R Core Team (2013), R: A language and environment for statistical computing. Vienna, Austria: R foundation for statistical computing; 2013.

Santhi, C., J. G. Arnold, J. R. Williams, W. A. Dugas, R. Srinivasan, and L. M. Hauck (2001), Validation of the SWAT model on a large riwer basin with point and nonpoint sources, *JAWRA Journal of the American Water Resources Association*, *37*(5), 1169–1188, doi:10.1111/j.1752-1688.2001.tb03630.x.

Sawicz, K., T. Wagener, M. Sivapalan, P. Troch, and G. Carrillo (2011), Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA, *Hydrology and Earth System Sciences*, *15*, 2895–2911, doi:10.5194/hess-15-2895-2011.

Schmalz, B., F. Tavares, N. Fohrer (2008), Modelling hydrological processes in mesoscale lowland river basins with SWAT-capabilities and challenges, *Hydrological Sciences Journal*, *53*, 989–1000, doi:10.1623/hysj.53.5.989.

Schmalz, B., and N. Fohrer (2009), Comparing model sensitivities of different landscapes using the ecohydrological SWAT model, *Advances in Geosciences*, *21*, 91–98, doi:10.5194/adgeo-21-91-2009.

Smakhtin, V. (2001), Low flow hydrology: A review, *Journal of Hydrology*, *240*(3), 147–186, doi:10.1016/S0022-1694(00)00340-1.

Soetaert, K., T. Petzoldt, et al. (2010), Inverse modelling, sensitivity and Monte Carlo Analysis in R using package FME, *Journal of Statistical Software*, *33*(3), 1–28, doi:10.1.1.160.5179.

Tallaksen, L. M., H. Madsen, and B. Clausen (1997), On the definition and modelling of streamflow drought duration and deficit volume, *Hydrological Sciences Journal*, *42*(1), 15–33, doi:10.1080/02626669709492003.

Tattari, S., J. Koskiaho, I. Barlund, and E. Jaakkola (2009), Testing a river basin model with sensitivity analysis and autocalibration for an agricultural catchment in SW Finland, *Agricultural and Food Science*, *18*(3-4), 3–4.

Thielen, J., J. Bartholmes, M. Ramos, and A. De Roo (2009), The european flood alert system–part I: Concept and development, *Hydrology and Earth System Sciences*, *13*(2), 125, doi:10.5194/hess-13-125-2009.

van Griensven, A., T. Meixner, S. Grunwald, T. Bishop, M. Diluzio, and R. Srinivasan (2006), A global sensitivity analysis tool for the parameters of multi-variable catchment models, *Journal of Hydrology*, *324*(1), 10–23, doi:10.1016/j.jhydrol.2005.09.008.

van Werkhoven, K., T. Wagener, P. Reed, and Y. Tang (2009), Sensitivity-guided reduction of parametric dimensionality for multi-objective calibration of watershed models, *Advances in Water Resources*, *32*(8), 1154–1169, doi:10.1016/j.advwatres.2009.03.002.

Vogel, R. M., and N. M. Fennessey (1994), Flow-duration curves. I: New interpretation and confidence intervals, *Journal of Water Resources Planning and Management*, *120*(4), 485–504, doi:10.1061/(ASCE)0733-9496(1994)120:4(485).

Wagener, T., and H. V. Gupta (2005), Model identification for hydrological forecasting under uncertainty, *Stochastic Environmental Research and Risk Assessment*, *19*(6), 378–387, doi:10.1007/s00477-005-0006-5.

Wagener, T., D. P. Boyle, M. J. Lees, H. S. Wheater, H. V. Gupta, S. Sorooshian, et al. (2001), A framework for development and application of hydrological models, *Hydrology and Earth System Sciences*, *5*(1), 13–26, doi:10.5194/hess-5-13-2001.

White, K. L., and I. Chaubey (2005), Sensitivity analysis, calibration, and validations for a multisite and multivariable SWAT model, *JAWRA Journal of the American Water Resources Association*, *41*(5), 1077–1089, doi:10.1111/j.1752-1688.2005.tb03786.x.

Winchell, M., R. Srinivasan, M. DiLuzio, and J. G. Arnold (2010), ArcSWAT interface for SWAT2009, users guide, *Blackland Research and Extension Center*, pp. 1–495.

Wu, K., and C. A. Johnston (2007), Hydrologic response to climatic variability in a great lakes watershed: A case study with the SWAT model, *Journal of Hydrology*, *337*(1), 187–199, doi:10.1016/j.jhydrol.2007.01.030.

Yadav, M., T. Wagener, and H. Gupta (2007), Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, *Advances in Water Resources*, *30*(8), 1756–1774, doi:10.1016/j.advwatres.2007.01.005.

Yilmaz, K. K., H. V. Gupta, and T. Wagener (2008), A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, *Water Resour. Res*, *44*(9), W09,417, doi:10.1029/2007WR006716.

Yokoo, Y., and M. Sivapalan (2011), Towards reconstruction of the flow duration curve: Development of a conceptual framework with a physical basis, *Hydrology and Earth System Sciences*, *15*(9), 2805–2819, doi:10.5194/hess-15-2805-2011.

Zambrano-Bigiarini, M. (2012), hydroTSM: Time series management, analysis and interpolation for hydrological modelling, *R package version 0.3-3*.

Zhang, C., J. Chu, and G. Fu (2012), Sobol's sensitivity analysis for a distributed hydrological model of Yichun river basin, China, *Journal of Hydrology*, *480*, 58–68, doi:10.1016/j.jhydrol.2012.12.005.

Zhang, H., G. H. Huang, D. Wang, and X. Zhang (2011), Multi-period calibration of a semi-distributed hydrological model based on hydroclimatic clustering, *Advances in Water Resources*, *34*(10), 1292–1303, doi:10.1016/j.advwatres.2011.06.005.

4 Temporal parameter sensitivity guided verification of process dynamics

Pfannerstill, M., Guse, B., D.Reusser and Fohrer, N.: Temporal parameter sensitivity guided verification of process dynamics, Hydrol. Earth Syst. Sci. Discuss., 12, 1729-1764, doi:10.5194/hessd-12-1729-2015, 2015.

Received: 11 December 2014 - Accepted: 23 January 2015

Abstract

To ensure reliable results of a hydrological model, it is essential that the model reproduces the hydrological processes adequately. Information about process dynamics is provided by looking at the temporal sensitivities of the corresponding model parameters. For this, the temporal dynamics of parameter sensitivity are used to describe the dominance of parameters for each time step. The parameter dominance is then related to the corresponding hydrological process, since the temporal parameter sensitivity represents the modelled hydrological process. For a reliable model application it has to be verified that the modelled hydrological processes match the expectations of real-world hydrological processes.

We present a framework, which distinguishes between a verification of single model components and of the overall model behaviour. We analyse the temporal dynamics of parameter sensitivity of a modified groundwater component of a hydrological model. The results of the single analysis for the modified component show that the behaviour of the parameters of the modified groundwater component is consistent with the idea of the structural modifications. Additionally, the appropriate simulation of all relevant hydrological processes is verified as the temporal dynamics of parameter sensitivity represent these processes according to the expectations. Thus, we conclude that temporal dynamics of parameter sensitivity are helpful for verifying modifications of hydrological models.

4.1 Introduction

Hydrological models are driven by different interacting processes that are implemented into the model. To investigate the reliability of model results, it is essential to understand how these processes are represented. It needs to be analysed whether the model results are consistent with the hydrological processes in the catchment. These analyses are performed for the model structure, which is described by the model equations and different model parameters.

Knowledge about the model structures is crucial, especially when hydrological processes that control a response variable are not simulated appropriately (Hrachowitz et al., 2014). Model diagnostic analyses as proposed by (Gupta et al., 2008; Yilmaz et al., 2008) determine the appropriateness of process descriptions in the model structure. Thus, diagnostic methods help to detect failures in models and the corresponding components that need to be improved (Fenicia et al., 2008; Reusser and Zehe, 2011; Guse et al., 2014).

A first step to evaluate modifications to the model structures is the comparison between simulated and observed discharge. However, this comparison is not sufficient. It is essential to investigate if the newly introduced parameters match the expected sequence of processes. More specifically, there is the need to analyse how well they represent the corresponding real-world processes.

As stated by Yilmaz et al. (2008), a systematic approach is needed to analyse the adequacy of model structure and model improvements. There is a need to diagnose, if the modified model structures and

their newly introduced parameters are consistent with the expected sequence of hydrological processes according to the model concept. This is a step towards a general framework for model accuracy verification as emphasized by Wagener et al. (2001) and Yilmaz et al. (2008).
The relevance of model structure analysis for model improvement is highlighted by Clark et al. (2011) since the processes are not always reproduced appropriately. According to Massmann et al. (2014), the detection of periods in which a parameter or a set of parameters controls the model output provides diagnostic information. Guse et al. (2014) showed that this information is obtained by TEmporal Dynamics of PArameter Sensitivity (TEDPAS, Sieber and Uhlenbrook, 2005; Reusser et al., 2011). TEDPAS detects dominant parameters by analysing their sensitivity in a high temporal resolution. Since typical patterns of temporal parameter sensitivity change over time, the parameters can be related to corresponding hydrological processes. These hydrological processes and discharge phases vary temporally and hence the dominance of model components (Boyle et al., 2000, 2001; Wagener et al., 2003, 2009; Reusser et al., 2011; Garambois et al., 2013; Guse et al., 2014). The high temporal resolution supports the confirmation of the expected sequence of processes that is related to changing hydrological conditions.
In this context, Guse et al. (2014) used TEDPAS (Reusser et al., 2011) and temporal model performance analysis (TIGER (Reusser et al., 2009)) to detect the component of a hydrological model, which was responsible for poorly simulated baseflow in dry years. Although the simulated sequence of temporal parameter sensitivity was reasonable, the model performed poor for several performance metrics in phases of groundwater dominance (Guse et al., 2014). Based on this temporal diagnostic analysis, Pfannerstill et al. (2014a) modified the aquifer structure of the model to emphasise non-linear dynamics of the groundwater processes. The analysis of Pfannerstill et al. (2014b) showed that the modification improved the simulation of the discharge with respect to different performance metrics. Despite the well fitted discharge, there is the need to analyse if the hydrological processes are adequately represented by the model structure.
To fill this gap in knowledge, we present a framework that makes use of TEDPAS to verify improvements when model components were modified. TEDPAS provides temporal sensitivities of the newly introduced parameters. Furthermore, the sequence of high temporal parameter sensitivity can be interpreted to a sequence of processes. These results are then used for the verification. Hypotheses of the expected sequence of parameter sensitivity are derived from the model structure and the expected sequence of hydrological processes are derived from observations and known processes within the modelled catchment. The verification is performed by comparing the simulation results with the hypotheses of expected sequence of parameter sensitivity and expected sequence of hydrological processes within the catchment. For this, we assume that hypotheses for the sequence of parameter sensitivity sequence and hypothesised hydrological processes represent expectations derived by the analysis of the model structure and the known processes wihthin the catchment. The framework distinguishes between verification of single model components ($\text{TEDPAS}_{\text{single}}$) and verification of the overall model behaviour ($\text{TEDPAS}_{\text{all}}$). $\text{TEDPAS}_{\text{single}}$ is used to assess the consistency between expected and simulated sequence of temporal parameter sensitivity for a single, newly introduced model component. $\text{TEDPAS}_{\text{all}}$ is used to verify if the implementation of the modified component into the model structure is appropriate by analysing the sequence of processes of the modified component in relation to the other model components. For both approaches, the expectations for the verification are hypothesised on the basis of model structure and hydrological processes within the studied catchment.

Since the parameter sensitivities are related to the hydrological processes, the consistency in representing the whole hydrological system is investigated. For this, we propose a general framework for the verification of hydrologically consistent model modifications which are in principal applicable to any model in any catchment. We demonstrate:

- how a single component of a model, which was modified or newly introduced, can be verified by relating the sequence of high temporal parameter sensitivity to the expected sequence according to its underlying process equations ($\text{TEDPAS}_{\text{single}}$).
- how temporal parameter sensitivities can be used to assess the consistency between expected and simulated sequence of processes by analysing the model component based overall hydrological process representation ($\text{TEDPAS}_{\text{sall}}$).

4.2 Methods

The general idea to achieve hydrologically consistent model structures with model diagnostics and model improvements includes three steps (Fig. 17). Firstly, the reason for poor model performance for distinct discharge periods is detected (cf. Guse et al., 2014). The temporal parameter sensitivities are used to identify the corrsponding model component, which is responsible for the poor model performance due to model structure deficiencies (cf. Guse et al., 2014). Secondly, the structure of a single model component that is responsible for the poor hydrological process representation is modified to improve the model performance (cf. Pfannerstill et al., 2014a). Thirdly, the modified model component is verified by comparing simulated and hypothesised temporal parameter sensitivities using a framework, which is demonstrated in this study. This framework integrates two elements of consecutive TEDPAS analyses that is described in the following. In this context, we define TEDPAS as a diagnostic method, which provides results in terms of temporal dynamics of parameter sensitivity.

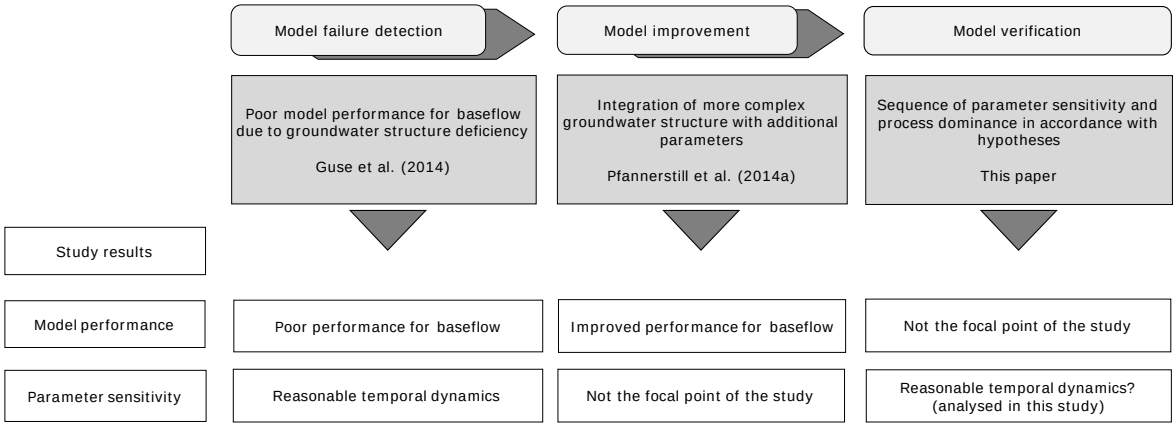

Figure 17: Steps for a hydrologically consistent model improvement.

4.2.1 TEDPAS methods

As shown in recent studies (Gupta et al., 2008; Yilmaz et al., 2008; Herbst et al., 2009; Reusser et al., 2009; van Werkhoven et al., 2009; Garambois et al., 2013; Herman et al., 2013; Pfannerstill et al., 2014b; Guse et al., 2014), a high temporal resolution is essential for proper diagnostic model evaluation. Therefore, TEDPAS aims to improve the understanding of model dynamics and to identify temporal dynamics of parameter sensitivity. For each time step, the sensitivity of the model output (e.g. discharge) is calculated on different parameters (cf. Reusser et al., 2009; Guse et al., 2014).

The temporal parameter sensitivities are related to hydrological processes. It is assumed that the parameter sensitivity represents the hydrological process that is described by process equations of the model and the corresponding parameters. The temporal dynamics of parameter sensitivity can be attributed to the temporal dynamics of hydrological processes. Accordingly, the dominant model processes for different periods of time can be determined (Sieber and Uhlenbrook, 2005; Cloke et al., 2008; Reusser et al., 2011).
There are three distinguishable goals in sensitivity analysis, namely factor prioritisation, factor fixing and factor mapping (Saltelli et al., 2006). The presented study focuses on the factor prioritisation setting to identify dominant model processes. These processes can be related to parameters that are dominant for the analysed time series (Reusser and Zehe, 2011). Thereby, periods of time that are especially useful for model calibration can be determined (Guse et al., 2014). The first-order partial variance is estimated to determine a measure of sensitivity (Saltelli et al., 2006). Parameters are simultaneously modified during partial variance estimations. Thus, TEDPAS investigates how a variation in model parameter values influences the variance of the model output (Eq. 11, from Reusser and Zehe (2011)). According to Reusser and Zehe (2011), the first-order partial variance is estimated by dividing the changes due to a specific parameter with the total variance V that is described by all model runs.

$$V = \sum_i V_i + \sum_{i<j} V_{ij} + \cdots + V_{1,2,3,\cdots,n} \tag{11}$$

V = total variance
V_i = variance of parameter θ_i (first order variance)
V_{ij} = covariance of θ_i (second order variance) and θ_j and higher order terms

For all parameters, the first-order partial variance is summed up. The sum of all partial variances cannot be higher than one by definition. However, it can be smaller than one due to parameter interactions. This is the case for the sensitivity of one parameter that is affected by other parameters. As shown by Saltelli et al. (2006); Nossent et al. (2011); Reusser and Zehe (2011); Sudheer et al. (2011); Herman et al. (2013); Massmann et al. (2014), the (extended) Fourier Amplitude Sensitivity Test (FAST) and Sobol′s method are applicable to determine the effect of parameter interactions. In this study, the FAST method was used. The FAST method considers non-linearities as an important factor in hydrology (Cukier et al., 1973, 1975, 1978) and has a high computational efficiency. In contrast with other methods such as Sobol's, the number of required model runs is lower, which is of particular relevance for complex models (Saltelli and Bolado, 1998; Reusser and Zehe, 2011). Since this algorithm has been implemented in the R-package FAST (Reusser, 2012), all analyses were made within the R environment. Readers are referred to Reusser and Zehe (2011) for further details.

