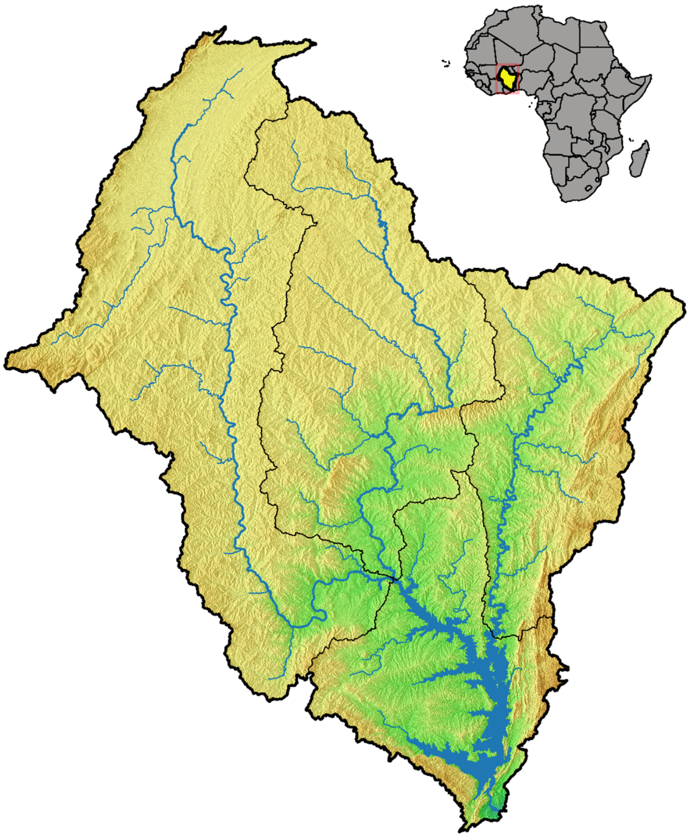
Improving the performance of distributed conceptual hydrological models using the spatio-temporal patterns of RS observations

A Hydrological Modelling Thesis



Daan te Witt

A thesis presented for the degree of
Master of Science in Civil Engineering

Department Name:	Faculty of Civil Engineering and Geosciences
University Name:	Delft University of Technology
Country:	The Netherlands
Date:	24-06-2021

Improving the performance of distributed conceptual hydrological models using the spatio-temporal patterns of RS observations

A Hydrological Modelling Thesis

Daan te Witt

A thesis presented for the degree of
Master of Science in Civil Engineering

Thesis Committee

Dr.	Markus Hrachowitz	Delft University of Technology
MSc.	Michel Zuijderwijk	Witteveen+Bos
Dr.ir.	Rolf Hut	Delft University of Technology
Dr.ir.	Gerrit Schoups	Delft University of Technology
Dr.	Moctar Dembélé	University of Lausanne
PhD. Cand.	Jerom Aerts	Delft University of Technology

Institution

Department Name:	Faculty of Civil Engineering and Geosciences
University Name:	Delft University of Technology
Country:	The Netherlands
Date:	24-06-2021

Preface

This document contains a complete description of my MSc thesis project and was written to obtain the degree of Master of Science in Civil Engineering - Water Management at Delft University of Technology. I carried out this research project with the help of my supervisors from Delft University of Technology, but also with the help of the group Water Nuisance at Witteveen+Bos, where I did an internship during the research period. The thesis project is about the effect of the use of spatio-temporal patterns of remote sensing data in the calibration of conceptual distributed hydrological models on the overall model performance. I was inspired by the work of Moctar Dembélé to do my thesis within this specific field of research.

My thesis can be read by anyone interested in hydrological modelling in general, and especially anyone interested in the use of remote sensing data in hydrological models. A basic background in hydrology may however be needed to fully understand all the discussed topics. I learned a lot from doing this thesis project, especially because this was only one of my first hydrological studies, but also because I had an amazing team of experienced supervisors who I could ask for advice on any problem I encountered. One of the most important things I learned is that in the field of hydrological modelling it is essential to gain experience. A starting hydrologist can only really learn in this field by applying his knowledge to real world cases. Therefore, I also realise that there is still much to learn in future projects.

During my thesis project, I had a lot of help from many people at Delft University of Technology and Witteveen+Bos. Therefore I would like to thank several people for their contributions to my project, whether big or small. I want to thank Mónica Estébanez Camarena, for making available rain gauge data in northern Ghana for comparison to remotely sensed precipitation data. I want to thank Nick van de Giesen, for his advice in the early stage of my research project, and Ivar Abas and Ingrid van den Brink for their help in choosing a thesis topic. I am also thankful to Nathalie Rouché, for making available streamflow data from the SIEREM database, and Moctar Dembélé, for making available streamflow data he used in his research. I also want to thank Maurits Ertsen for learning me how to set up a research project and how to analyse the literature within a specific research field. I also really appreciate the help from Pieter Hazenberg from Deltares, with helping me to fully understand the wflow hbv model.