4.2.2 TEDPAS as a framework for the verification of model improvements

The presented framework for the verification of model improvements is based on the main assumption that the provided information about high parameter sensitivity in a certain time period indicates the dominance of the corresponding model component. The presented framework for a TEDPAS-based verfication aims to provide insights into the modelled hydrological system in a high temporal resolution by using generally available data (e.g. daily discharge). In general, TEDPAS is applicable with or without measured data.
Parameters with a strong impact on the selected model output are assumed to be relevant for the process description in the model and can be related to model components. The provided diagnostic information is then used for two different TEDPAS-based analyses, $\text{TEDPAS}_{\text{single}}$ and $\text{TEDPAS}_{\text{all}}$.

4.2.2.1 TEDPAS$_{\text{single}}$

TEDPAS$_{\text{single}}$ aims to analyse the temporal parameter sensitivity within a modified or newly introduced model component. The main outcome of this analysis is a sequence of temporal parameter sensitivity, which is compared with the concept of the analysed model component. Focusing on the parameters of an individual model component, the relevance of each parameter can be identified precisely since possible interactions with parameters of other model components are excluded.

4.2.2.2 TEDPAS$_{\text{all}}$

TEDPAS$_{\text{all}}$ is used to verify the simulated sequence of hydrological processes using knowledge about the real processes. The main assumption of the TEDPAS$_{\text{all}}$ is that the sequence of hydrological processes is represented by temporal parameter sensitivities of different model components. A high temporal parameter sensitivity of a model component is assumed to reflect the hydrological process that is simulated by the model component. By applying TEDPAS$_{\text{all}}$, an accurate process implementation can be verified, especially for the modified or newly introduced model component.

4.2.2.3 Expected temporal parameter sensitivity and expected sequence of processes

To verify the single model component using TEDPAS$_{\text{single}}$, it is necessary to firstly define hypotheses about the expected temporal parameter sensitivity of the model. These hypotheses are derived from the concept of the model structure. By comparing the calculated parameter sensitivities with the hypothesised parameter sensitivities, the consistency between model parameter behaviour and the idea of the improved model structure is estimated.

To determine the hydrological consistency for the whole hydrological model with respect to the modified, single model component, the results of TEDPAS$_{\text{all}}$ are analysed. Therefore, the expected sequence of processes is hypothesised based on the knowledge of general hydrological and catchment specific processes. The hypotheses are compared with the results of TEDPAS$_{\text{all}}$, which provide information about the simulated sequence of hydrological processes.

4.3 Framework demonstration example

4.3.1 Catchment description and data

The Kielstau catchment comprises an area of about 50 km^2 and is located in the federal state of Schleswig-Holstein in the North Germany. It is a subbasin of the Treene catchment to which TEDPAS has been applied by Guse et al. (2014). The catchment is characterised by a maritime climate with a mean annual precipitation of 918,9 mm and mean annual temperature of 8,2° (Station: Gluecksburg-Meierwik, period: 1961 - 1990; DWD, 2012).

As reported by Kiesel et al. (2010), the catchment has a high water retention potential. However, due to the flat topography (27 m to 78 m above mean sea level), the water tables are very high in this region (Kiesel et al., 2010) and a high fraction of the agricultural area is drained (Fohrer et al., 2007). The installed tile drainages conribute to fast runoff and consequently increase peak flows, especially in winter (Kiesel et al., 2010). During drier periods decreasing tile drainage flow has been observed from April and May before tile drainage flow stops in summer months (Kiesel et al., 2009).

Another main characteristic of the Kielstau catchment is the close interaction between river and groundwater, which is due to high groundwater water tables that are directly connected to the river (Schmalz et al., 2008). The near-surface groundwater is controlled by precipitation, especially in winter (Schmalz et al., 2008). A more detailed description of the catchment can be found in Fohrer and Schmalz (2012).

Catchment specific input data for the model includes a soil map (resolution 1:200.000, BGR, 1999)

and a digital elevation model (resolution 5 m; LVermA, 1995). To define land use and crop rotations, data from mapping campaigns of 2011/2012 and 2012/2013 were available from Pfannerstill et al. (2014a,b). The soil and crop databases, and the spatial distribution of tile drainages were obtained from Fohrer et al. (2013, 2007).
Precipitation data was provided by the Gluecksburg-Meierwik weather station located north of the Kielstau catchment (DWD, 2012). Additional weather input from the STAtistical Regional model (STAR, Orlowsky et al. 2008) was used to fill gaps of needed data. The STAR data were already used as recent climate data for the SWIM model (e.g. Huang et al., 2010; Martinkova et al., 2011). In this study, wind speed, temperature, solar radiation, and humidity of STAR were used to fill data gaps.

4.4 Model description and setup

In the following, the hydrological model is described, which was used to exemplarily show the application of TEDPAS for verification of a modified model component. The semi-distributed, eco-hydrological SWAT model (Arnold et al., 1998) uses distinct spatial positions for the subbasins within the catchment. Within the subbasins, Hydrological response units (HRU) are used to describe areas of the same land use, slope and soil. The different components of the SWAT model have an empirical and process-oriented character. Due to the incorporation of several model components, there is high number of parameters which increase the complexity of the SWAT model (Cibin et al., 2010).
The water balance is driven mainly by the processes of precipitation, evapotranspiration, runoff, soil water percolation, drainage and groundwater flow. Runoff is routed through the main reaches of the subbasins to the catchment outlet. A detailed description of process implementation and the theory about the SWAT model can be found in Neitsch et al. (2011).
To set up the model, 36 subbasins and 2214 HRUs, which were determined using three slope classes ($< 2{,}6\,\%$, 2.6 - 4,6 % and $> 4{,}6\,\%$), were defined with ArcSWAT interface (version 2012.10.1.6). For the application of the TEDPAS-based model verification, the SWAT_{3S} version (Pfannerstill et al., 2014a) with its modified groundwater structure was used. Therefore, the groundwater input files were reprocessed using a script in the R environment (R Core Team, 2013) to add the additional groundwater input parameters required by SWAT_{3S}. To obtain equilibrium for the different storages of the model, a warm-up period from 1997 to 2000 was chosen. The temporal sensitivity analysis was performed for the hydrological years of 2001 to 2004.

4.4.1 Demonstration of verification framework

The verification framework for a modified model component is demonstrated by applying $\text{TEDPAS}_{\text{single}}$ and $\text{TEDPAS}_{\text{all}}$ to the modified groundwater component of SWAT_{3S}. $\text{TEDPAS}_{\text{single}}$ was used to verify the sequence of temporal parameter sensitivity for the groundwater module. With $\text{TEDPAS}_{\text{all}}$ the expected sequence of processes of surface runoff, tile drainage flow, evaporation and soil water storage is analysed. For $\text{TEDPAS}_{\text{single}}$ and $\text{TEDPAS}_{\text{all}}$, the model parameters (Tab. 10) and their ranges were selected according to previous SWAT model studies (Guse et al., 2014; Pfannerstill et al., 2014a).

4.4.1.1 $\text{TEDPAS}_{\text{single}}$ for model parameter verification

In the following, the SWAT_{3S} groundwater component with its parameters (Fig. 18) is briefly described. Also, we formulate hypotheses about the expected sequence of temporal parameter sensitivity. These hypotheses are the basis for the verification with $\text{TEDPAS}_{\text{single}}$. For the detailed process equations we refer to the appendix. A brief description of the main groundwater processes of the original SWAT version can be found in the supplement.

Table 10: Selection of parameters and its ranges for the temporal sensitivity analyses. The variation type distinguishes between replacing (r), multiplication (m) and addition/subtraction (as). The parameters are assigned according to the hydrological process including surface runoff (SR), soil water storage (SW), drainage flow (DF), evapotranspiration (ETP), and groundwater flow (GW)

Parameter name	Abbreviation	Process	Range	Type
Curve number	CN2	SR/SW	-15 - 15	as
Surface runoff lag coefficient	SURLAG	SR	0.2 - 4.0	r
Available soil water capacity	SOL_AWC	SW	-0.07 - 0.10	as
Tile drain lag time	GDRAIN	DF	0.5 - 2.0	m
Distance between two tile drains	SDRAIN	DF	10000 - 45000	r
Multiplication factor for K_e	LATKSATF	DF	0.6 - 2.0	r
Soil evaporation compensation	ESCO	ETP	0.5 - 1.0	r
Delay fast shallow aquifer	GW_DELAY_{fsh}	GW	1 - 15	r
Recession fast shallow aquifer	$ALPHA\_BF_{fsh}$	GW	0.3 - 1	r
Percolation slow shallow aquifer	$RCHRG_{ssh}$	GW	0.65 - 0.80	r
Delay slow shallow aquifer	GW_DELAY_{ssh}	GW	15 - 60	r
Recession slow shallow aquifer	$ALPHA\_BF_{ssh}$	GW	0.0001 - 0.3000	r
Percolation deep aquifer	$RCHRG_{dp}$	GW	0.1 - 0.4	r

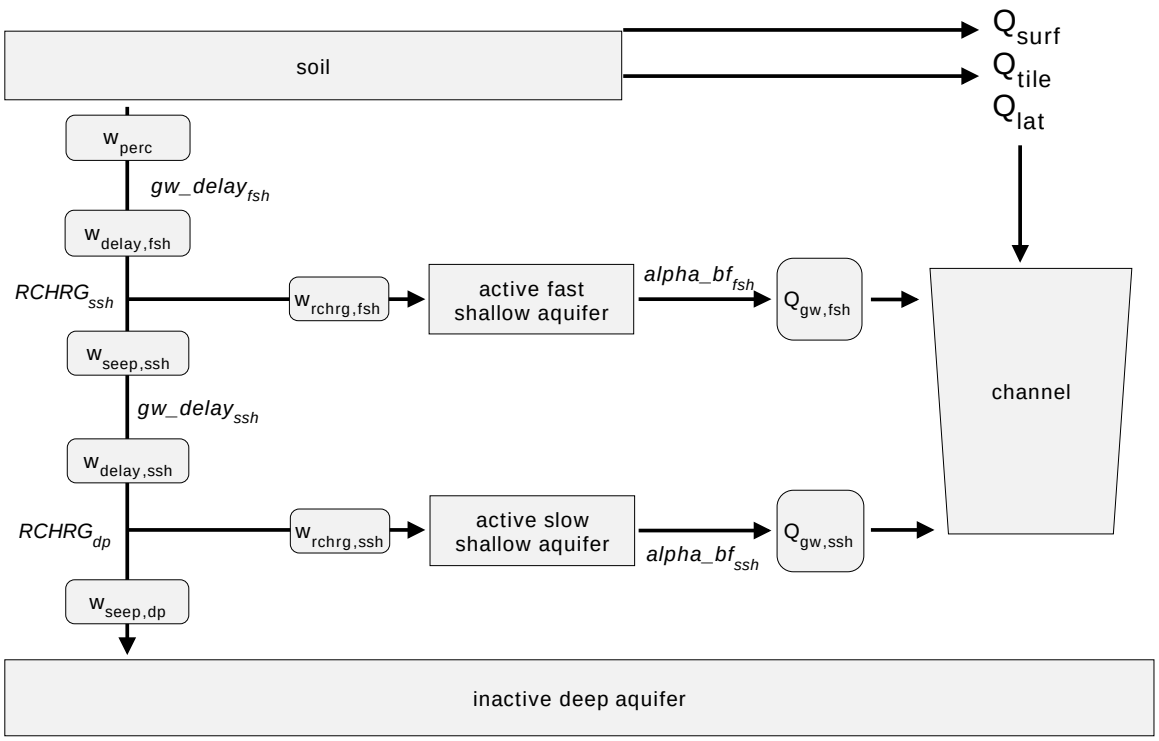

Figure 18: Description of the main groundwater processes and its parameters (highlighted in italic) of $SWAT_{3S}$ (cf. Pfannerstill et al. (2014a))

According to the concept for the $SWAT_{3S}$ groundwater module, the delay in the recharge of groundwater (GW_DELAY_{fsh}) is expected to be the first sensitive parameter. The delayed recharge is then partitioned ($RCHRG_{ssh}$) into a recharge to a fast shallow aquifer and to conceptually underlying aquifers (slow shallow and deep). Next, a recession constant ($ALPHA\_BF_{fsh}$) controls the contribution of the fast shallow aquifer to the stream. Based on the groundwater model concept, we expect the sequence of temporal parameter sensitivity to follow the order: GW_DELAY_{fsh}, $RCHRG_{ssh}$ and $ALPHA\_BF_{fsh}$ for the fast shallow aquifer (Hypothesis H1: sequence fast).
$SWAT_{3S}$ simulates also a delayed recharge (GW_DELAY_{ssh}) for aquifers conceptually located beneath the fast shallow aquifer (slow shallow and deep aquifer). The delayed recharge is partitioned ($RCHRG_{dp}$) into a recharge to a slow shallow and a deep aquifer. Finally, the contribution of the slow shallow aquifer is controlled by a recession constant ($ALPHA\_BF_{ssh}$). Consequently, we expect the temporal dynamics of parameter sensitivity to be similar to the expected sequence of the shallow aquifer parameters (H2: GW_DELAY_{ssh}, $RCHRG_{dp}$ and $ALPHA\_BF_{ssh}$ for sequence slow).
In general, the fast shallow aquifer was implemented to represent fast reacting groundwater processes in times of high discharge. In contrast, the slow shallow aquifer is intended to control the low flow phases by contributing delayed groundwater recharge. This concept should lead to an explicit sequence of temporal parameter sensitivity for the different aquifers. We hypothesise, that the parameters controlling the fast shallow aquifer (GW_DELAY_{fsh}, $RCHRG_{ssh}$, and $ALPHA\_BF_{fsh}$) are most relevant directly after a precipitation event, before the parameters controlling the slow shallow aquifer (GW_DELAY_{ssh}, $RCHRG_{dp}$, and $ALPHA\_BF_{ssh}$) become dominant later (H3: relation fast to slow).

4.4.1.2 $TEDPAS_{all}$ for model component verification

The consistency between the expected and the simulated sequence of processes is verified with $TEDPAS_{all}$. The sensitivity of parameters controlling the processes of surface runoff, tile drainage flow, evaporation and soil water storage is related to the groundwater processes. Thereby, the groundwater component is verified in the context of the overall process representation of the hydrological cycle.
The results of $TEDPAS_{all}$ are compared with hypotheses of temporal process patterns, which were developed for the case study catchment. These hypotheses are based on the concept of vertical water redistribution (Yilmaz et al., 2008) and on qualitative knowledge of the catchment processes. The vertical redistribution of water after excess rainfall between faster and slower runoff components is one of the primary functions of the watershed system (Yilmaz et al., 2008). Accordingly, we distinguish between the different processes of surface runoff, tile drainage flow, fast (primary) and slow (secondary) groundwater flow and evapotranspiration (Fig. 19).
Based on Fig. 19 and findings of Kiesel et al. (2010) for the study catchment, it is hypothesised that the surface runoff (CN2) and the surface runoff lag (SURLAG) are relevant during the whole year whenever the amount of precipitation exceeds the soil infiltration capacity (H4: surface runoff upon rainfall).
The amount of water that does not run off on the surface infiltrates into the soil and is stored (SOL_AWC) for a limited time and depending on soil water storage capacity. The storage capacity is directly connected with tile drainage and groundwater dynamics as shown by Kiesel et al. (2009, 2010); Schmalz et al. (2008) for the study catchment. In winter, groundwater tables are high which results in a high potential for groundwater extraction through the tile drainages (Kiesel et al., 2010). Based on the observations of Kiesel et al. (2009), it is expected that tile drainage flow leads to peak flows in winter due to groundwater ponding and a high soil water content. Consequently, we hypothesise that the effective lateral hydraulic conductivity factor (LATKSATF), the spacing for tile drainages

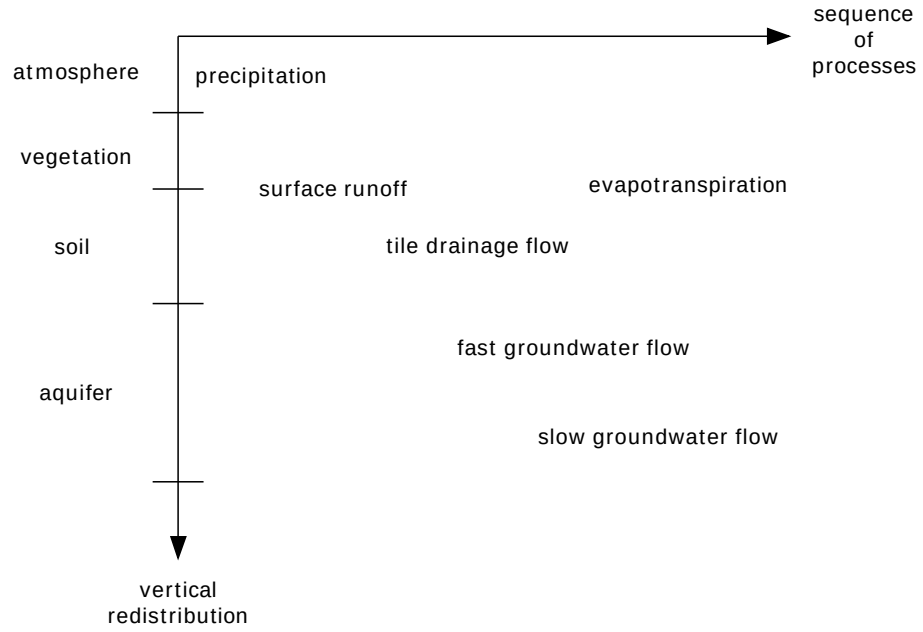

Figure 19: Schema of the expected sequence of processes after a precipitation event based on the concept of vertical water redistribution.

(SDRAIN), and their storage and lag time (GDRAIN) to be of high relevance mainly in winter (H5: tile drainage flow in winter).
In addition, high groundwater tables are the most important characteristic in the study catchment. During winter periods, the groundwater dynamics are mainly controlled by precipitation inputs due to a direct hydraulic connection between groundwater and river (Schmalz et al., 2008). In contrast, the dynamics of groundwater interaction decreases in summer but groundwater storage remains the main contributor of flow to the river. Based on these assumptions, we hypothesise a high relevance of fast groundwater flow represented by GW_DELAY$_{fsh}$, RCHRG$_{ssh}$, and ALPHA_BF$_{fsh}$ in winter and high relevance of the slow groundwater flow represented by GW_DELAY$_{ssh}$, RCHRG$_{dp}$, and ALPHA_BF$_{ssh}$ in the beginning of summer (H6: variable recession slope).
More specifically, GW_DELAY$_{fsh}$ is expected to be the first dominant parameter controlling fast groundwater recharge during high discharge periods in winter. This fast groundwater recharge is followed by increasing dominance of RCHRG$_{ssh}$ and ALPHA_BF$_{fsh}$ which control the outflow from the aquifer at decreasing high discharge (H7: fast groundwater flow at high discharge). At the beginning of the recession, the delayed recharge (GW_DELAY$_{ssh}$) is expected the be the main process controlling the discharge generation, followed by an increasing relevance of RCHRG$_{dp}$, and ALPHA_BF$_{ssh}$ (H8: flat recession at low discharge).
Since Kiesel et al. (2009) observed that tile drainage flow decreases during April and May before tile drainages run completely dry in the summer period, we expect decreasing relevance of the drainage model component. Also, due to the climatic conditions in the Kielstau catchment, the summer periods are characterized by dry soil layers and extraction of soil water by vegetation (Kiesel et al., 2010). As a consequence, groundwater recharge is very limited and the dominance of the groundwater module is decreasing. Based on this observation, we hypothesise high relevance of the soil water storage capacity (SOL_AWC) and the soil evaporation compensation (ESCO) in dry summer months until the beginning of resaturation phases (H9: evaporation at resaturation).

Table 11: Hypotheses for model verification, derived from model concept, theory of vertical water redistribution and known hydrological processes within the catchment with related model parameters

Abbreviation	Description	Source	Parameter
H1	sequence fast	model concept	GW_DELAY$_{fsh}$, ALPHA_BF$_{fsh}$, RCHRG$_{ssh}$
H2	sequence slow	model concept	GW_DELAY$_{ssh}$, ALPHA_BF$_{ssh}$, RCHRG$_{dp}$
H3	relation fast to slow	model concept	GW_DELAY$_{fsh\ +\ ssh}$, ALPHA_BF$_{fsh\ +\ ssh}$, RCHRG$_{ssh\ +\ dp}$
H4	surface runoff upon rainfall	vertical water redistribution	CN2, SURLAG
H5	tile drainage flow in winter	observation in catchment	GDRAIN, SDRAIN, LATKSATF
H6	variable recession slope	observation in catchment	GW_DELAY$_{,fsh}$, GW_DELAY$_{ssh}$
H7	fast groundwater flow at high discharge	vertical water redistribution	GW_DELAY$_{fsh}$, ALPHA_BF$_{fsh}$, RCHRG$_{ssh}$
H8	flat recession at low discharge	vertical water redistribution	GW_DELAY$_{ssh}$, ALPHA_BF$_{ssh}$, RCHRG$_{dp}$
H9	evaporation at resaturation	observation in catchment, vertical water redistribution	ESCO, SOL_AWC

4.5 Description and discussion of the results

4.5.1 Temporal sensitivity of groundwater parameters ($\mathbf{TEDPAS_{single}}$)

The sensitivity of all six groundwater parameters varied considerably between different discharge phases (Fig. 20). Based on the temporal parameter sensitivities, a pattern of parameter relevance could be observed. The delay time of the fast shallow aquifer ($\mathrm{GW\_DELAY}_{fsh}$) had the strongest effect on high discharge events caused by large amounts of precipitation (Fig. 20a). Next, the relevance for controlling the percolation to the fast and slow shallow aquifers (RCHRG_{ssh}) increased. Finally, the recession constant $\mathrm{ALPHA\_BF}_{fsh}$ was sensitive at the end of high discharge phases (Fig. 20c). Thus, hypothesis H1 is verified, as the expected sequence of temporal parameter sensitivity of $\mathrm{GW\_DELAY}_{fsh}$, RCHRG_{ssh}, and $\mathrm{ALPHA\_BF}_{fsh}$ was confirmed.
Regarding hypothesis H2, the expected sequence of temporal parameter sensitivity of $\mathrm{GW\_DELAY}_{ssh}$, RCHRG_{dp}, and $\mathrm{ALPHA\_BF}_{ssh}$ were confirmed as well. In comparison to the fast shallow aquifer, this sequence is much clearer as the sensitivity of the parameters of the slow shallow aquifer showed a much higher temporal variability (Fig. 20).
The most important finding about the overall parameter sensitivity is an earlier reaction of the parameters for the fast shallow aquifer compared to the slow shallow aquifer, which was the expectation of the modified groundwater concept and hypothesised with H3. Overall, the fast aquifer parameters were most sensitive during recession phases following discharge peaks. In contrast, the slow aquifer parameters dominated the low flow periods. Thus, hypothesis H3 was confirmed as well, which expected an earlier reaction of the parameters controlling the response of the fast shallow aquifer to precipitation events compared to the parameters controlling the response of the slow shallow aquifer.

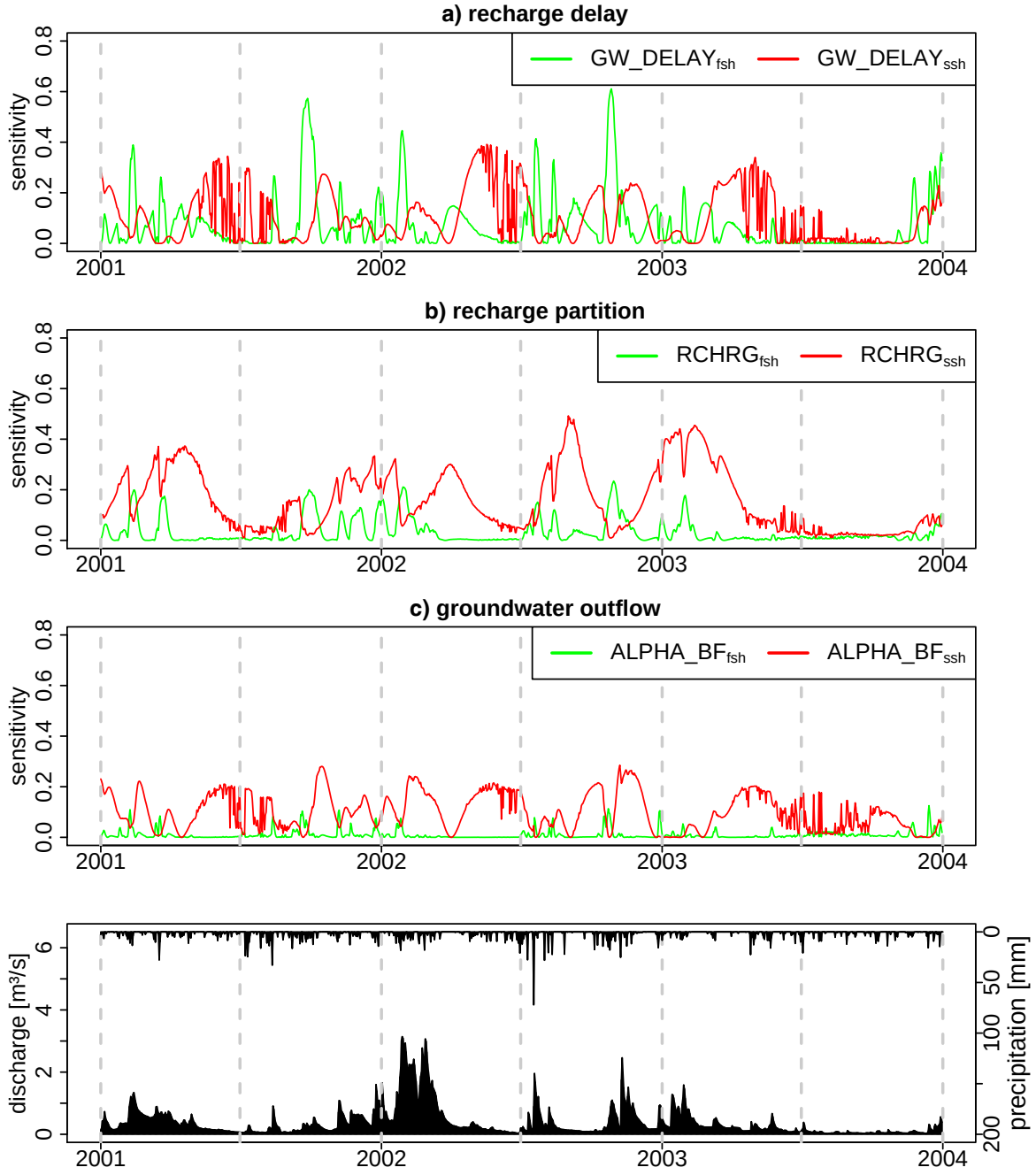

Figure 20: Temporal sensitivity for groundwater parameters of the fast (fsh) and the slow shallow aquifer (ssh). The different groundwater processes are seperated into recharge delay (a), the partition of groundwater recharge (b), and the groundwater outflow (c) according to the fast (fsh) and the slow shallow aquifer (ssh). The observed discharge and precipitation are shown additionally from 2001 to 2004 in the last subplot.

4.5.2 Overall process verification (TEDPAS$_{all}$)

TEDPAS$_{all}$ is used to determine the sequence of processes by analysing the temporal sensitivities of the different model components. The results show that the impact of the different components on discharge changed remarkably over time.

The impact of the model component controlling surface runoff (SURLAG and CN2) was observed during discharge peaks throughout the year (Fig. 21). The model component for surface runoff is the first component to become sensitive during a rainfall event, which confirms hypothesis H4. The expected sequence of processes, which was based on the observations of Kiesel et al. (2010) for the study catchment are confirmed by the sensitivity of the two parameters, which is clearly linked to short peak flow events during the whole simulation period (Fig. 21).

All other parameters showed a characteristic sequence of parameter sensitivity, which depends on to the discharge magnitude and the moisture conditions. The impact of tile drainages (GDRAIN, SDRAIN and LATKSATF) was very low in phases of low discharge during summer. This finding verifies hypotheses H5 and H9: tile drainages are inactive due to low water tables, which do not rise during the short and low precipitation events in summer periods. The highest dynamic of sensitivity and influence on the discharge was observed during wet periods in winter and spring (Fig. 21), where rising water tables are expected due to sufficient precipitation.

The low impact of the tile drainages can be further explained by the groundwater dominance at low flow periods, which is the next step in the sequence of processes as described by the concept of vertical water redistribution (see Fig. 19). The high impact of groundwater on discharge for the studied lowland catchment is particularly visible at the beginning and the end of the long lasting low flow periods, which is in accordance with hypothesis H8.

Additionally, there is a clear separation for the relevance of the fast and the slow shallow aquifers. The time delay for recharge of the fast shallow aquifer (GW_DELAY$_{fsh}$) becomes less relevant as soon as the influence of the time delay parameter of the slow shallow aquifer (GW_DELAY$_{ssh}$) increases. This result was expected, as the model structure expects a recharge to the fast shallow aquifer at high discharge with fast groundwater contribution (ALPHA_BF$_{fsh}$), followed by a delayed recharge to the slow shallow aquifer at recession phases with slow groundwater contribution (ALPHA_BF$_{fsh}$, hypotheses H6, H7, H8). Consequently, the low flow during dry periods is controlled by flow from the slow shallow aquifer to the channel (Fig. 21). This finding supports hypothesis H6, which expects a high relevance of the slow shallow aquifer parameters in the beginning of the low flow period in summer but low relevance in winter.

In general, the fast shallow aquifer had very limited impact on the discharge. In comparison to the results of TEDPAS$_{single}$, the impact of the fast shallow aquifer is lower, because the tile drainage flow controls the water amount for the groundwater recharge. Consequently, the process of fast discharge generation is controlled by both, the tile drainage flow and the fast shallow aquifer. This result was partly expected, since the parameters of the fast shallow aquifer were expected to be mainly relevant in winter (H5). Due to the low parameter sensitivity of the fast shallow aquifer, hypothesis H5 is partly verified. The overlap of high sensitivity of the parameters controlling tile drainage flow and the fast shallow aquifer emphasizes the relevance of a single model component analysis as performed with TEDPAS$_{single}$.

The partitioning of recharge of the slow shallow and the deep aquifer (RCHRG$_{dp}$) was especially important at the beginning of recession phases (Fig. 21), because it controls the water amount available for groundwater flow. According to the model structure, the total amount of recharge to the slow shallow and deep aquifers is affected by the partitioning of the recharge in the fast shallow aquifer. The more water flows into the fast shallow aquifer, the less is available for the slow shallow and the inactive deep aquifer. This behaviour is consistent with the model concept since the recharge to

the fast shallow aquifer is intended to be more important during wet phases with fast groundwater recharge (H6, H7). In contrast, the slow shallow aquifer is designed to control the slow recharge before recession phases (H6, H8).
The processes expected to become relevant last according to the concept of vertical water redistribution (Fig. 19) is the storage function of the soils and evaporation. The evaporation and soil water availability (ESCO and SOL_AWC) are most relevant during low flow periods in late summer and during phases of resaturation in the beginning of autumn. During these periods, the influence of all other processes is very limited. This highlights the relevance of additional storages besides the groundwater storages for the generation of baseflow in dry periods. Since the parameter sensitivities of the groundwater component is very low in these periods, hypothesis H9 is verified.

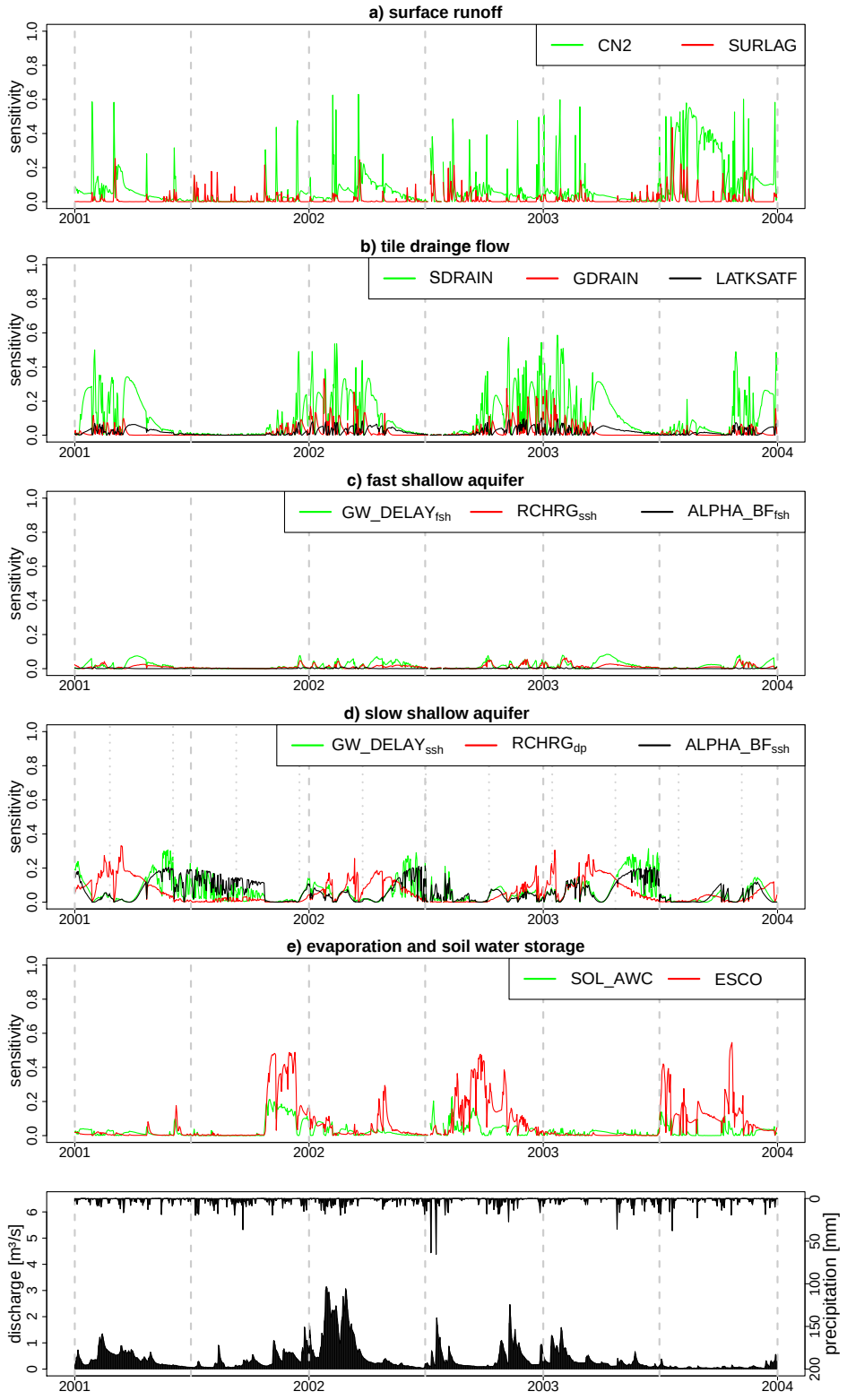

Figure 21: Temporal sensitivities for the groundwater parameters together with additional parameters for surface runoff (SURLAG, CN2), tile drainage flow (GDRAIN, SDRAIN and LATKSATF) and evaporation (ESCO, SOL_AWC). The parameters are ordered according to the processes of surface runoff (a), tile drainage flow (b), the process dynamics of the fast shallow aquifer (c) and the slow shallow aquifer (d), and the evaporation together with soil water storage (e). The observed discharge and precipitation are shown additionally from 2001 to 2004 in the last subplot.