I am also very thankful for the chance to do an internship at the group Water Nuisance at Witteveen+Bos. I want to thank all members of this group, being Josje van Houwelingen, Ivar Abas, Bart Dekens and Michel Zuijderwijk, for their help and advice and for giving me the chance to learn from and about their work. I am also very grateful for the help of the eWaterCycle team, consisting of Jerom Aerts, Niels Drost and Rolf Hut. They provided my with the technical tools to set up my model and to carry out my research project, and also put a lot of time in helping me with the computational problems I encountered within this project. Above all, I want to thank my daily supervisor from Delft University of Technology, Markus Hrachowitz, for the many zoom-sessions we had about hydrological modelling in which I really learned a lot every time, and for providing me with very practical insights within the research field, but also for keeping me motivated during the difficult time in which I wrote this thesis. The same holds for my daily supervisor from Witteveen+Bos, Michel Zuijderwijk, who I especially want to thank for the many interesting talks we had about both our projects, and for giving me the chance to learn about real-world applications and projects. Last but not least, I want to thank the other members of my thesis committee, consisting of Gerrit Schoups, Moctar Dembélé, Jerom Aerts and Rolf Hut, for their time, advice and feedback and their contribution to this project.

Daan te Witt, June 2021, Delft

Summary

Hydrological models are used for all kinds of water management applications. Detailed hydrological simulations are needed to solve the hydrological problems of the 21^{st} century, especially in developing countries. However, sufficient hydrological and meteorological data is often not available. The use of remote sensing (RS) datasets may offer a solution to this problem. RS datasets can perfectly be applied in distributed conceptual hydrological models. In this study, several RS datasets are applied in the calibration of a distributed conceptual hydrological model, and the influence of this approach on the overall model performance is assessed.

The RS data applied in this study include terrestrial water storage anomaly (TWSA) data, normalized difference vegetation index (NDVI) data, and soil moisture (SM) data. Also the input and forcing data for the hydrological model consists of datasets based on satellite observations. This data is used in a wflow hbv model, which is applied to the Volta basin in Western Africa as a case-study. In this study, not only the effect of including RS data in the calibration of a distributed hydrological model on streamflow is assessed, but also the effect on a set of internal components of the system, directly related to the datasets used for calibration. These internal stocks and fluxes are the TWSA, the actual evapotranspiration (AET) and the amount of soil moisture in the unsaturated zone. Together with streamflow, the assessment of these stocks and fluxes make up the overall model performance of the system.

The effect on the overall model performance is examined using different scenarios, in which different combinations of datasets are used for calibration. Not the absolute values, but the spatio-temporal patterns of the remote sensing datasets are used for model assessment. This is done using the spatial pattern efficiency metric (E_{SP}). Model optimization was done using the Dynamically Dimensioned Search (DDS) algorithm.

The results show that the hydrological model developed for this case-study is already able to simulate streamflow and the temporal patterns of the RS datasets quite well, when it is calibrated on streamflow only. However, the spatial pattern representation of the RS datasets was found to be inadequate and the differences in streamflow simulation performance for the different subcatchments is large. When SM or TWSA data was added to the calibration procedure, the temporal and spatial pattern representations only changed minimally, which is attributed to limited model complexity and flexibility. However, generally a trade-off effect was observed in which the spatial and temporal pattern representation improved, but the streamflow performance decreased. This effect was stronger for the addition of the SM dataset to the calibration than for the addition of TWSA dataset. Although there is definitely a strong connection between NDVI and AET, the physical relation between the two variables was found to be too weak to be used for hydrological model calibration, even when only the spatial and temporal pattern information was used.

The overall model performance did improve most in the calibration catchments in the scenario in which Q, SM and TWSA data were combined in the calibration procedure, but the differences with the baseline scenario were only small. For the streamflow performance however, the differences between the scenarios are quite significant. It was shown that calibrating a hydrological model on the spatial and temporal patterns of RS data only (non-Q calibration) can accurately represent the temporal pattern of streamflow observations, but not the magnitude of the flow values. It is recommended to repeat this study using a more complex and more flexible model setup, which allows the model to use the freedom it is given to better represent the spatial patterns observed with RS.