4.6 Relevance of TEDPAS for the verification of model modifications

TEDPAS is a central method for model diagnostics and the verification of model improvements (Fig. 17). We build a framework with two different TEDPAS applications. In the following, it is discussed, whether the results of the presented TEDPAS framework provides diagnostic information for model verification upon modified or newly introduced model components. In this context, it is discussed if the application of TEDPAS can be interpreted as the last step for model verifications.
In this study, we exemplify the analysis of a modified model in regard to two different aspects: (i) the hydrological consistency within the model and (ii) the hydrological processes within a catchment. The general application of this framework is shown by abstracting our findings into a more general context. We hypothesise that this framework is applicable for any hydrological model in any catchment, which needs further demonstration.
Based on our analysis results of the modified model component, it was shown that there is the necessity to analyse the role of the newly introduced parameters. We interpret the results of the demonstration example to focus on the hydrological processes which are identified with high temporal resolution.
Due to the daily resolution, the hydrological processes of a single model component were clearly identified (fast and slow reacting aquifer). According to the model structure and our derived hypotheses, $\text{TEDPAS}_{\text{single}}$ confirmed the expected sequence of parameter sensitivity. Furthermore, the case study results revealed a simulated sequence of processes that is consistent with the concept of vertical water redistribution (Fig. 19) and according to our knowledge based process understanding for the study catchment. The simulated sequence of processes consistently exhibited the order with surface runoff as first process, followed by tile drainage. Finally, this sequence of processes continues with fast groundwater flow and slow groundwater flow (Fig. 20 and 21). However, the low sensitivity of the parameters for the fast shallow aquifer limits the verification to a small extent. Nonetheless, the sequence of processes is identifiable. Consequently, the confirmation of the consistency is the core result of the diagnostic analysis. It indicates that the simplified representation of the groundwater processes is in accordance with the concept of vertical process dynamics.
In this study, $\text{TEDPAS}_{\text{single}}$ and $\text{TEDPAS}_{\text{all}}$ were applied using commonly available, daily observed discharge data. The high temporal resolution facilitated the diagnosis of the model structure and its ability to simulate the processes occurring in the catchment. Thereby, TEDPAS provided additional diagnostic information to understand the representation of processes within the analysed model. Additionally, the presented example highlights the potential of the TEDPAS framework to evaluate the consistency of parameters and process structure using qualitative data. We used observed processes occurring the catchment, as well as the concept of vertical water redistribution (Fig. 19) and the theoretical foundations of the modified model structure (Fig. 18) to derive hypotheses for the model verification. This procedure can be transferred to any model and can be performed for studies in any catchment.
The results of this study show, that $\text{TEDPAS}_{\text{single}}$ and $\text{TEDPAS}_{\text{all}}$ are needed for the extraction of comprehensive model diagnostic information. The $\text{TEDPAS}_{\text{single}}$ method is used to check the consistency between expected and simulated sequence of temporal parameter sensitivity for the modified or newly introduced model component. With this approach, the role of each parameter can be clearly identified, especially due to the high temporal resolution. The application of $\text{TEDPAS}_{\text{all}}$ in our demonstration example revealed, that the highest sensitivity of single parameters of a modified model component and parameters of other model components may occur simultaneously. This finding emphasizes the importance of $\text{TEDPAS}_{\text{all}}$, since this method is able to identify the overlapping dominance of different model components and the corresponding hydrological processes.

4.7 Conclusion

The main capability of model diagnostics is the determination of the adequacy of process descriptions in model structures. In this study, we used temporal dynamics of parameter sensitivities (TEDPAS) as a verification method in model diagnostics. We propose three steps for model diagnostics and the verification of model improvements. Firstly, inappropriate model structures are detected (cf. Guse et al. (2014)) and secondly, the related process description within the model is modified to improve the representation of hydrological processes (cf. Pfannerstill et al. (2014a). The third step is the model verification with a TEDPAS-based framework, which is presented in this study.
Based on our results, we propose TEDPAS as a method to provide relevant diagnostic information after a modification of a model component. The presented framework includes the application of $\text{TEDPAS}_{\text{single}}$ and $\text{TEDPAS}_{\text{all}}$. In a high temporal resolution, $\text{TEDPAS}_{\text{single}}$ aims to provide information about the reasonable sequence of temporal parameter sensitivities within a single model component. Thereby, the intended role of parameters within a modified or newly introduced model component is verified. $\text{TEDPAS}_{\text{all}}$ is applied to analyse the sequence of processes including not only the modified, but all model components.
The main outcomes of this study are:

- TEDPAS provides diagnostic information for the verification of the consistency between the expected and simulated sequence of processes. The expected sequence of processes is derived from the model concept, qualitative knowledge of the catchment, and the concept of vertical water redistribution.
- $\text{TEDPAS}_{\text{single}}$ provides the sequence of temporal parameter sensitivity within a single modified or newly introduced model component.
- $\text{TEDPAS}_{\text{all}}$ provides the simulated sequence of processes of the whole model for the verification with the expected sequence of processes.

We recommend the use of TEDPAS as a part of a verification framework for model diagnostics, since it provides relevant information, which leads to an improved understanding of the relationship between modified model structure and the processes occurring in a catchment.

4.8 Appendix

4.8.1 The groundwater component for SWAT_{3S}

The idea of the modified groundwater component of SWAT_{3S} (Pfannerstill et al., 2014a) is the integration of two aquifers that may contribute to the river and one aquifer that accounts for percolation into deep geologic formations. For this, the shallow aquifer was split into a fast and a slow reacting storage. A detailed description of the groundwater processes of SWAT_{3S} can be found in Pfannerstill et al. (2014a). For comparisons with the original SWAT version, the governing process equations are described in the supplement.
In the following, the modified groundwater processes of SWAT_{3S} are briefly described. In a first step, a delay for soil water that percolates out of the soil $w_{perc,i}$ (mm H_2O) is considered in Eq. 13. The parameter GW_DELAY_{fsh} (days) describes the time delay for percolating water, entering the geologic formation of the fast shallow aquifer. The amount of water, percolating to the aquifer on the day before $(i-1)$ is represented by $w_{delay,fsh,i-1}$ (mm H_2O).

$$w_{delay,fsh,i} = (1 - exp[\frac{-1}{GW\_DELAY_{fsh}}]) \cdot w_{perc,i} \tag{12}$$

$$+ \ exp[\frac{-1}{GW\_DELAY_{fsh}}] \cdot w_{delay,fsh,i-1} \tag{13}$$

SWAT_{3S} considers the delayed percolation water $w_{delay,fsh,i}$ (mm H_2O) which is split into recharge of the fast shallow aquifer and into recharge that is entering the geologic formation of the slow shallow aquifer. Water percolating to the slow shallow aquifer is represented by $w_{seep,ssh,i}$ (mm H_2O, Eq. 14). The parameter RCHRG_{ssh} is a partitioning coefficient, which is used to calculate the percolation into the slow shallow aquifer. The recharge of the fast shallow aquifer $w_{rchrg,fsh,i}$ (mm H_2O) is calculated by subtracting the water that is percolating into the geologic formation of the slow shallow aquifer with Eq. 15. :

$$w_{seep,ssh,i} = RCHRG_{ssh} \cdot w_{rchrg,i} \tag{14}$$

$$w_{rchrg,fsh,i} = w_{rchrg,i} - w_{seep,ssh,i} \tag{15}$$

The concept of SWAT_{3S} assumes a delay of the calculated seepage to the slow shallow aquifer $w_{seep,ssh,i}$ (Eq. 14, Eq. 17). Thereby, the time delay of recharge due to different geologic formations is described with GW_DELAY_{ssh} (days).

$$w_{delay,ssh,i} = (1 - exp[\frac{-1}{GW\_DELAY_{ssh}}]) \cdot w_{seep,ssh,i} \tag{16}$$

$$+ \ exp[\frac{-1}{GW\_DELAY_{ssh}}] \cdot w_{delay,ssh,i-1} \tag{17}$$

To consider percolation to the slow shallow aquifer on the day before, the parameter $w_{delay,ssh,i-1}$ (mm H_2O) is used. SWAT_{3S} incorporates the simulation of groundwater recharge to deep geologic formations. The percolation to the deep aquifer $w_{seep,dp,i}$ (mm H_2O) is calculated with Eq. 18:

$$w_{seep,dp,i} = RCHRG_{dp} \cdot w_{delay,ssh,i} \tag{18}$$

The delayed recharge to the slow shallow aquifer $w_{rchrg,ssh,i}$ (mm H_2O) is then simulated with Eq. 19:

$$w_{rchrg,ssh,i} = w_{delay,ssh,i} - w_{seep,dp,i} \tag{19}$$

Finally, the groundwater flow into the stream is calculated. As SWAT_{3S} considers two contributing groundwater storages, there are two equations for the simulation of groundwater flow. The groundwater flow out of the fast shallow aquifer is calculated with Eq. 21. The parameter ALPHA_BF_{fsh} (1/days), which is the baseflow recession constant, is used to describe the outflow of the aquifer ($Q_{gw,fsh,i}$, mm H_2O):

$$Q_{gw,fsh,i} = Q_{gw,fsh,i-1} \cdot exp[-ALPHA\_BF_{fsh} \cdot \Delta t] \tag{20}$$

$$+ \ w_{rchrg,fsh,i} \cdot (1 - exp[-ALPHA\_BF_{fsh} \cdot \Delta t]) \tag{21}$$

The contribution of the slow shallow aquifer to the discharge is calculated with Eq. 23:

$$Q_{gw,ssh,i} = Q_{gw,ssh,i-1} \cdot exp[-ALPHA\_BF_{ssh} \cdot \Delta t] \quad (22)$$

$$+ w_{rchrg,ssh,i} \cdot (1 - exp[-ALPHA\_BF_{ssh} \cdot \Delta t]) \quad (23)$$

The modified SWAT$_{3S}$ calculates the groundwater contribution of the slow shallow aquifer to the river $Q_{gw,ssh,i}$ (mm H_2O) using ALPHA_BF$_{ssh}$ (1/days), which is the baseflow recession constant for the slow shallow aquifer. The recharge of the slow shallow aquifer is described with the parameter $w_{rchrg,ssh,i}$ (mm H_2O).

4.9 Acknowledgements

The Government-Owned Company for Coastal Protection, National Parks and Ocean Protection (LKN-SH) of Schleswig-Holstein provided the discharge data for this study. The digital elevation model and the river net were obtained from the land survey office of Schleswig-Holstein. We thank the German Weather Service (DWD) for providing the climate data and the Potsdam Institute for Climate Impact Research (PIK) for providing the STAR data.
The first author was supported by a scholarship of the German Environmental Foundation (DBU). The DFG funded project GU 1466/1-1 (Hydrological consistency in modeling) supported the work of the second author. Dominik Reusser was supported by the BMBF via its initiative Potsdam Research Cluster for Georisk Analysis, Environmental Change and Sustainability (PROGRESS – Grant: 03IS2191B). We want to thank the community of the open source software R, which was used for the calibration of the SWAT model and following analysis.

4.10 References

Arnold, J. G., Srinivasan, R., Muttiah, R. S., and Williams, J. R.: Large area hydrologic modeling and assessment part I: Mmodel development, J. Am. Water Resour. As., 1, 73–89, doi:10.1111/j.1752-1688.1998.tb05961.x, 1998.

BGR: Bundesanstalt fuer Geowisschenschaften und Rohstoffe - Bodenuebersichtskarte im Maßstab 1:200.000. Verbreitung der Bodengesellschaften, 1999.

Boyle, D. P., Gupta, H. V., and Sorooshian, S.: Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, Water. Resour. Res., 36, 3663–3674, doi:10.1029/2000wr900207, 2000.

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

Cibin, R., Sudheer, K. P., and Chaubey, I.: Sensitivity and identifiability of stream flow generation parameters of the SWAT model, Hydrol. Process., 24, 1133–1148, doi:10.1002/hyp.7568, 2010.

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

Cloke, H., Pappenberger, F., and Renaud, J.-P.: Multi-method global sensitivity analysis (MMGSA) for modelling floodplain hydrological processes, Hydrol. Process., 22, 1660–1674, doi:10.1002/hyp.6734, 2008.

Cukier, R. I., Fortuin, C. M., Shuler, K. E., Petschek, A. G., and Schaibly, J. H.: Study of sensitivity of coupled reaction systems to uncertainties in rate coefficients .1. Theory, J. Chem. Phys., 59(8), 3873–3878, doi:10.1063/1.1680571, 1973.

Cukier, R. I., Schaibly, J. H., and Shuler, K. E.: Study of sensitivity of coupled reaction systems to uncertainties in rate coefficients .3. Analysis of approximations, J. Chem. Phys., 63(3), 1140–1149, doi:10.1063/1.431440, 1975.

Cukier, R. I., Levine, H. B., and Shuler, K. E.: Non-linear sensitivity analysis of multi-parameter model systems, J. Comput. Phys., 26, 1–42, doi:10.1016/0021-9991(78)90097-9, 1978.

DWD: Weather and climate data from the German Weather Service (DWD) of the station Flensburg (1961-1990), Online climate data, 2012.

Fenicia, F., Savenije, H., and Winsemius, H.: Moving from model calibration towards process understanding, Phys. Chem. Earth, 33, 1057–1060, doi:10.1016/j.pce.2008.06.008, 2008.

Fohrer, N. and Schmalz, B.: Das UNESCO Oekohydrologie-Referenzprojekt Kielstau-Einzugsgebiet - Nachhaltiges Wasserressourcenmanagement und Ausbildung im laendlichen Raum, Hydrol. Wasserbewirts., 4, 160–168, doi:10.5675/HyWa_2012,4_1, 2012.

Fohrer, N., Schmalz, B., Tavares, F., and Golon, J.: Modelling the landscape water balance of mesoscale lowland catchments considering agricultural drainage systems, Hydrol. Wasserbewirts., 51, 164–169, 2007.

Fohrer, N., Dietrich, A., Kolychalow, O., and Ulrich, U.: Assessment of the Environmental Fate of the Herbicides Flufenacet and Metazachlor with the SWAT Model, J. Environ. Qual., 42, 1–11, doi:10.2134/jeq2011.0382, 2013.

Garambois, P. A., Roux, H., Larnier, K., Castaings, W., and Dartus, D.: Characterization of process-oriented hydrologic model behavior with temporal sensitivity analysis for flash floods in Mediterranean catchments, Hydrol. Earth Syst. Sci., 17, 2305–2322, doi:10.5194/hess-17-2305-2013, 2013.

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

Guse, B., Reusser, D. E., and Fohrer, N.: How to improve the representation of hydrological processes in SWAT for a lowland catchment - Temporal analysis of parameter sensitivity and model performance, Hydrol. Process., 28, 2651–2670, doi:10.1002/hyp.9777, 2014.

Herbst, M., Gupta, H. V., and Casper, M. C.: Mapping model behaviour using Self-Organizing Maps, Hydrol. Earth Syst. Sci., 13, 395–409, doi:10.5194/hess-13-395-2009, 2009.

Herman, J. D., Reed, P. M., and Wagener, T.: Time-varying sensitivity analysis clarifies the effects of watershed model formulation on model behavior, Water Resour. Res., 49, doi:10.1002/wrcr.20 124, doi:10.1002/wrcr.20124, 2013.

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., Freer, J., Savenije, H., and Gascuel-Odoux, C.: Process consistency in models: The importance of system signatures, expert knowledge, and process complexity, Water Resources Research, 50, 7445–7469, doi:10.1002/2014wr015484, 2014.

Huang, S., Krysanova, V., Österle, H., and Hattermann, F. F.: Simulation of spatiotemporal dynamics of water fluxes in Germany under climate change, Hydrol. Process., 24, 3289–3306, doi:10.1002/hyp.7753, 2010.

Kiesel, J., Schmalz, B., and Fohrer, N.: SEPAL - A simple GIS-based tool to estimate sediment pathways in lowland catchments, Advances in Geosciences, 21, 25–32, doi:10.5194/adgeo-21-25-2009, URL http://dx.doi.org/10.5194/adgeo-21-25-2009, 2009.

Kiesel, J., Fohrer, N., Schmalz, B., and White, M. J.: Incorporating landscape depressions and tile drainages of a northern German lowland catchment into a semi-distributed model, Hydrol. Process., 24, 1472–1486, doi:10.1002/hyp.7607, 2010.

LVermA: Landesvermessungsamt Schleswig-Holstein Digitales Geländemodell fuer SchleswigHolstein. Quelle: TK25. Gitterweite 25 m x 25 m und TK50 Gitterweite 50 m x 50 m sowie ATKIS-DGM2-1 m x 1 m Gitterweite und DGM 5 m x 5 m Gitterweite, abgeleitet aus LiDAR-Daten, 1995.

Martinkova, M., Hesse, C., Krysanova, V., Vetter, T., and Hanel, M.: Potential impact of climate change on nitrate load from the Jizera catchment (Czech Republic), Phys. Chem. Earth, 36, 673–683, doi:10.1016/j.pce.2011.08.013, URL http://dx.doi.org/10.1016/j.pce.2011.08.013, 2011.

Massmann, C., Wagener, T., and Holzmann, H.: A new approach to visualizing time-varying sensitivity indices for environmental model diagnostics across evaluation time-scales, Environ. Model. Softw., 51, 190–194, doi:10.1016/j.envsoft.2013.09.033, 2014.