Table of Contents

Preface i

Summary ii

1 Introduction 1

2 Data & Study Area 5

2.1 Data 5

2.1.1 Static input data: DEM & land use 5

2.1.2 Dynamic input data: Precipitation and PET 6

2.1.3 Calibration / Evaluation data: Q, TWSA, NDVI and SM 8

2.2 Study Area 12

3 Methodology 15

3.1 Model description: wflow hbv 15

3.2 Model Implementation & Scenarios 19

3.3 Model assessment: A multivariate approach 24

3.3.1 The calibration algorithm: DDS 24

3.3.2 Streamflow assessment 25

3.3.3 TWSA, AET and SM assessment 28

3.3.4 Overall model assessment 31

4 Results 33

4.1 Results Scenario 1: Q-only Calibration 33

4.2 Results scenario 2: Q + SM Calibration 40

4.3 Results scenario 3: Q + TWSA Calibration 44

4.4 Results scenario 4: Q + SM + TWSA Calibration 46

4.5 Results scenario 5: SM + TWSA Calibration 48

4.6 Overview Results 52

5 Discussion 59

5.1 Changes to the Methodology 59

5.2 Interpretation of the Results 62

5.3 Discussion of the Data and Methodology 67

5.4 What did we learn? 70

6 Conclusions & Recommendations 72

Appendices 77

A Brief history and applications of the hbv model 77

B CHIRPS precipitation comparison with rain gauge data 78

C Spatio-temporal patterns of RS data 80

D Water balance method used for the determination of T and $S_{u,max}$ 82

E Baseflow recession curve analysis 87

F Amount of model runs per calibration 89

G Results Scenario 1: Q-only Calibration 90

H Results Scenario 2: Q+SM Calibration 96

I Results Scenario 3: Q+TWSA Calibration 103

J Results Scenario 4: Q+SM+TWSA Calibration 112
K Results Scenario 5: SM+TWSA Calibration 120
L Results Scenarios Compared 127

1 Introduction

Hydrological models are used for all kinds of water management applications. These applications include crucial public services like flood warnings and drought monitoring, but also provide important information on the state of river systems in less extreme situations, for instance for reservoir management. These services are especially vital for farmers in arid and semi-arid climates, people living in areas prone to river flooding, and people dependant on a reservoir for either their water use, electricity or flood protection. Since the majority of the human population lives within 3 kilometres of a fresh water body for one of these or other services (Kummu et al., 2011), it is evident that hydrological models must be able to accurately simulate river flow.

The classical approach of setting up a hydrological model is as follows. A hydrologist takes meteorological data, like precipitation and temperature measurements from a weather station, as forcing for the model. The free parameters of the model are determined by calibration against discharge data from a gauging station, and this parameter set is tested against another part of the discharge timeseries for evaluation. If the model simulates the observed discharge well in both periods, the model is accepted and can be used for a specific water management application.

However, the modelling procedure described here is often not applicable and is also increasingly not sufficient to solve the hydrological problems of the 21^{st} century. In developing countries in arid and semi-arid regions, where water related problems are largest, basic meteorological and discharge observations are seldom available, and if they are, the quality is often poor. Furthermore, classical lumped hydrological models only model discharge at the outlet of a catchment, instead of at every place from river outlet to source. This classical but simple approach of hydrological modelling is often not sufficient to deal with the highly variable meteorological conditions and the hydrological extremes that are common in countries at low latitudes.

Therefore, detailed hydrological simulations are needed to identify and simulate hydrological processes at the sub-catchment scale. Since discharge is only the sum of all hydrological processes upstream of the river outlet (Beven et al., 1999), it is of fundamental importance that a hydrological model captures these hydrological processes at the sub-catchment scale. If the internal processes are simulated incorrectly but the discharge simulation performance is acceptable, then this is the result of model parameters correcting for each other. These type of models should be rejected because their performance will eventually collapse under different conditions.

The problem of parameters compensating for each other is known as the equifinality problem (Beven, 2006; Savenije, 2001) and the problem of simulating streamflow right but doing so without simulating internal processes right has been described as 'getting the right answers for the wrong reasons' (Kirchner, 2006). The models that do get the right answers for the wrong reasons are known as 'mathematical marionettes' because of their ever lasting ability to compensate for parameter errors because of their overparameterization. It has been shown that hydrological models should be "as complex as necessary and as simple as possible" (Hrachowitz et al., 2014). Among other things this means that a model should be able to simulate the most important hydrological processes and signatures correctly, but should also have an as small as possible number of parameters. Models that are not able to capture the internal processes are also more susceptible to climate and land use changes and are unfit for transfer to catchments or time periods other than the calibration period.

A solution to the modelling problems identified above is to use distributed hydrological models (DHM's) forced with remote sensing (RS) data. Distributed hydrological models are process explicit models that account for the most important hydrological processes and do so, separately, for each grid cell. These grid cells are generally much smaller than most catchments and therefore include information on the hydrological states and fluxes within the modelled catchment. DHM's can be forced with gridded forcing data, generated by a set of meteorological observations or observed by satellites. Over the last few decades, more and more RS data products have become available that can be used as hydrological forcing. This development offers great opportunities, particularly in regions where accurate meteorological data are scarce. DHM's provide hydrological information on every grid cell in the model domain and this makes it possible to assess the internal dynamics of the model.

RS data can also be used for hydrological model calibration (Dembele, Hrachowitz, et al., 2020; Dembele, Zwart, et al., 2020; Demirel et al., 2019; Dezetter & Ruelland, 2012; Koch et al., 2018; Mendiguren et al., 2017; Nijzink et al., 2018; Rakovec, Kumar, Attinger, & Samaniego, 2016; Stisen et al., 2011, 2018; Tangdamrongsub et al., 2015; Zink et al., 2018). Calibration of multiple hydrological stocks and fluxes (multivariate calibration) will result in the rejection of model parameter sets with low performing simulations for any of the assessed variables. This calibration approach can thus be seen as an extra model assessment compared to model assessment based on streamflow only, thereby allowing to invalidate (Beven, 1993) more models (or parameter sets). Doing so results in better informed models, that capture the internal system dynamics better than classic lumped models, therefore suffering less from equifinality problems and mathematical curve-fitting (if the model parameterization is kept the same!). These models will perform better in temporal and spatial evaluations and will be better able to retain high performance in climate or land use change scenarios. Taking this one-step further, DHM's forced and assessed with RS data could be key in building hydrological models that are able to do predictions in ungauged basins (Hrachowitz et al., 2013).

As was already partly explained in the last section, RS data can not only be used to force a model, but also to update model states or assess model performance. RS datasets often have a global spatial coverage, but the accuracy of observations is generally poor. It's most important strength subsists in the use of the observed spatio-temporal patterns of hydrological states and fluxes in hydrological models (Dembele, Hrachowitz, et al., 2020; Demirel, Koch, et al., 2018; Demirel, Mai, et al., 2018; Koch et al., 2017, 2018; Mendiguren et al., 2017). RS data is already being used extensively for calibration in recent research (Hulsman et al., 2020, 2021; Bouaziz et al., 2021)). Often, temporal mean spatial patterns of different hydrological states and fluxes are extracted from observations, and the hydrological model is optimized to simulate those patterns (Dembele, Hrachowitz, et al., 2020; Demirel, Mai, et al., 2018; Koch et al., 2018).

Optimizing for spatial patterns here does not mean optimizing for the absolute values of a hydrological state or flux for each grid cell, but only for the relative value of such a state or flux in one cell, compared to all other cells. In this way the spatio-temporal value of satellite observations can be exploited, without including the weakness of their poor absolute values. Recent research also demonstrates that the inclusion of multiple RS data sources within the model calibration improves the overall model performance (Dembele, Hrachowitz, et al., 2020; Nijzink et al., 2018; Rakovec, Kumar, Attinger, & Samaniego, 2016; Stisen et al., 2018).

The goal of this research is to investigate how the use of the spatio-temporal patterns of RS data on the highest resolution available in the calibration of conceptual distributed hydrological models influences the overall prediction skill of those models. The prediction skill or model performance in this research focuses not only on streamflow, as is common in classical hydrological modelling studies, but also on a selection of other hydrological states and fluxes of the model. The combined performance on streamflow and the selection of hydrological states and fluxes is hereafter referred to as 'overall performance'. This leads to the following research question:

How does the inclusion of the spatio-temporal patterns of RS data products in the calibration of a distributed conceptual hydrological model influence the overall model performance?

Hence, the specific goal of this research is to find out how model calibration, using a combination of streamflow and different RS data sets, influences the performance of the model in simulating a selection of hydrological stocks and fluxes. It is hypothesized that in general, implementation of spatio-temporal pattern information in the calibration will improve the performance of the internal hydrological process simulations, but decrease the streamflow performance. This trade-off effect was also observed by Dembele, Hrachowitz, et al. (2020) and can be explained by the simple fact that in this multivariate approach, streamflow is not the only hydrological flux you are optimizing for anymore. It is expected that with the inclusion of more data in the calibration, more and more parameter sets can be falsified because they do not perform satisfactory for the internal hydrology of the system. This will especially become clear in the model evaluation.

The goal of this research project is not to develop the best possible hydrological model for a certain region, but to gain insight in the benefits and trade-offs of using RS data for DHM calibration. Hence, the final model itself is not the most important deliverable of this research project, but this is the calibration method and its results and insights. Therefore, this research project could have been performed and can be repeated with every DHM available. It is expected that the results will be very similar.

The project does not include physically based models, data driven models or stochastic models, but focuses on conceptual process-based models. This choice was made deliberately because the underlying goal of this research is to improve internal model dynamics, i.e. all the hydrological stocks and fluxes within the hydrological system that together form the streamflow response. Conceptual hydrological models aim to simulate those internal model dynamics and are therefore most consistent with the research goal.