Neitsch, S. L., Arnold, J. G., Kiniry, J. R., and Williams, J. R.: SWAT Theoretical Documentation Version 2009, Grassland, Soil and Water Research Laboratory, Agricultural Research Service. Blackland Research Center, Texas Agricultural Experiment Station, 2011.

Nossent, J., Elsen, P., and Bauwens, W.: Sobol' sensitivity analyses of a complex environmental model, Environ. Model. Softw., 26, 1515–1525, doi:10.1016/j.envsoft.2011.08.010, 2011.

Orlowsky, B., Gerstengarbe, F. W., and Werner, P. C.: A resampling scheme for regional climate simulations and its performance compared to a dynamical RCM, Theor. Appl. Climatol., 92, 209–223, doi:10.1007/s00704-007-0352-y, 2008.

Pfannerstill, M., Guse, B., and Fohrer, N.: A multi-storage groundwater concept for the SWAT model to emphasize nonlinear groundwater dynamics in lowland catchments, Hydrol. Process., 28, 5599–5621, doi:10.1002/hyp.10062, 2014a.

Pfannerstill, M., Guse, B., and Fohrer, N.: Smart low flow signature metrics for an improved overall performance evaluation of hydrological models, J. Hydrol., 510, 447–458, doi:10.1016/j.jhydrol.2013.12.044, 2014b.

R Core Team: R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013, 2013.

Reusser, D.: Implementation of the Fourier Amplitude Sensitivity Test (FAST), R-package, 0.61, 2012.

Reusser, D., Blume, T., Schaefli, B., and Zehe, E.: Analysing the temporal dynamics of model performance for hydrological models, Hydrol. Earth Syst. Sci., 13, 999–1018, doi:10.5194/hess-13-999-2009, 2009.

Reusser, D. E. and Zehe, E.: Inferring model structural deficits by analyzing temporal dynamics of model performance and parameter sensitivity, Water Resour. Res., 47(7), doi:10.1029/2010WR009 946, 2011.

Reusser, D. E., Buytaert, W., and Zehe, E.: Temporal dynamics of model parameter sensitivity for computationally expensive models with FAST (Fourier Amplitude Sensitivity Test), Water Resour. Res., 47(7), doi:10.1029/2010WR009 947, 2011.

Saltelli, A. and Bolado, R.: An alternative way to compute Fourier amplitude sensitivity test (FAST), Comput. Stat. Data Anal., 26(4), 445–460, doi:10.1016/S0167-9473(97)00043-1, 1998.

Saltelli, A., Ratto, M., Tarantola, S., and Campolongo, F.: Sensitivity analysis practices: Strategies for model-based inference, Reliab. Eng. Syst. Safe., 91 (10-11), 1109–1125, doi:10.1016/j.ress.2005.11.014, 2006.

Schmalz, B., Springer, P., and Fohrer, N.: Interactions between near-surface groundwater and surface water in a drained riparian wetland., in: Proceedings of International Union of Geodesy and Geophysics XXIV General Assemble" A New Focus on Integrated Analysis of Groundwater/Surface Water Systems", Perugia, Italy, 11-13 July 2007., pp. 21–29, IAHS Press, 2008.

Sieber, A. and Uhlenbrook, S.: Sensitivity analyses of a distributed catchment model to verify the model structure, J. Hydrol., 310(1-4), 216–235, doi:10.1016/j.jhydrol.2005.01.004, 2005.

Sudheer, K. P., Lakshmi, G., and Chaubey, I.: Application of a pseudo simulator to evaluate the sensitivity of parameters in complex watershed models, Environ. Model. Softw., 26, 135–143, doi:10.1016/j.envsoft.2010.07.007, 2011.

van Werkhoven, K., Wagener, T., Reed, P., and Tang, Y.: Sensitivity-guided reduction of parametric dimensionality for multi-objective calibration of watershed models, Adv. Water Resour., 32, 1154–1169, doi:10.1016/j.advwatres.2009.03.002, 2009.

Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V., and Sorooshian, S.: A framework for development and application of hydrological models, Hydrol. Earth Syst. Sci., 5, 13–26, doi:10.5194/hess-5-13-2001, 2001.

Wagener, T., McIntyre, N., Lees, M., Wheater, H., and Gupta, H.: Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis, Hydrol. Process., 17, 455–476, doi:10.1002/hyp.1135, 2003.

Wagener, T., Reed, P., van Werkhoven, K., Tang, Y., and Zhang, Z.: Advances in the identification and evaluation of complex environmental systems models, J. Hydroinform., 11, 266, doi:10.2166/hydro.2009.040, 2009.

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resour. Res., 44, W09 417, doi:10.1029/2007WR006716, 2008.

4.11 Supplement: The original groundwater module of the SWAT model

The detailed process equations of the groundwater processes for the original SWAT model are summarized in Neitsch et al. (2011). In the following description, the main groundwater processes of the original SWAT version are shown. For this, we focus on the equations that control the groundwater contribution to the discharge of the river.
In the SWAT model, soil water percolates into a conceptual aquifer structure, which is divided into a shallow and a deep aquifer. The shallow aquifer represents an unconfined aquifer that may discharge into the channel. The deep aquifer is described as a confined aquifer. As a consequence, the deep aquifer does not contribute to the streamflow within the watershed. Thus, the deep aquifer is considered as inactive, because it does not deliver water back to the modeled catchment.
Water that percolates out of the soil is delayed with an exponential delay function before it recharges ($w_{rchrg,i}$ mm H_2O) the groundwater system (Eq. 24, cf. Neitsch et al. (2011)):

$$w_{rchrg,i} = (1 - exp[\frac{-1}{\delta_{gw}}]) \cdot w_{seep} + exp[\frac{-1}{\delta_{gw}}] \cdot w_{rchrg,i-1} \tag{24}$$

The parameter δ_{gw} (days) describes the delay of recharge that accounts for geologic formations. Eq. 24 considers the percolation out of the last soil layer on day i (w_{seep}, mm H_2O) and the parameter $w_{rchrg,i-1}$ (mm H_2O) which represents the amount of water that enters the aquifer on the day before $(i-1)$.
With Eq. 25 (cf. Neitsch et al. (2011)), the daily recharge $w_{rchrg,i}$ is partitioned by a percolation coefficient β_{dp} (-) to calculate the recharge of the deep aquifer ($w_{seep,dp,i}$, mm H_2O):

$$w_{seep,dp,i} = \beta_{dp} \cdot w_{rchrg,i} \tag{25}$$

The remaining amount of recharge ($w_{rchrg,sh,i}$, mm H_2O) that enters the shallow aquifer is then calculated with Eq. 26 (cf. Neitsch et al. (2011)):

$$w_{rchrg,sh,i} = w_{rchrg,i} - w_{seep,dp,i} \tag{26}$$

In the SWAT model, groundwater contribution to the river ($Q_{gw,i}$, mm H_2O) is simulated with the shallow aquifer (Eq. 27, cf. Neitsch et al. (2011)). The recharge of the shallow aquifer ($w_{rchrg,sh,i}$) is used together with the parameter α_{gw} (1/days) and a one day time step (Δt) to describe a recession-based outflow out of the aquifer storage.

$$Q_{gw,i} = Q_{gw,i-1} \cdot exp[-\alpha_{gw} \cdot \Delta t]) + w_{rchrg,sh,i} \cdot (1 - exp[-\alpha_{gw} \cdot \Delta t]) \tag{27}$$

5 Summarising discussion and conclusion

5.1 Summary of key achievements

The reliablity of models to investigate the hydrological system strongly depends on the ability of the model structure to capture the relevant hydrological processes of the study catchment. This ability can be tested by (i) applying appropriate evaluation criteria during model calibration and (ii) by ensuring coincidence between simulated and observed hydrological processes. This thesis demonstrates how this goal can be achieved by making use of diagnostic model analyses.
The motivation of this thesis is based on a model structure deficiency of the SWAT model that was identified by Guse et al. (2014) with TEDPAS and TIGER. The groundwater model component with its parameters were found to be the main reason for poor model performance at low flow phases.
The main aim of this thesis was the development of a guideline for hydrologically consistent models, which is the synthesis of three research questions. The core results of the three research questions that were raised to build up the proposed guideline are answered in a summary to finally resolve the answer for the main question.

- **How can deficient model structures be improved with information of diagnostic model analyses?**

The findings of several studies that reported poor model performance for the low flow phases of SWAT and the diagnostic model analyses of Guse et al. (2014) were utilised to adapt the model structure of the SWAT groundwater component. As described in Chapter 2, the presented model modification emphasises the non-linearity of the SWAT groundwater component by integrating an additional groundwater storage. The new SWAT$_{3S}$ version integrates two active groundwater storages that contribute to the discharge and one inactive groundwater storage that accounts percolation into deep geologic formations.
The raised research question of how the information of diagnostic model analyses can be used to improve deficient model structures was answered by realising a more suitable groundwater description with SWAT$_{3S}$. Based on a literature review, the common recommendation of incorporating multiple storages into hydrological models to describe non-linear groundwater processes was transferred to the SWAT model. According to this recommendation, a more complex storage structure with an improved suitability for lowland catchments was achieved. The results of this modification show the improvement of low flow reproduction by SWAT$_{3S}$.
The developed groundwater component of SWAT$_{3S}$ is characterised by a flexible structure that allows consideration of spatial heterogeneity of groundwater processes and the description of groundwater processes itself. Individual parameter settings can be used to adapt the simulated groundwater processes to the catchments (e.g. emphasised groundwater relevance vs. no groundwater contribution). Furthermore, groundwater aquifers can be activated and deactivated on the subcatchment and HRU level to account for heterogeneity within large catchments.
Referring to the performance evaluation of the original SWAT and the new SWAT$_{3S}$, common performance measures (NSE, PBIAS) were applied. Furthermore, the presented study of Chapter 2 describes the utilisation of the FDC to evaluate different discharge phases. The applied evaluation method provided a first possibility to determine model performance for the mid flow and the low flow segment of the FDC. However, since the adapted groundwater component of SWAT$_{3S}$ aims to improve the low flow reproduction, the need to determine the model performance for low flow in a more detailed way was emphasised. Based on this finding further research to address combined evaluation criteria that allow the detailed evaluation of the low flow phases together with the remaining phases of the hydrograph was motivated. Consequently, the second questions focused on:

- **How can modified model structures be evaluated by considering all relevant discharge phases?**

Chapter 3 describes the evaluation of the modified groundwater component of $SWAT_{3S}$ with a newly developed multi-metric framework. This framework integrates metrics for different discharge phases to achieve a satisfying reproduction of the overall discharge. According to Gupta et al. (1998), Boyle et al. (2000), Madsen (2000), Wagener and Gupta (2005) and Gupta et al. (2008) the application of one single evaluation criterion is insufficient to take into account the representation of all relevant processes. As a consequence, the newly developed multi-metric framework makes use of a stepwise intersection of best model runs that are determined separately for each performance and signature metric to make sure that all phases of the hydrograph are considered.
As the low flow phases were found to be an important indicator for the proper simulation of groundwater contribution to the river, the development of the multi-metric framework integrates additional FDC segments in comparison to the common FDC segmentation (e.g. Yilmaz et al., 2008). Low flow phases are further separated into a low flow segment and a very low flow segment that considers the flow exceedance probability between 95 % and 100 %. In this way, the model performance evaluation for discharge volume reproductions at low flow phases is improved. Regarding the common FDC segmentation with a very high flow segment with the flow exceedance probability of 5 %, extreme events of very high and very low flow are equally weighted.
The main outcome of the multi-metric evaluation is that the application of FDC segments is a helpful strategy to evaluate the model performance for the whole discharge period including very high and very low flows. As it requires minimal effort for a more accurate model evaluation, this method is proposed to be integrated into model performance analysis. With respect to the evaluation of the modified model structure of $SWAT_{3S}$, the additional segmentation of the FDC with very low and low performance metrics improves the evaluation of low flows.
Despite of improved evaluation of simulated discharge phases, satisfying model performance for discharge reproduction does not automatically guarantee hydrologically realistic process reproduction. According to a more processed based evaluation of model behaviour as proposed by Yilmaz et al. (2008) and Pokhrel et al. (2012), further diagnostic analysis are recommendable. For this, additional diagnostic information needs to be extracted from simulated and observed hydrological data to evaluate hydrological models in a way that goes beyond model performance only for discharge simulation. This finding leads to the third question:

- **How can modified model structures be verified with diagnostic model analyses and expert knowledge?**

As shown in Chapter 3, the application of a multi-metric framework is useful to consider all phases of the hydrograph and consequently it is determined if a (modified) model structure is able to simulate hydrological characteristics (e.g. discharge) properly. However, as mentioned by Clark et al. (2008) and Hrachowitz et al. (2014) the appropriate discharge simluation does not guarantee for appropriate reproduction of hydrological processes. For this, Chapter 4 proposes the analysis of the modified $SWAT_{3S}$ groundwater model component in regard to two different aspects: (i) the hydrological consistency within the model and (ii) the consistent simulation of hydrological processes for the study catchment. In a first step, TEDPAS was used to analyse the modified groundwater model component with respect to the implemented model parameters and their temporal relevance. Hypotheses about the temporal parameter sensitivity and their temporal sequence were derived from the modified model structure. The results of TEDPAS for the groundwater model component confirmed the expected sequence of parameter sensitivity: parameters of the fast reacting groundwater storage were sensitive before the parameters of the slow reacting groundwater storage. In the second step, TEDPAS was

used to analyse the process sequence according the hypotheses of vertical water redistribution concept and additional expert-knowledge of hydrological processes within the study catchment. The results revealed a simulated process sequence that is consistent with the derived hypotheses for the study catchment. The simulated process sequence consistently exhibited the order with surface runoff as first process, followed by tile drainage. This process sequence continued with fast groundwater flow and slow groundwater flow. Finally the simulated process sequence showed evapotranspiration and soil water storage as the last relevant process.
Considering the third question, the results of TEDPAS were combined with the knowledge about the hydrological processes and their occurrence timing for the study catchment. The derived hypotheses were used to evaluate the modified model structure in a qualitative way. By proving the coincidence between simulated and hypothesised sequence of parameter sensitivity and hydrological process sequence, the presented approach can be intepreted to a final step of hydrological model improvements. With this step, the hydrological consistency is confirmed. Finally all three questions may be combined to answer the main research question of this thesis:

- **How can hydrologically consistent models be achieved?**

The results that were obtained from the three research questions are transferred into a more general context to propose the demonstrated procedure as a general guideline for hydrologically consistent models. Diagnostic model analyses can be interpreted as the foundation of this guideline. As exemplarily shown for the SWAT groundwater component, diagnostic model analyses may be used to identify model structure deficiencies that need improvement for satisfying model performance. Chapter 2 illustrates how the information of model diagnostic analyses can be transformed into the idea for a model structure improvement.
However, the modifications of individual model components need further analysis to make sure that the improvements lead to improved model performance together with hydrologically consistent simulation of the hydrological processes. For this, the thesis proposes two different approaches. In a first step, the modified model is evaluated with appropriate performance and signature metrics. A prerequisite for this evaluation is the selection of proper evaluation criteria as shown in Chapter 3. Additionally, the hydrological consistent simulation of the hydrological processes needs to be verified. The consistency is verified with hypotheses about the parameter relevance and the hydrological process sequence for specific discharge phases of the study catchment. This approach makes use of TEDPAS which can be assigned to tools of diagnostic model analyses. The provided results allow the qualitative comparison between simulated and hypothesised sequence of parameter sensitivity and observed hydrological processes of the study catchment.
Complying the proposed steps of model improvement due to identified model structure deficiencies, performance evaluation and process verification, the guideline leads to a hydrologically consistent model. In general, this guideline can be applied in principal to any hydrological model in any catchment.

5.2 Discussion

According to the summarised main achievements of this thesis, the results of Chapter 2, 3, and 4 are discussed in a more general context. Finally, the discussion provides the foundation for the synthesis that leads to the main outcome of this thesis: A guideline for hydrologically consistent models. The thesis is then closed by pointing out aspects that are recommendable for future research.