The model used in this research is the wflow hbv model. This is a distributed version of the Hydrologiska Byrans Vattenbalansavdelning (HBV) (Lindström et al., 1997; Bergström, 1992) rainfall-runoff model. wflow is an open-source distributed hydrological modeling platform from Deltares and includes several distributed versions of established and widely used hydrological models, like HBV. wflow makes it possible to build hydrological models that purely rely on satellite input data, making the models very suitable for this research project. The HBV model is chosen because of its proven applicability in all kinds of environments (Wetterhall, 2014) (See Appendix A) and because of it's relatively simple structure, with enough parameters to allow for a comprehensive optimization.

The wflow model will be run using eWaterCycle. eWaterCycle is a framework still under development by TU Delft and the eScience Center. Its goal is to make it very easy to compare the results of one hydrological model to other hydrological models in the same area, even if these models are written in another programming language. Some wflow models, like the wflow hbv model, are already incorporated in the eWaterCycle framework and therefore using this framework makes it possible to run lots of different scenarios, within the time available. eWaterCycle will thereby allow me to focus more on hydrology and less on programming. From eWaterCycle it is also possible to migrate my model and calibration scripts to the Cartesius supercomputer to reduce the computation time of the model runs.

To answer the research question, several RS data products are selected and used to calibrate a DHM. The selected RS data products are the Terrestrial Water Storage Anomaly (TWSA), a Normalized Difference Vegetation Index (NDVI) dataset and a soil moisture (SM) dataset. These datasets are used as calibration data on top of the ground-based discharge observations and next to the static and dynamic input data like a digital elevation model (DEM), a land use map, and gridded precipitation and potential evapotranspiration (PET) maps.

The case-study in which the research project was performed is the Volta basin. The reason for choosing this basin is threefold. The first reason is that in Western Africa there is a lack of high-quality ground-based meteorological and hydrological data, which makes the region suitable for the use of RS data. The second reason is its large size, which allows for the application of all kinds of RS data, which is not always possible in small basins because of the low spatial resolution of the products. The third reason is the presence of a reservoir, which creates the opportunity to also apply the model for simple reservoir management applications.

2 Data & Study Area

This Chapter first gives an overview of the data used in this study and then gives background information on the study area. The used datasets are almost all RS products. For the most important ones, a short explanation is given on the observation techniques of the sensors on board of the satellites that performed the observations for each dataset. Additionally, a justification is given for the specific datasets used in this study, since often multiple datasets are available. The section on the study area analyses the morphological and meteorological features of the Volta basin, and also elaborates on the mean spatial and temporal patterns of the hydrological stocks and fluxes inside the basin observed with RS.

2.1 Data

The data used in this study can be subdivided into three categories, namely static input data, dynamic input data and calibration and evaluation data. The static input data is assumed to be stable over the modelled time period. This includes a DEM and land use data. Dynamic input consists of meteorological forcing data that does change over time, i.e. precipitation and PET data. The calibration and evaluation data are all dynamic datasets used for model assessment. These datasets include ground-based streamflow (Q) observations and RS data products of the TWSA, the NDVI and SM within the catchment.

2.1.1 Static input data: DEM & land use

The static input data consists of elevation and land use data. This data is assumed to be constant over the modelled time period. For elevation data, this assumption is justified because this morphological feature does change very little over time. However, land use can change considerably over time and as a matter of fact this is happening in West Africa in the form of deforestation and cropland expansion, with considerable hydrological effects (Li et al., 2007). The simplification of using a static land use map is therefore less valid than for the elevation map, but still necessary to keep modelling of the hydrology of the Volta feasible. Normally, also a soil map is used in the setup of a wflow hydrological model, but this map was left out in this project because the map did not contribute to the final parameterization of the model. A landscape classification was introduced instead of a soil classification. This classification will be introduced in Chapter 3.

The DEM used in this model is the Global Multi-resolution Terrain Elevation Data 2010 (GMTED 2010) (Danielson & Gesch, 2011). The dataset was developed by the U.S. Geological Survey (USGS) and the National Geospatial-Intelligence Agency (NGA) and includes several products containing different elevation characteristics. In this study, the breakline emphasis product was used, which is most suitable for hydrological applications such as watershed delineation. The dataset is available at 3 different resolutions of approximately 250 m (7.5 arc seconds), 500 m (15 arc seconds) and 1000 m (30 arc seconds). In this study, the 250 m product was used. This resolution is deemed sufficient for application in a hydrological model of the complete Volta basin.

The land use dataset used in this study is the GlobCover 2009 product (Bontemps et al., 2011). This product is an initiative from the European Space Agency (ESA) together with multiple other organizations to deliver global land cover maps. This version of the product covers the period January till December 2009 and is derived from the Medium Resolution Imaging Spectrometer Instrument (MERIS) Fine Resolution (FR) surface reflectance observations. The land cover maps include 22 classes, which are defined by the United Nations (UN) Land Cover Classification System (LCCS) and have a resolution of approximately 300 m. See Figure 1 for an overview of the DEM and land use map of the Volta basin.

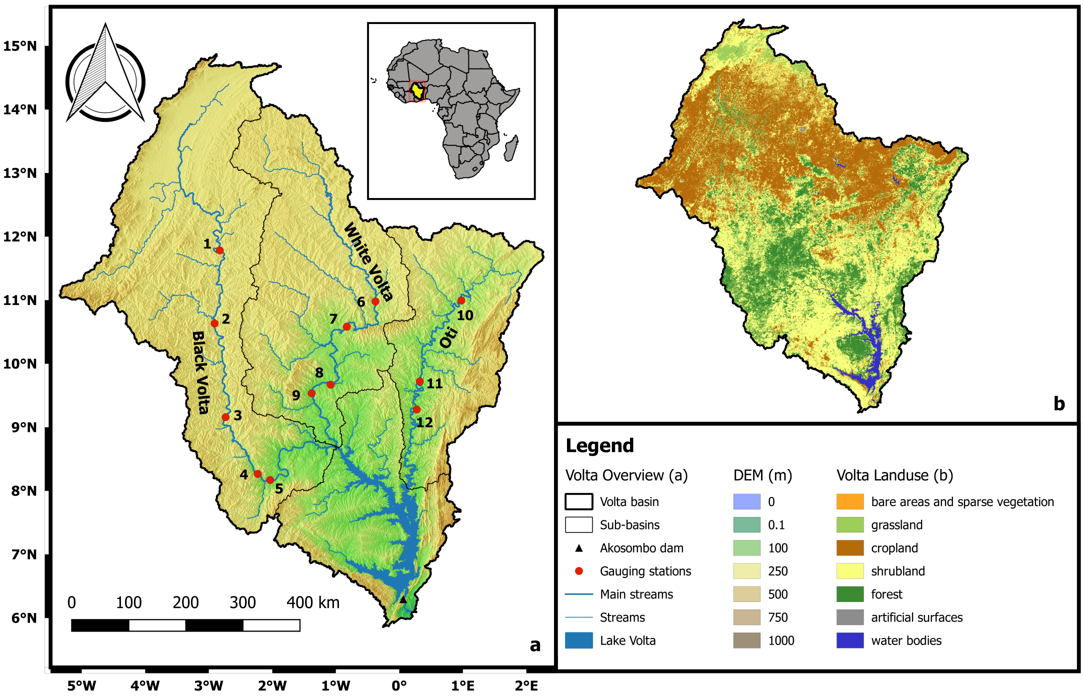


Figure 1: (a) DEM of the Volta basin and its most important sub-basins with the main rivers and streams, Lake Volta and the locations of the gauging stations. The names of the gauging stations are: 1. Boromo, 2. Lawra, 3. Chache, 4. Bui Amont, 5. Bamboi, 6. Yaragu, 7. Pwalagu, 8. Nawuni, 9. Daboya, 10. Porga, 11. Saboba and 12. Sabari. (b) Land use map of the Volta basin.

2.1.2 Dynamic input data: Precipitation and PET

Precipitation

Since a dense network of high quality precipitation observation stations is missing in the Volta basin, RS data offers a solution. Several remotely sensed precipitation products are available, each with their own ad- and disadvantages, depending on the application. Some datasets are more suitable for application in drought monitoring while other datasets can best be used in models developed for flood warnings (Stisen & Sandholt, 2010; Dembele & Zwart, 2016). Research has also shown that regional datasets tend to yield higher performance when applied in hydrological models within that specific region compared to global datasets (Stisen & Sandholt, 2010). Therefore the choice for a specific precipitation dataset can really affect the simulation results. Next to ground-based and RS precipitation observations, there are also reanalysis datasets, which model the precipitation of the past decades. Most precipitation products nowadays do not include only ground-based or remotely sensed observations, but are a combination of those two, and possibly also reanalysis data. Datasets that actually are a combination of different types of observations tend to perform better than datasets including only one type of data (Stisen & Sandholt, 2010; Dembele & Zwart, 2016). RS observations are generally modified (or even calibrated) to match ground-based observations because the spatio-temporal pattern of RS observations is mostly accurate, but the absolute values are not.

The properties of RS precipitation products are mostly determined by the observation technique used. Two common methods for observing precipitation from satellites exist. The first method uses thermal infrared (TIR) sensors to observe brightness temperatures at the top of the cloud deck. From this temperature it can be deduced if it is raining beneath the cloud deck or not. This technique therefore accurately predicts if it is raining, but since the sensor does not penetrate the cloud deck, it tells very little about how much water is actually falling down. A second method uses passive microwave (PM) observations, which do penetrate the cloud deck and are therefore much better at estimating rainfall amounts. However, this technique has a much poorer spatial and temporal resolution compared to TIR sensors (Dembele & Zwart, 2016).