5.2.1 The modified groundwater structure of SWAT$_{3S}$ as an example for model improvements

The modification of the groundwater component for the new SWAT$_{3S}$ was motivated by the analyses of Guse et al. (2014) with TEDPAS and TIGER. This diagnostic model analysis and the following interpretation of diagnostic information to derive requirements for model modifications agree with recommendations of Gupta et al. (2008). As proposed by Gupta et al. (2008), diagnostic model analyses should clearly depict the aspects of a model that need to be improved and how the model should be improved. According to this recommendation, Chapter 2 describes the development of a more complex groundwater component incorporating knowledge and theories about groundwater processes. According to the presented results of this thesis and according to the application by Haas et al. (2015), the groundwater structure of SWAT$_{3S}$ with two contributing storages was proved to better capture complex groundwater processes, especially in lowlands. There is a clear separation of groundwater storage activity for the study catchments: the fast reacting groundwater flow controls the recession phase. By constrast, the slow groundwater flow controls the baseflow with its delayed answer of groundwater recharge, which can be further controlled with a loss to the deep aquifer.
The newly developed groundwater structure leads to an improved discharge simulation, but there are still some aspects that need further discussion. Despite of the emphasised complexity of SWAT$_{3S}$, the model is still characterised by a conceptual model structure with simplified process descriptions. As shown by Orth et al. (2015) an increased complexity may lead to over-parameterisation but a too simple model structure may suffer from an incomplete representation of relevant processes. Consequently, the models need to have simple structures but adequate complexity (Orth et al., 2015). Referring to the new groundwater structure of SWAT$_{3S}$ it has to be discussed if the increased complexity is justifiable with respect to the available information that are provided by knowledge of the catchment and model input data.
Depending on the knowledge of the catchment and available model input data, the different groundwater storages of SWAT$_{3S}$ may be activated or deactivated. This flexibility takes up the idea of flexible model structure approaches for hydrological modeling as introduced by Fenicia et al. (2011). The introduced framework of Fenicia et al. (2011) proposes the selection of appropriate model structures with respect to the known processes within the study catchment. In this context, the opportunity of SWAT$_{3S}$ to depict groundwater processes with adapted model structures provides a high degree of flexibility. SWAT$_{3S}$ is able to represent groundwater processes with one or two storages that may contribute to the river. Furthermore, groundwater percolation into deep geologic formations may be considered with a third storage. Consequently, the modeler is free to decide which degree of complexity for the groundwater and how many model parameters are needed. As shown in this thesis and by Haas et al. (2015), the groundwater structure of two contributing storages and one storage for percolation into deep geologic formations is favourable to depict the nonlinear processes of typical lowlands. However, other catchments may not need this complex groundwater structure due to less complex groundwater processes. In this case, the number of used groundwater storages and consequently the number of parameters may be reduced. To conclude, SWAT$_{3S}$ provides opportunities to apply more complex model structures if needed. However, this flexibility increases the responsibility of the modeler. The complexity of the applied groundwater structure need to be selected carefully to prevent unrealisitic process simulation.
In this context, Hrachowitz et al. (2014) found out that increasing model complexity does not necessarily leads to improved model performance and process simulation. This is especially the case when the parameter values for complex model structures are not reasonably constrained. As a conclusion, increased complexity of model structures should not be the ultimate goal of model development. De-

pending on available information of landscape features and knowlegde about hydrological processes within the catchment, it has to be ensured that the model structure suits to the catchment (Euser et al., 2013; Hrachowitz et al., 2014). Additionally, the main processes have to be captured while maintaining minimum levels of complexity (Fenicia et al., 2008).
Considering the hydrological processes of the catchment, the complexity can be further interpreted in a spatial point of view. For example, a typical pathway for water flow is driven by gradients from hills to valleys. Additionally there are interactions between the river and the subsurface system of the valley. The HRU concept of SWAT does allow the integration of landscape heterogeneity during the model-setup and by setting parameter values according to expected groundwater processes. In this way, different characteristics that are related to groundwater processes (e.g. soil permeability, tile drainages, slope) may be taken into account at the HRU generation to emphasise spatial complexity. Recent developments of SWAT lead to integrations of landscape routing options (Volk et al., 2007; Bosch et al., 2010; Arnold et al., 2010) and grid-based model setups (Rathjens et al., 2015). However, the advantage of integrating more spatial complexity is often limited due to inavailability of appropriate data. Consequently, the modeller has to make sure that the complexity of the applied model structure matches the available data to derive reasonable spatial distributed parameter values and to achieve reasonable process simulation.
Referring to reasonable process simulation, this thesis points out the advantage of $SWAT_{3S}$ for the study catchment since improved discharge reproduction was achieved with emphasised groundwater relevance on the discharge. The presented application of $SWAT_{3S}$ revealed a major improvement that can be identified for the simulation of the low flow segment with adequate representation of the mid flow segment. The groundwater structure of two contributing groundwater storages is essential for the study catchment to depict the groundwater dominance which is typical for lowland catchments. The diagnostic information that is provided by Guse et al. (2014) was used to improve the model through the inclusion of particular processes as recommended by Orth et al. (2015). To investigate the applicability and to proof the flexibility of $SWAT_{3S}$ to represent contrary hydrological conditions, additional applications in catchments with other hydrological characteristics are recommendable.

5.2.2 Diagnostic model analyses - the key to hydrologically consistent models?

To perform diagnostic model analyses, several methods and tools are available to extract information about model behaviour and to identify the reasons for specific behaviour (e.g. Wagener et al., 2003; Clark et al., 2008; Fenicia et al., 2008; Herbst et al., 2009; Clark et al., 2011; Massmann and Holzmann, 2012; Euser et al., 2013; Garambois et al., 2013). In this thesis, it is shown how diagnostic model analyses help to achieve hydrologically consistent model improvements making use of model performance evaluation and temporal parameter sensitivity analysis. The applied temporal parameter sensitivity analysis was found to be a valuable method to extract diagnostic information. As shown by Haas et al. (2015), the application of temporal parameter sensitivity analyses is not limited to hydrological aspects. The study of Haas et al. (2015) clearly showed, that this kind of diagnostic model analysis is helpful to understand nitrate processes of $SWAT_{3S}$.
To apply diagnostic model analyses, multiple evaluation criteria can be used to compare simulated and observed hydrological processes. This kind of model evaluation aims to identify specific phases of the hydrograph or specific hydrological processes that are poorly reproduced by the model. Considering the provided example of Guse et al. (2014), the poor model performance is systematically evaluated to extract additional diagnostic information. Referring to the investigation of very low flow phases, this thesis provides a newly developed evaluation criterion for this specific discharge phase. Since the very low flow is of high interest such as for water resource management or eco-hydrology (e.g. Smakhtin, 2001; Laaha and Blöschl, 2006) the newly developed evaluation criterion particulary helps

to investigate the model performance for this specific discharge phase.
However, it has to be mentioned that the application of the newly developed evaluation criterion is just one step in diagnostic model analyses. Beside of model evaluation, additional methods need to be applied to extract information about accurate process integration within the model structures. Temporal parameter sensitivity analyses on the model's output (e.g. discharge) provide further insights into possible structural model deficits (e.g. Guse et al., 2014). In this way, it is investigated which parameter controls which discharge phase. Considering the motivation and the presented studies of this thesis, the relevance of temporal parameter sensitivity analysis is highlighted. The preliminary work of Guse et al. (2014) revealed that the groundwater parameters were highly sensitive on the discharge for the low flow phases. On the modeler's perception these results are in accordance with the intended model behaviour to represent hydrological processes of a lowland. Despite of reasonable temporal parameter sensitivity, the applied model was not able to simulate the baseflow driven low flow phases properly. The groundwater parameters were found to have the highest potential for model improvement by improving the groundwater process representation. Consequently, TEDPAS is a recommendable method to identify model structure deficiencies.
However, the role of TEDPAS for diagnostic model analyses is not limited to the identification of model structure deficiencies. In Chapter 4 of this thesis, TEDPAS was further developed to a method that provides information to investigate the hydrological consistency of a modified model. After modifying the model component that is responsible for poor model performance, the process representation was improved. The improvement was quantified and verified by integrating the newly developed evaluation criterion for very low flows. Nonetheless, satisfying model performance for one hydrological process (e.g. discharge) is not always related to a high consistency if other processes are inappropriately represented (Yilmaz et al., 2008; Euser et al., 2013; Hrachowitz et al., 2014). Consequently, in following diagnostic model analyses the ultimate goal was to ensure hydrological consistency for the modified model to make sure that the model results are improved for the right reason (Clark et al., 2008; Martinez and Gupta, 2011; Euser et al., 2013). Again it was investigated if all model parameters are relevant at the desired time and if the hydrological process are properly simulated. According to Gupta and Nearing (2014), this analysis aims to investigate if consistency between the model and the real world is achieved (Gupta et al., 1998, 2008; Martinez and Gupta, 2011).
The decision if the hydrological processes of the real world have been achieved by the applied model requires high diagnostic information content that demands for more power in identification methods (Wagener and Gupta, 2005). In this context, future developments of diagnostic model analyses that were applied in this thesis need to be discussed. The demonstrated TEDPAS application within this thesis provided daily time series of partial parameter sensitivity on the discharge. Individual parameter sensitivities of model components were analysed and parameter sensitivities were interpreted into simulated processes. However, up to now there are more opportunities to increase the information content of the temporal parameter sensitivities. A first step is to consider different temporal resolutions as presented by Herman et al. (2013) and Massmann et al. (2014). With these approaches it is highlighted that individual processes are relevant at different time scales. Consequently, the parameter sensitivity is expected to be observed according to the individual time scale. In addition, there is room for further visualisation techniques. As presented in this thesis, the evaluation is based on daily time series. Future work could focus on the visualisation of temporal parameter sensitivity with respect to different discharge events or discharge magnitudes together with seasonal aspects of the discharge magnitude to extract information about seasonality of parameter sensitivity more clearly.
Discussing the model improvement that was presented in this thesis, the newly developed evaluation framework with an newly developed low flow evaluation criterion and the investigated temporal parameter sensitivity analysis suggest hydrologically consistent model results. As a conclusion, the

raised question if diagnostic model analyses can be interpreted to the key for hydrologically consistent models can be positively answered in terms of model performance and parameter timing. However, this conclusion relies to a huge extent on the simulated discharge and on the modeler's perception, when parameters should be sensitive. As shown in Chapter 4, the additional utilisation of real world observations provides further crucial information for the decision if hydrological consistency has been achieved or not. The knowledge of hydrological processes of the real world can lead to a more well-founded final decision. For this, a way of integrating expert-knowledge into diagnostic model analyses is needed.

5.2.3 Advanced diagnostic model analyses with expert-knowledge

Model evaluation and verification is preferably performed with measured data, which comprises mainly the easily available discharge data for the studied catchment. Based on the discharge, the most appropriate model structures for a given problem are identified (e.g. Clark et al., 2008; Fenicia et al., 2008; Stoelzle et al., 2015) by analysing the reliability of model results with diagnostic analysis of the model structures. However, considering the discharge as evaluation and verification criterion limits the validity of the performed diagnostic model analysis. Of course, the ultimate goal should be the integration of additional measured data such as soil water content, groundwater levels or the separated discharge components. For example, Rathjens et al. (2015) recommend the utilisation of remote sensing data information such as evapotranspiration or soil water content to validate the grid-based landscape SWAT version. The separation of discharge components with the help of tracers or diatoms (e.g. Klaus et al., 2015) may help to determine the appropriateness of the different modelled discharge components. Especially for the cases of model structure modifications, the inclusion of additional observational information is essential (Krause et al., 2005). However, a spatial distribution of this additional and valuable information is often not available since gathering this data is expensive and unaffordable even for small catchments.

Considering this limitation, solutions have to be found to analyse models and model structures beyond discharge reproduction. The first step of improving the diagnostic model analyses is achieved by extracting as much information as possible from simulated and observed data. As mentioned by Rathjens et al. (2015), rapidly developing techniques of geographical information systems and remote sensing provide an increasing amount of spatially and temporally detailed data. For example, remotely sensed land cover data may be used together with observed land cover data to interpolate data gaps to finally obtain a complete time series of land cover change (Rathjens et al., 2014).

However, it has to be questioned if these approaches are able to provide a complete data set about the hydrological processes within the catchment. A detailed investigation of the model behaviour provides a first idea of how processes are simulated. For this, performance metrics can be linked to hydrological processes (e.g. Yilmaz et al., 2008; Hrachowitz et al., 2014). This approach is especially important to make sure that the well reproduced discharge was achieved by reproducing the hydrological processes properly (Clark et al., 2008). However, this linkage between hydrological processes and performance metrics is not the common practice. Up to now, there are no general recommendations for a comprehensive library of performance metrics that are capable to evaluate hydrological process reproduction. Due to this limitation, the certainty about a well reproduced hydrological system is still limited. For example, temporal aspects for the hydrological process occurrence are important to make a final decision about the hydrologically consistent model improvement. Up to this point, the decision for hydrologically consistent model improvement is derived by considering the real world with respect to quantification of reproduced discharge magnitude and overall timing.

For the case that additional measured data is unavailable, it needs to be discussed how the decision about hydrological consistency can be further assisted. Seibert and McDonnell (2002) and Hrachowitz

et al. (2014) propose the integration of soft data into the modeling process, which can be defined as qualitative data that is derived from visual observation of hydrological processes within the catchment. According to Chapter 4, the proposed concept of qualitative data can be further interpreted to a hypothesis test for TEDPAS-based diagnostic model analyses. According to Clark et al. (2011) hypotheses about the processes within the catchment are in this context not a formal statistical test but a qualitative evaluation of expert knowledge.
The idea of the presented TEDPAS-based diagnostic model analysis is based on the assumption that patterns of temporal parameter sensitivity can be intepretetd to a simulated process. However, the innovation of Chapter 4 is the linkage between simulated hydrological processes and the observed hydrological processes of the real world. In this context, observed hydrological processes of the real world are defined as expert-knowledge of the catchment that is derived from previous studies. Despite of the expert-knowledge of process occurrence, the provided information remains qualitative since the process occurrence is not assigned to a certain date of the discharge series but assigned to a hydrological condition of the catchment as a result of pre-conditions and events within the catchment.
Although this expert-knowledge has a qualitative character, it is valuable information as shown in model calibration approaches that aimed for improved model performance together with improved hydrological consistency (e.g. Seibert and McDonnell, 2002; Yilmaz et al., 2008; Hrachowitz et al., 2014). This concept was further developed in this thesis (Chapter 4) by verifying the ability of a model to reproduce and describe the temporal aspects of hydrological processes within the catchment. At the same time, it is verified that the model structure is suitable to the study catchment.
This idea of TEDPAS-based model analyses is in the same line with the examples provided by Reusser et al. (2011), Massmann and Holzmann (2012), Herman et al. (2013) and Guse et al. (2014) who interpreted the visualised temporal parameter sensitivities on the discharge. A further development was presented by Massmann and Holzmann (2012) by applying temporal parameter sensitivity analysis on additional model outputs such as discharge components. This is a first step to increase the information content that is extracted out of the model and in accordance with Gupta et al. (1998) who propose the evaluation of several output fluxes of the model to identify shortcomings in representing hydrological processes. However, it still has to be mentioned that the results of this interpretation are still limited in their usability as long as measurements are not available for this model output.
To conclude, expert-knowledge can be used for model diagnostics to overcome the lack of measured data within the catchment. Observed events and hydrological processes are intepreted into model verification data to compare simulated and observed behaviour of the hydrological system. In this way, the hydrological consistency of the model can be determined in a qualitative way.

5.2.4 Synthesis - a guideline for hydrologically consistent model improvements

Based on the main achievements and preliminary studies, a synthesis is carried out to derive a general guideline for hydrologically consistent model improvements. According to the presented studies and conclusions, different steps were identified for this guideline comprising model failure detection and model modification followed by model evaluation and the final verification of hydrological consistency (Fig. 22) The first step of this guideline is focused on the model failure detection aiming to identify the model component that needs structural improvements (e.g. Reusser and Zehe, 2011; Guse et al., 2014). In the particular case, TIGER and TEDPAS are used to identify the reasons for poor model performance. These methods take up the ideas of Gupta et al. (1998) and Gupta et al. (2008), who propose additional information for diagnostic model analyses that is extracted from model output and related to real world processes. Consequently, models can be assessed to improve the understanding of the hydrological processes in the study catchment and their incorporation into models (Fenicia et al., 2008; Reusser et al., 2009).

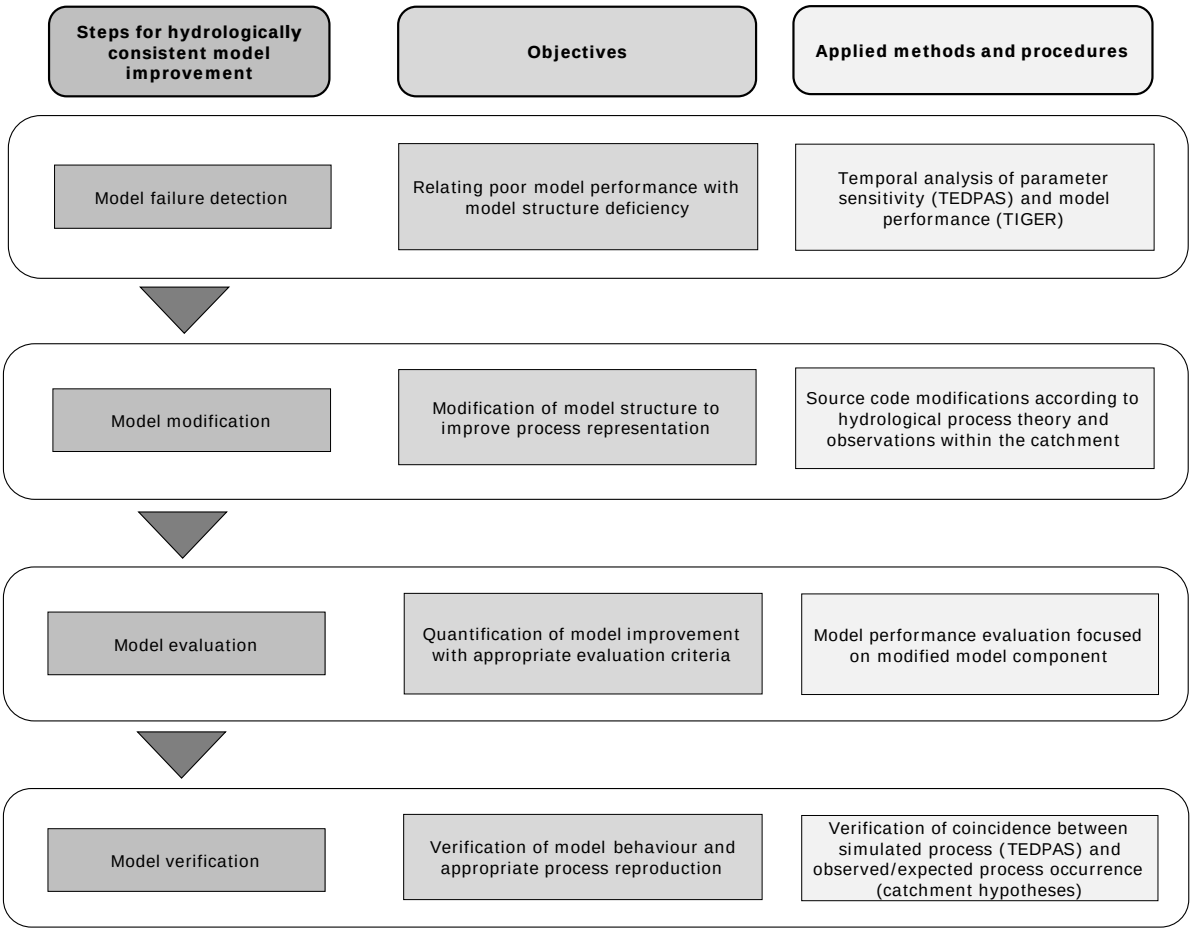

Figure 22: Steps for a hydrological consistent model improvement.