The precipitation dataset used in this study is the Climate Hazards group Infrared Precipitation with Stations (CHIRPS) dataset (Funk et al., 2015). This dataset is based on Cold Cloud Duration (CCD) estimates, calibrated with the Tropical Rainfall Measurement Mission (TRMM) Multi-satellite Precipitation Analysis (TMPA) product, which combines several satellite precipitation observation products. The CCD product is based on TIR sensors and the rainfall amounts from those observations are compared to rainfall observations from TMPA to be able to infer rainfall amounts from CCD observations. Also station data is included in the blending algorithm of the datasets. CHIRPS is also very suitable for operational applications like reservoir management because it is a near real-time dataset. This means that observations are available very soon after the event. CHIRPS data are available from $50°N$ to $50°S$ in the period 1981 till present on different spatial en temporal resolutions. In this study the 0.05° daily resolution product of Africa was used. This product is compared to station data in a point-to-pixel analysis. The results of this analysis can be found in Appendix B.

PET

While precipitation is a directly observable quantity, this is much less true for potential evapotranspiration (PET). Evapotranspiration (ET) is a bulk term and is the sum of all types of evaporation (interception evaporation, soil evaporation and open water evaporation) and transpiration, which are two physically distinctive processes (Savenije, 2004). The PET is, given the energy available for ET, the maximum amount of ET that will occur if enough water is available. PET cannot be directly measured so its value needs to be modelled or estimated using physical relationships via other observed meteorological quantities, like temperature. In this project the choice was made to use estimations of PET via (a simplification of) a physical relationship, rather than a model. This is done to stay close to the observations and to prevent using modelled data that is possibly based on the same or similar data as the data used in the model assessment, like soil moisture. This also avoids using the output of one model as input for another model, which may include an accumulation of model uncertainties.

PET estimation for application in hydrological models is generally done using one of the following methods; The Penman-Monteith method (Monteith, 1965), the Priestly and Taylor method (Priestley & Taylor, 1972) the Makkink method (Makkink, 1957) or the Hargreaves and Samani method (Hargreaves & Samani, 1985). The Penman-Monteith method needs a lot of meteorological input variables, which are generally not available from RS observations. The Priestley and Taylor method is a simplification of the Penman-Monteith method and uses radiation as input to estimate PET. The Makkink method is used a lot in the Netherlands. The method that will be used in this research is the Hargreaves and Samani method, because the only input variable needed is temperature and because good results were obtained using this method in other research projects (Demirel, Mai, et al., 2018; Dembele, Hrachowitz, et al., 2020).

Temperature is a quantity that can be observed using RS techniques and therefore different RS temperature datasets exist. The available datasets differ in their temporal and spatial coverage and their temporal and spatial resolution. The dataset that is used in this study is the E2O dataset (Sperna Weiland et al., 2015). This is a high resolution (0.05°) global daily temperature dataset with high temporal coverage (1979 - 2012). The dataset is a downscaled version of the WATCH Forcing Data ERA Interim (WFDEI) dataset (Weedon et al., 2014) and is developed by Deltares. The data can be accessed with options to already derive the PET from temperature, using either the Penman-Monteith method, the Priestley and Taylor method or the Hargreaves and Samani method. In this research, the PET estimated using the Hargreaves and Samani method was used. For practical model applications, it is also convenient to have all dynamic input data at the same resolution and with the same spatial coverage, because this input data is generally the limiting factor for the model resolution. In this research project, the model resolution is therefore set at a daily timescale with a 0.05° cell size, in order to use the dynamic input at their original resolution.

2.1.3 Calibration / Evaluation data: Q, TWSA, NDVI and SM

This research uses multiple (satellite) datasets for which the parameter space of the hydrological model is optimized by calibration. This approach is called multivariate calibration and has shown its added value in the last decade (Demirel, Koch, et al., 2018; Dembele, Hrachowitz, et al., 2020; Nijzink et al., 2018; Rakovec, Kumar, Attinger, & Samaniego, 2016; Rakovec, Kumar, Mai, et al., 2016; Stisen et al., 2018). Apart from streamflow, all the datasets used for calibration will be RS datasets, because by doing so this research project will contribute to doing hydrological predictions in ungauged basins (Hrachowitz et al., 2013) from space. However, it has been shown in recent research (Dembele, Hrachowitz, et al., 2020; Demirel et al., 2019; Rakovec, Kumar, Mai, et al., 2016) that it is not yet possible to exclude ground-based streamflow observations from the calibration procedure without large reductions in streamflow performance, or without a gauged donor basin from which the parameter set can be transferred (Rakovec, Kumar, Attinger, & Samaniego, 2016; Rakovec, Kumar, Mai, et al., 2016). This section elaborates on the calibration and evaluation data selected for this research. The choice for a specific RS data product was made depending on the spatial and temporal availability and resolution of the products in combination with results obtained in previous research. An overview of all selected datasets for this research is given in Table 3. First the sources, length and quality of the streamflow observations are discussed, and later the details of the selected RS datasets, including a basic description of the techniques used for doing the observations.