The second step of the guideline is based on the diagnostic information of TIGER and TEDPAS. As shown in Chapter 2, model structures with their equations and parameters are modified to overcome limitations of process representation. The provided example lead to increased complexity of the groundwater component of SWAT to realise a more complex process representation. In this way, the diagnostic information is used to improve the model through the inclusion of particular processes for a specific model component (Orth et al., 2015).
After modifying the model structure, the third step of the presented guideline needs to be applied. It has to be ensured that the modification leads to improved model results. For this, the model has to be evaluated with appropriate evaluation criteria. The evaluation should be preferably based on multiple criteria (Gupta et al., 1998; Krause et al., 2005) to evaluate the model performance for model output. Additionally, there should be a second focus on the ability of the model to capture the hydrological processes and functions of the catchment (Yilmaz et al., 2008; Sawicz et al., 2011).
In a last step, the proper reproduction of the hydrological processes is further investigated. This step verifies that hydrological consistency is achieved with good model performance and proper process simulation. According to Hrachowitz et al. (2014), this step follows the idea of a systematic use of hydrological signatures and expert knowledge to achieve model consistency. Observed processes of the catchment are interpreted as expert-knowledge to compare simulated and observed hydrological processes in a qualitative way. In this way, it is verified that the improved model performance of the modified model is achieved by reproducing the hydrological processes of the catchment.

With respect to consistency in hydrological modelling, the approaches of Euser et al. (2013) and Hrachowitz et al. (2014) are focused on model calibration and testing of model structure suitability for the catchment. However, the new guideline can be seen as a more broader approach as it guides through a complete sequence starting with model structure identification and ending by the verification of hydrological consistency for a modified model structure. Consequently the proposed guideline can be interpreted as an additional component in diagnostic model analysis for the case that poor model performance is related to model structure deficiencies. Due to the general description of how the different steps of this guideline have to be applied, it can be hypothesised that it is applicable for any hydrological model in any catchment.

5.3 Further research questions

Based on the previous discussions, future research questions are derived in the following to encourage further applications and developments of $SWAT_{3S}$ and the presented diagnostic methods. According to the discussion, the demand and recommendation for further studies about the applicability of the new $SWAT_{3S}$ is highlighted. The modified groundwater component provides a highly flexible structure to depict temporal characteristics of groundwater processes and their spatial heterogeneity within the catchments. Up to now, $SWAT_{3S}$ was applied especially to lowland catchments. Further studies should focus on different catchments with groundwater processes that are different to lowland catchments. Furthermore, additional efforts to integrate spatial information about landscape features to finally derive spatial heterogeneity of groundwater processes are of high relevance to verify the flexibility of $SWAT_{3S}$.

To evaluate the model performance and reliability at low flow conditions, the newly developed performance metric explicitly focuses on the very low flow phase. The very low flow metric is based on discharge magnitude probabilities and highly relevant to ensure model reliability for drought and baseflow simulations. The presented multi-metric framework evaluates the model performance by weighting all discharge phases equally. The applicability of this framework would benefit from investigations of how this framework can be used in a more flexible way to emphasise specific discharge phases that might be of higher interest with respect to other discharge phases.

At the same time, it has to be mentioned that the very low flow reflects a catchment answer under specific hydrological conditions. These conditions are caused by specific hydrological processes. For further research, it is recommendable to link the very low flow segment of the FDC and the associated performance metric explicitly with the hydrological process to evaluate the simulated catchment behaviour. This is a step towards a more process-based model evaluation contrary to the discharge focused model evaluation.

In addition to the model evaluation, the presented TEDPAS application takes up the idea of a detailed process analysis. As shown in this thesis, TEDPAS can be used to verify a realistic model behaviour. However, there is still space for further developments. To perform a fully process-oriented diagnostic model analysis, it has to be questioned how the different temporal and spatial scales of hydrological processes can be considered. On the one hand, the investigated processes have to be analysed at the appropriate temporal resolution which might be influenced by the spatial extent where the process takes place. On the other hand, the investigated processes have to be analysed by the appropriate model ouput since discharge does not include all information to analyse each process in detail. Consequently, the ultimate goal is to better focus on the verification and evaluation of the simulated processes by using available information that is beyond discharge.

The integration of additional catchment information into model diagnostics was exemplarily shown with the proposed guideline of this thesis. Observed processes and hydrological conditions are utilised to increase the available information content for the diagnostic model analysis. In this context, the

proposed guideline can be seen as one aspect in model diagnostic analyses. It is hypothesised that the general procedure is applicable to any model and any catchment. For this, future work should focus on providing an open-source package that makes available all methods that are needed for the described structured guideline. This freely available package would support the needed testing of the hypothesis that the structured guideline can be used universally. Since testing of this hypothesis is limited for complex models due to huge data requirements and computational demands, the proof has to be performed in further studies with other hydrological models for other catchments.

5.4 References

Arnold, J. G. and Fohrer, N.: SWAT2000: Current capabilities and research opportunities in applied watershed modelling, Hydrological Processes, 19, 563–572, doi:10.1002/hyp.5611, 2005.

Arnold, J. G., Srinivasan, R., Muttiah, R. S., and Williams, J. R.: Large area hydrologic modeling and assessment part I: Model development, Journal - American Water Works Association, 1, 73–89, doi:10.1111/j.1752-1688.1998.tb05961.x, 1998.

Arnold, J. G., Allen, P. M., Volk, M., Williams, J. R., and Bosch, D. D.: Assessment of different representations of spatial variability on SWAT model performance, Transactions of the ASABE, 53, 1433–1443, doi:10.13031/2013.34913, 2010.

Bekele, E. G. and Nicklow, J. W.: Multi-objective automatic calibration of SWAT using NSGA-II, Journal of Hydrology, 341, 165–176, doi:10.1016/j.jhydrol.2007.05.014, 2007.

Bennett, N. D., Croke, B. F. W., Guariso, G., Guillaume, J. H. A., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T. H., Norton, J. P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A. A., Fath, B. D., and Andreassian, V.: Characterising performance of environmental models, Environmental Modelling & Software, 40, 1–20, doi:10.1016/j.envsoft.2012.09.011, 2013.

Beven, K.: Changing ideas in hydrology - The case of physically-based models, Journal of Hydrology, 105, 157–172, doi:10.1016/0022-1694(89)90101-7, 1989.

Beven, K. and Binley, A.: The future of distributed models: Model calibration and uncertainty prediction, Hydrological Processes, 6, 279–298, doi:10.1002/hyp.3360060305, 1992.

Borah, D. K. and Bera, M.: Watershsd-scale hydrologic and nonpoint-source pollution models: Review of mathematical bases, Transactions of the ASAE, 46, 1553–1566, doi:10.13031/2013.15644, 2003.

Bosch, D. D., Arnold, J. G., Volk, M., and Allen, P. M.: Simulation of a low-gradient coastal plain watershed using the SWAT landscape model, Transactions of the ASABE, 53, 1445–1456, doi:10.13031/2013.34899, 2010.

Boyle, D. P., Gupta, H. V., and Sorooshian, S.: Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, Water Resources Research, 36, 3663–3674, doi:10.1029/2000WR900207, 2000.

Bronstert, A., Kolokotronis, V., Schwandt, D., and Straub, H.: Comparison and evaluation of regional climate scenarios for hydrological impact analysis: General scheme and application example, International Journal of Climatology, 27, 1579–1594, doi:10.1002/joc.1621, 2007.

Cibin, R., Sudheer, K. P., and Chaubey, I.: Sensitivity and identifiability of stream flow generation parameters of the SWAT model, Hydrological Processes, 24, 1133–1148, doi:10.1002/hyp.7568, 2010.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., and Hay, L. E.: Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, Water Resources Research, 44, W00B02, doi:10.1029/2007WR006735, 2008.

Clark, M. P., McMillan, H. K., Collins, D. B. G., Kavetski, D., and Woods, R. A.: Hydrological field data from a modeller's perspective: Part 2: process-based evaluation of model hypotheses, Hydrological Processes, 25, 523–543, doi:10.1002/hyp.7902, 2011.

Conradt, T., Koch, H., Hattermann, F. F., and Wechsung, F.: Spatially differentiated management-revised discharge scenarios for an integrated analysis of multi-realisation climate and land use scenarios for the Elbe River basin, Regional Environmental Change, 12, 633–648, doi:10.1007/s10113-012-0279-4, 2012.

Dawson, C. W., Abrahart, R. J., and See, L. M.: : A web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts, Environmental Modelling & Software, 22, 1034–1052, doi:10.1016/j.envsoft.2006.06.008, 2007.

Duan, Q. Y., Gupta, V. K., and Sorooshian, S.: Shuffled complex evolution approach for effective and efficient global minimization, Journal of Optimization Theory and Applications, 76, 501–521, doi:10.1007/BF00939380, 1993.

Eckhardt, K., Haverkamp, S., Fohrer, N., and Frede, H. G.: SWAT-G, a version of SWAT99.2 modified for application to low mountain range catchments, Physics and Chemistry of the Earth, Parts A/B/C, 27, 641–644, doi:10.1016/S1474-7065(02)00048-7, 2002.

Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., and Savenije, H. H. G.: A framework to assess the realism of model structures using hydrological signatures, Hydrology and Earth System Sciences, 17, 1893–1912, doi:10.5194/hess-17-1893-2013, 2013.

Fenicia, F., McDonnell, J. J., and Savenije, H. H. G.: Learning from model improvement: On the contribution of complementary data to process understanding, 44, doi:10.1029/2007WR006386.

Fenicia, F., Savenije, H. H. G., Matgen, P., and Pfister, L.: Understanding catchment behavior through stepwise model concept improvement, Water Resources Research, 44, W01 402, doi:10.1029/2006WR005563, 2008.

Fenicia, F., Kavetski, D., and Savenije, H. H. G.: Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development, Water Resources Research, 47, W11 510, doi:10.1029/2010WR010174, 2011.

Garambois, P. A., Roux, H., Larnier, K., Castaings, W., and Dartus, D.: Characterization of process-oriented hydrologic model behavior with temporal sensitivity analysis for flash floods in Mediterranean catchments, Hydrology and Earth System Sciences, 17, 2305–2322, doi:10.5194/hess-17-2305-2013, 2013.

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

Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, Water Resources Research, 34, 751–763, doi:10.1029/97WR03495, 1998.

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

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., and Ye, M.: Towards a comprehensive assessment of model structural adequacy, Water Resources Research, 48, W08 301, doi:10.1029/2011WR011044, 2012.

Guse, B., Reusser, D. E., and Fohrer, N.: How to improve the representation of hydrological processes in SWAT for a lowland catchment - temporal analysis of parameter sensitivity and model performance, Hydrological Processes, 28, 2651–2670, doi:10.1002/hyp.9777, 2014.

Guzman, J., Moriasi, D., Gowda, P., Steiner, J., Starks, P., Arnold, J., and Srinivasan, R.: A model integration framework for linking SWAT and MODFLOW, Environmental Modelling & Software, 73, 103–116, doi:10.1016/j.envsoft.2015.08.011, 2015.

Haas, M. B., Guse, B., Pfannerstill, M., and Fohrer, N.: Detection of dominant nitrate processes in ecohydrological modeling with temporal parameter sensitivity analysis, Ecological Modelling, 314, 62–72, doi:10.1016/j.ecolmodel.2015.07.009, 2015.

Herbst, M., Gupta, H. V., and Casper, M. C.: Mapping model behaviour using Self-Organizing Maps, Hydrology and Earth System Sciences, 13, 395–409, doi:10.5194/hess-13-395-2009, 2009.

Herman, J. D., Reed, P. M., and Wagener, T.: Time-varying sensitivity analysis clarifies the effects of watershed model formulation on model behavior: Time-varying sensitivity of watershed models, Water Resources Research, 49, 1400–1414, doi:10.1002/wrcr.20124, 2013.

Hesse, C., Krysanova, V., and Voß, A.: Implementing in-stream nutrient processes in large-scale landscape modeling for the impact assessment on water quality, Environmental Modeling & Assessment, 17, 589–611, doi:10.1007/s10666-012-9320-8, 2012.

Hesse, C., Krysanova, V., Vetter, T., and Reinhardt, J.: Comparison of several approaches representing terrestrial and in-stream nutrient retention and decomposition in watershed modelling, Ecological Modelling, 269, 70–85, doi:10.1016/j.ecolmodel.2013.08.017, 2013.

Hrachowitz, M., Savenije, H., Blöschl, G., McDonnell, J., Sivapalan, M., Pomeroy, J., Arheimer, B., Blume, T., Clark, M., Ehret, U., Fenicia, F., Freer, J., Gelfan, A., Gupta, H., Hughes, D., Hut, R., Montanari, A., Pande, S., Tetzlaff, D., Troch, P., Uhlenbrook, S., Wagener, T., Winsemius, H., Woods, R., Zehe, E., and Cudennec, C.: A decade of Predictions in Ungauged Basins (PUB) - A review, Hydrological Sciences Journal, 58, 1198–1255, doi:10.1080/02626667.2013.803183, 2013.

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., Freer, J., Savenije, H. H. G., and Gascuel-Odoux, C.: Process consistency in models: The importance of system signatures, expert knowledge, and process complexity, Water Resources Research, 50, 7445–7469, doi:10.1002/2014WR015484, 2014.

Huang, S., Krysanova, V., and Hattermann, F. F.: Projection of low flow conditions in Germany under climate change by combining three RCMs and a regional hydrological model, Acta Geophysica, 61, 151–193, doi:10.2478/s11600-012-0065-1, 2013.

Hunter, N. M., Bates, P. D., Horritt, M. S., and Wilson, M. D.: Simple spatially-distributed models for predicting flood inundation: A review, Geomorphology, 90, 208–225, doi:10.1016/j.geomorph.2006.10.021, 2007.

Kiesel, J., Fohrer, N., Schmalz, B., and White, M. J.: Incorporating landscape depressions and tile drainages of a northern German lowland catchment into a semi-distributed model, Hydrological Processes, 24, 1472–1486, doi:10.1002/hyp.7607, 2010.

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

Klaus, J., Wetzel, C. E., Martinez-Carreras, N., Ector, L., and Pfister, L.: A tracer to bridge the scales: on the value of diatoms for tracing fast flow path connectivity from headwaters to meso-scale catchments, Hydrological Processes, doi:10.1002/hyp.10628, 2015.

Koch, H., Liersch, S., and Hattermann, F. F.: Integrating water resources management in eco-hydrological modelling, Water Science & Technology, 67, 1525, doi:10.2166/wst.2013.022, 2013a.

Koch, S., Bauwe, A., and Lennartz, B.: Application of the SWAT model for a tile-drained lowland catchment in North-Eastern Germany on subbasin scale, Water Resource Management, 27, 791–805, doi:10.1007/s11269-012-0215-x, 2013b.

Kollat, J. B., Reed, P. M., and Wagener, T.: When are multiobjective calibration trade-offs in hydrologic models meaningful?, Water Resources Research, 48, W03 520, doi:10.1029/2011WR011534, 2012.

Koren, V., Smith, M., and Cui, Z.: Physically-based modifications to the Sacramento Soil Moisture Accounting model. Part A: Modeling the effects of frozen ground on the runoff generation process, Journal of Hydrology, 519, Part D, 3475–3491, doi:10.1016/j.jhydrol.2014.03.004, 2014.

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

Krysanova, V. and Arnold, J. G.: Advances in ecohydrological modelling with SWAT: A review, Hydrological Sciences Journal, 53, 939–947, doi:10.1623/hysj.53.5.939, 2008.

Krysanova, V., Müller-Wohlfeil, D.-I., and Becker, A.: Development and test of a spatially distributed hydrological/water quality model for mesoscale watersheds, Ecological Modelling, 106, 261–289, doi:10.1016/S0304-3800(97)00204-4, 1998.

Laaha, G. and Blöschl, G.: Seasonality indices for regionalizing low flows, Hydrological Processes, 20, 3851–3878, doi:10.1002/hyp.6161, 2006.

Laaha, G. and Blöschl, G.: A national low flow estimation procedure for Austria, Hydrological Sciences Journal, 52, 625–644, doi:10.1623/hysj.52.4.625, 2007.

Legates, D. R. and McCabe, G. J.: Evaluating the use of goodness-of-fit measures in hydrologic and hydroclimatic model validation, Water Resources Research, 35, 233–241, doi:10.1029/1998WR900018, 1999.

Lenhart, T., Eckhardt, K., Fohrer, N., and Frede, H.-G.: Comparison of two different approaches of sensitivity analysis, Physics and Chemistry of the Earth, Parts A/B/C, 27, 645–654, doi:10.1016/s1474-7065(02)00049-9, 2002.

Luo, Y., Arnold, J., Allen, P., and Chen, X.: Baseflow simulation using SWAT model in an inland river basin in Tianshan Mountains, Northwest China, Hydrology and Earth System Sciences, 16, 1259–1267, doi:10.5194/hess-16-1259-2012, 2012.

Madsen, H.: Automatic calibration of a conceptual rainfall-runoff model using multiple objectives, Journal of Hydrology, 235, 276–288, doi:10.1016/S0022-1694(00)00279-1, 2000.

Martinez, G. F. and Gupta, H. V.: Hydrologic consistency as a basis for assessing complexity of monthly water balance models for the continental United States: Hydrologic consistency and model complexity, Water Resources Research, 47, doi:10.1029/2011WR011229, 2011.

Massmann, C. and Holzmann, H.: Analysis of the behavior of a rainfall-runoff model using three global sensitivity analysis methods evaluated at different temporal scales, Journal of Hydrology, 475, 97–110, doi:10.1016/j.jhydrol.2012.09.026, 2012.

Massmann, C., Wagener, T., and Holzmann, H.: A new approach to visualizing time-varying sensitivity indices for environmental model diagnostics across evaluation time-scales, Environmental Modelling & Software, 51, 190–194, doi:10.1016/j.envsoft.2013.09.033, 2014.

McMillan, H., Gueguen, M., Grimon, E., Woods, R., Clark, M., and Rupp, D. E.: Spatial variability of hydrological processes and model structure diagnostics in a 50 km2 catchment, Hydrological Processes, doi:10.1002/hyp.9988, 2013.

McMillan, H. K., Clark, M. P., Bowden, W. B., Duncan, M., and Woods, R. A.: Hydrological field data from a modeller's perspective: Part 1. Diagnostic tests for model structure, Hydrological Processes, 25, 511–522, doi:10.1002/hyp.7841, 2011.

Merz, R., Parajka, J., and Blöschl, G.: Time stability of catchment model parameters: Implications for climate impact analyses, Water Resources Research, 47, W02 531, doi:10.1029/2010WR009505, 2011.

Moriasi, D. N., Arnold, J. G., and Green, C. H.: Incorporation of Hooghoudt and Kirkham tile drain equations into SWAT2005, 4th International SWAT Conference Proceedings, pp. 139–147, 2007a.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L.: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, Transactions of the ASABE, 50, 885–900, 2007b.

Moriasi, D. N., Rossi, C. G., Arnold, J. G., and Tomer, M. D.: Evaluating hydrology of the Soil and Water Assessment Tool (SWAT) with new tile drain equations, Journal of Soil and Water Conservation, 67, 513–524, doi:10.2489/jswc.67.6.513, 2012.

Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models part I - A discussion of principles, Journal of Hydrology, 10, 282–290, doi:10.1016/0022-1694(70)90255-6, 1970.

Neitsch, S. L., Arnold, J. G., Kiniry, J. R., and Williams, J. R.: SWAT Theoretical Documentation Version 2009, Grassland, Soil and Water Research Laboratory, Agricultural Research Service. Blackland Research Center, Texas Agricultural Experiment Station, 2011.

Niehoff, D., Fritsch, U., and Bronstert, A.: 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, 80–93, doi:10.1016/S0022-1694(02)00142-7, 2002.

Nossent, J. and Bauwens, W.: Multi-variable sensitivity and identifiability analysis for a complex environmental model in view of integrated water quantity and water quality modeling, Water Science & Technology, 65, 539, doi:10.2166/wst.2012.884, 2012.

Nossent, J., Elsen, P., and Bauwens, W.: Sobol sensitivity analysis of a complex environmental model, Environmental Modelling & Software, 26, 1515–1525, doi:10.1016/j.envsoft.2011.08.010, 2011.

Orth, R., Staudinger, M., Seneviratne, S. I., Seibert, J., and Zappa, M.: Does model performance improve with complexity? A case study with three hydrological models, Journal of Hydrology, doi:10.1016/j.jhydrol.2015.01.044, 2015.

Pechlivanidis, I. G., Jackson, B., McMillan, H., and Gupta, H.: Use of an entropy-based metric in multiobjective calibration to improve model performance, Water Resources Research, 50, 8066–8083, doi:10.1002/2013WR014537, 2014.

Perrin, C., Michel, C., and Andreassian, V.: Improvement of a parsimonious model for streamflow simulation, Journal of Hydrology, 279, 275–289, doi:10.1016/S0022-1694(03)00225-7, 2003.

Pokhrel, P., Yilmaz, K. K., and Gupta, H. V.: Multiple-criteria calibration of a distributed watershed model using spatial regularization and response signatures, Journal of Hydrology, 418-419, 49–60, doi:10.1016/j.jhydrol.2008.12.004, 2012.

Rathjens, H., Dörnhöfer, K., and Oppelt, N.: IRSeL - An approach to enhance continuity and accuracy of remotely sensed land cover data, International Journal of Applied Earth Observation and Geoinformation, 31, 1–12, doi:10.1016/j.jag.2014.02.010, 2014.

Rathjens, H., Oppelt, N., Bosch, D. D., Arnold, J. G., and Volk, M.: Development of a grid-based version of the SWAT landscape model, Hydrological Processes, 29, 900–914, doi:10.1002/hyp.10197, 2015.

Reusser, D. E. and Zehe, E.: Inferring model structural deficits by analyzing temporal dynamics of model performance and parameter sensitivity, Water Resources Research, 47, W07 550, doi:10.1029/2010WR009946, 2011.

Reusser, D. E., Blume, T., Schaefli, B., and Zehe, E.: Analysing the temporal dynamics of model performance for hydrological models, Hydrology and Earth System Sciences, 13, 999–1018, doi:10.5194/hess-13-999-2009, 2009.

Reusser, D. E., Buytaert, W., and Zehe, E.: Temporal dynamics of model parameter sensitivity for computationally expensive models with the Fourier amplitude sensitivity test, Water Resources Research, 47, W07 551, doi:10.1029/2010WR009947, 2011.

Saltelli, A., Chan, K., and Scott, E. M.: Sensitivity analysis, J. Wiley & sons, New York; Chichester; Weinheim [etc], 2000.

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A., and Carrillo, G.: Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA, Hydrology and Earth System Sciences, 15, 2895–2911, doi:10.5194/hess-15-2895-2011, 2011.

Seibert, J. and McDonnell, J. J.: On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration, Water Resources Research, 38, 1241, doi:10.1029/2001WR000978, 2002.

Shafii, M. and Tolson, B. A.: Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives, Water Resources Research, 51, 3796–3814, doi:10.1002/2014wr016520, 2015.

Sieber, A. and Uhlenbrook, S.: Sensitivity analyses of a distributed catchment model to verify the model structure, Journal of Hydrology, 310, 216–235, doi:10.1016/j.jhydrol.2005.01.004, 2005.

Smakhtin, V. U.: Low flow hydrology: a review, Journal of Hydrology, 240, 147–186, doi:10.1016/S0022-1694(00)00340-1, 2001.

Stoelzle, M., Weiler, M., Stahl, K., Morhard, A., and Schuetz, T.: Is there a superior conceptual groundwater model structure for baseflow simulation?, Hydrological Processes, 29, 1301–1313, doi: 10.1002/hyp.10251, 2015.

Strauch, M. and Volk, M.: SWAT plant growth modification for improved modeling of perennial vegetation in the tropics, Ecological Modelling, 269, 98–112, doi:10.1016/j.ecolmodel.2013.08.013, 2013.

Strauch, M., Lima, J. E. F., Volk, M., Lorz, C., and Makeschin, F.: The impact of best management practices on simulated streamflow and sediment load in a Central Brazilian catchment, Journal of Environmental Management, 127, 24–36, doi:10.1016/j.jenvman.2013.01.014, 2013.

Tallaksen, L. M., Madsen, H., and Clausen, B.: On the definition and modelling of streamflow drought duration and deficit volume, Hydrological Sciences Journal, 42, 15–33, doi:10.1080/02626669709492003, 1997.

Tetzlaff, D., McDonnell, J. J., Uhlenbrook, S., McGuire, K. J., Bogaart, P. W., Naef, F., Baird, A. J., Dunn, S. M., and Soulsby, C.: Conceptualizing catchment processes: simply too complex?, 22, 1727–1730, doi:10.1002/hyp.7069.

Thielen, J., Bartholmes, J., Ramos, M., and De Roo, A.: The European Flood Alert System–Part 1: Concept and development, Hydrology and Earth System Sciences, 13, 125, doi:10.5194/hess-13-125-2009, 2009.

van Griensven, A., Meixner, T., Grunwald, S., Bishop, T., Diluzio, M., and Srinivasan, R.: A global sensitivity analysis tool for the parameters of multi-variable catchment models, Journal of Hydrology, 324, 10–23, doi:10.1016/j.jhydrol.2005.09.008, 2006.

van Werkhoven, K., Wagener, T., Reed, P., and Tang, Y.: Sensitivity-guided reduction of parametric dimensionality for multi-objective calibration of watershed models, Advances in Water Resources, 32, 1154–1169, doi:10.1016/j.advwatres.2009.03.002, 2009.

Vogel, R. and Fennessey, N.: Flow-Duration Curves. I: New Interpretation and Confidence Intervals, Journal of Water Resources Planning and Management, 120, 485–504, doi:10.1061/(ASCE)0733-9496(1994)120:4(485), 1994.

Volk, M., Arnold, J. G., Bosch, D. D., Allen, P. M., and Green, C. H.: Watershed Configuration and Simulation of Landscape Processes with the SWAT Model, in: MODSIM07 International Congress on Modelling and Simulation "Land, Water and Environmental Management: Integrated Systems for Sustainability", Christchurch, 10-13 December 2007, pp. 2383–2389, 2007.

Volk, M., Liersch, S., and Schmidt, G.: Towards the implementation of the European Water Framework Directive?: Lessons learned from water quality simulations in an agricultural watershed, Land Use Policy, 26, 580–588, 2009.

Vrugt, J. A. and Ter Braak, C. J. F.: DREAM(D): an adaptive Markov Chain Monte Carlo simulation algorithm to solve discrete, noncontinuous, and combinatorial posterior parameter estimation problems, Hydrology and Earth System Sciences, 15, 3701–3713, doi:10.5194/hess-15-3701-2011, 2011.

Vrugt, J. A., Gupta, H. V., Bastidas, L. A., Bouten, W., and Sorooshian, S.: Effective and efficient algorithm for multiobjective optimization of hydrologic models, Water Resources Research, 39, 1214, doi:10.1029/2002WR001746, 2003a.

Vrugt, J. A., Gupta, H. V., Bouten, W., and Sorooshian, S.: A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters, Water Resources Research, 39, 1201, doi:10.1029/2002WR001642, 2003b.

Wagener, T. and Gupta, H. V.: Model identification for hydrological forecasting under uncertainty, Stochastic Environmental Research and Risk Assessment, 19, 378–387, doi:10.1007/s00477-005-0006-5, 2005.

Wagener, T., McIntyre, N., Lees, M. J., Wheater, H. S., and Gupta, H. V.: Towards reduced uncertainty in conceptual rainfall-runoff modelling: dynamic identifiability analysis, Hydrological Processes, 17, 455–476, doi:10.1002/hyp.1135, 2003.

White, K. L. and Chaubey, I.: Sensitivity analysis, calibration, and validation for a multisite and multivariable SWAT model, Journal of the American Water Resources Association, 41, 1077–1089, doi:10.1111/j.1752-1688.2005.tb03786.x, 2005.

Wu, K. and Johnston, C. A.: Hydrologic response to climatic variability in a Great Lakes Watershed: A case study with the SWAT model, Journal of Hydrology, 337, 187–199, doi:10.1016/j.jhydrol.2007.01.030, 2007.

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resources Research, 44, doi:10.1029/2007WR006716, 2008.

Zhang, X., Hörmann, G., Gao, J., and Fohrer, N.: Structural uncertainty assessment in a discharge simulation model, Hydrological Sciences Journal, 56, 854–869, doi:10.1080/02626667.2011.587426, 2011a.

Zhang, X., Srinivasan, R., Arnold, J., Izaurralde, R. C., and Bosch, D.: Simultaneous calibration of surface flow and baseflow simulations: a revisit of the SWAT model calibration framework, Hydrological Processes, 25, 2313–2320, doi:10.1002/hyp.8058, 2011b.

List of Figures

List of Tables

Curriculum vitae

Persönliche Angaben	
Name:	**Matthias Pfannerstill**
Familienstand:	**verheiratet**
Nationalität:	**deutsch**
Geburtstag:	**16.12.1983**
Geburtsort:	**Husum**

Werdegang	
Schule	
08/1994 - 06/2003	**Werner-Heisenberg-Gymnasium**, Heide: - Allgemeine Hochschulreife (06/2003; Note: 2,1)
08/2003 - 04/2004	**Zivildienst**, Kindertagesstätte Lunden
Hochschule	
10/2004 - 04/2010	**Studium der Geoökologie** an der Technischen Universität zu Braunschweig: - Diplom (Note: 1,4) - Vertiefungen: Umweltsystemanalyse, Ingenieurhydrologie, Geochemie, Umweltrecht
Berufliche Tätigkeiten	
06/2010 - 10/2011	**Wissenschaftlicher Angestellter** Institut für Natur- und Ressourcenschutz, Abteilung Hydrologie und Wasserwirtschaft, Christian-Albrechts-Universität zu Kiel Forschungsprojekt: Untersuchungen zur Wirkung von Reaktiven Grabensystemen auf die Nährstoffrückhaltung in Schleswig-Holstein
11/2011 - 12/2014	**Promotionsstipendiat der Deutschen Bundesstiftung Umwelt** Bearbeitung am Institut für Natur- und Ressourcenschutz, Abteilung Hydrologie und Wasserwirtschaft, Christian-Albrechts-Universität zu Kiel Thema: Evaluierung von Strategien zur Verminderung der Stickstoffbelastung in dränierten Gebieten
10/2014 - 11/2014	**Auslands-Stipendiat der Heinz-Wüstenberg-Stiftung** Forschungsaufenthalt beim USDA, Agricultural Research Service in Texas, USA Weiterentwicklung des Soil and Water Assessment Tools (SWAT)
seit 07/2015	**Wissenschaftlicher Angestellter** Landesamt für Landwirtschaft, Umwelt und ländliche Räume des Landes Schleswig-Holstein in Flintbek, Abteilung Gewässer, Dezernat Grundwasserhydrologie und Grundwasserschutz Forschungsprojekt: ReWaM - Verbundprojekt MUTReWa

Summary of peer-reviewed publications of the author

1. Pfannerstill, M., Hugenschmidt, C., Trepel, M., and Fohrer, N. (2012). Reaktive Grabensysteme zur Reduktion des diffusen Stickstoffeintrags aus drainierten landwirtschaftlichen Flächen. Hydrologie und Wasserbewirtschaftung, 56(4):203-214, DOI: 10.5675/HyWa_2012,4_5.

2. Holsten, B. , Bednarek, A., , Fier, A., Fohrer, N., Heckrath, G., Höper, H., Hugenschmidt, C., Kjaergard, C., Krause, B., Litz, N., Matzinger, A., Orlikowski, D., Perillon, C., Pfannerstill, M., Rouault, P., Schäfer, W., Trepel, M., Ubraniak, M., and Zalewski, M. (2011). Potentiale für den Einsatz von Nährstoff-Filtersystemen in Deutschland zur Verringerung der Nährstoffeinträge in Oberflächengewässer. Hydrologie und Wasserberwirtschaftung, 56(1):4-15,
DOI: 10.5675/HyWa_2012,1_1.

3. Pfannerstill, M., Guse, B., and Fohrer, N. (2013). A multi-storage groundwater concept for the SWAT model to emphasize nonlinear groundwater dynamics in lowland catchments. Hydrological Processes, 28(22):5599-5612, DOI: 10.1002/hyp.10062.

4. Pfannerstill, M., Guse, B., and Fohrer, N. (2014). Smart low flow signature metrics for an improved overall performance evaluation of hydrological models. Journal of Hydrology, 510:447-458,
DOI: 10.1016/j.jhydrol.2013.12.044.

5. Pfannerstill, M., Guse, B., Reusser, D., and Fohrer, N. (2015). Temporal parameter sensitivity guided verification of process dynamics, Hydrology and Earth System Sciences Discussion, 12:1729-1764, DOI: 10.5194/hessd-12-1729-2015.

6. Guse, B., Pfannerstill, M., and Fohrer, N. (2015). Dynamic modelling of land use change impacts on nitrate loads in rivers. Environmental Processesses, DOI: 10.1007/s40710-015-0099-x.

7. Haas, M., Guse, B., Pfannerstill, M., and Fohrer, N. (2015). Detection of dominant nitrate processes in ecohydrological modelling with temporal parameter sensitivity analysis. Ecological Modelling, 314:62-72, DOI: 10.1016/j.ecolmodel.2015.07.009.

8. Kiesel, J., Pfannerstill, M., Schmalz, B., Khoroshavin, V., Sheldukov, A., Veshkurseva, T., and Fohrer, N. (2015). Modelling of hydrological processes in snowmelt-governed meso-scale catchments of the Western Siberian Lowlands. Hydrological Processes, under review: HYP-15-0095.

9. Guse, B., Pfannerstill, M., Strauch, M., Reusser, D., Lüdtke, S., Volk, M., Gupta, H., and Fohrer, N. (2015). Parameter and process identification from temporal sensitivity patterns. Hydrological Processes, under review: : HYP-15-0554.

Declaration of authorship

Hiermit erkläre ich, dass ich die vorliegende Dissertation, abgesehen von der Beratung durch meine Betreuer, selbständig verfasst habe und keine weiteren Quellen und Hilfsmittel als die hier angegebenen verwendet habe. Diese Arbeit hat weder ganz noch in Teilen bereits an anderer Stelle einer Prüfungskommission zur Erlangung des Doktorgrades vorgelegen. Ich erkläre, dass die vorliegende Arbeit gemäß der Grundsätze zur Sicherung guter wissenschaftlicher Praxis der Deutschen Forschungsgemeinschaft erstellt wurde.

Kiel, 02.09.2015 Matthias Pfannerstill