Streamflow
The streamflow data was obtained from different sources. The first source is the Global Runoff Data Centre (GRDC, 2021), which is a database with long historical timeseries of worldwide streamflow observations maintained for scientific research. The second source is the SIEREM database (Boyer et al., 2006) from the research unit HydroSciences Montpellier. The SIEREM database contains all kinds of hydro-meteorological data timeseries for Africa. The third source is Mr. Dembélé, lead author of Dembele, Hrachowitz, et al. (2020), who used local streamflow timeseries in his research and who is willing to share some of his data for this research project.

The start of all timeseries is chosen at 01-01-2000 since before 2000, RS observations are scarce. All timeseries contain daily observations. The end of all GRDC streamflow timeseries in the Volta basin is at 28-02-2007. The SIEREM database has some timeseries that end on the same day as the GRDC timeseries, but also contains timeseries with data until the year 2016. The timeseries from Dembele, Hrachowitz, et al. (2020) are between 13 and 16 years long and mostly cover the periods with data from the GRDC ad SIEREM database.

Not only the length, but also the quality of the observations differs a lot per streamflow timeseries. Some timeseries miss a lot of individual observations or even completely lack some years of observations. This is especially true for the timeseries from the GRDC and SIEREM databases.

In Dembele et al. (2019), a gap-filling procedure is described, which is applied to the streamflow data obtained from Dembele, Hrachowitz, et al. (2020). This results in timeseries without gaps, as can be seen in Table 1, in which an overview of the streamflow data used in this study is given.

The choice was made to only use the streamflow data until 28-02-2007. This was done because this was the period with the least missing values and the only period in which all the streamflow timeseries had observations. The period was split-up in a calibration period from 01-01-2000 till 31-12-2003 and an evaluation period from 01-01-2004 till 28-02-2007. The evaluation period is a bit shorter than the calibration period but the data in this period is also more complete for most of the timeseries. A selection of discharge stations was made to use in the project based on the amount of observations available in the model period and their spatial distribution within the Volta basin. This resulted in the selection of stations as is presented in Table 1. The names and locations of the stations can be found in Figure 1.

Table 1: Streamflow data

Station	**SB**	**Source**	**Fill perc cal (%)**	**Fill perc eval (%)**
Boromo	BV	SIEREM	67.9	69.61
Lawra	BV	SIEREM	94.25	97.75
Chache	BV	Dembélé	100	100
Bui Amont	BV	Dembélé	100	100
Bamboi	BV	SIEREM	87.75	100
Yaragu	WV	GRDC	66.67	100
Pwalagu	WV	GRDC	75.29	100
Nawuni	WV	GRDC	97.94	98.27
Daboya	WV	Dembélé	100	100
Porga	Oti	SIEREM	100	100
Saboba	Oti	Dembélé	100	100
Sabari	Oti	GRDC	93.02	99.48

SB: Sub-Basin, BV: Black Volta, WV: White Volta, Dembélé: Moctar Dembélé from Dembélé et al., (2020), Fill perc: Filling percentage; The percentage of days with an observed Q value in the calibration (cal) and evaluation (eval) period

TWSA

The terrestrial water storage anomaly is the difference in the quantity of water in a certain land area with respect to a specified mean. The simulated water storage anomaly corresponds to the quantity of water in all reservoirs of the conceptual hbv model (See Figure 4), with respect to the mean amount of water in all those reservoirs in the model period. This very same water storage anomaly is also observed from space by the GRACE mission.

The GRACE mission quantifies both the static and time-variable gravity anomalies of the Earth. All gravity anomalies are caused by mass variations. The static gravity anomalies (compared to a smooth round sphere) are caused by for example mountains (positive anomalies) or ocean trenches (negative anomalies). In the first 111 days of the mission, GRACE produced a static gravity anomaly map. Time-variable gravity anomalies are mostly caused by mass changes in water, which circles between the oceans, atmosphere, land and glaciers and ice caps. The GRACE satellite actually consists of two satellites that orbit the Earth behind each other. Positive gravity anomalies cause an acceleration of the first satellite (and later also in the second satellite) and this increases the distance between the two. For negative gravity anomalies, it is the other way around. This distance is continuously measured and turned into mass anomalies. These mass anomalies can be attributed to one of the different time-variable causes mentioned above, such as the TWSA (Swenson, 2012; Landerer & Swenson, 2012; Swenson & Wahr, 2006).

3 solutions of the GRACE TWSA, in which the water storage mass anomaly over land is filtered out of the total time-variable mass anomaly, exist. These solution are provided by the Jet Propulsion Laboratory (JPL), the German Research Centre for Geosciences (GFZ) and the Center for Space Research of the University of Texas, Austin (CSR). The mean of these three solutions reduces noise the most, compared to individual solutions (Sakumura et al., 2014) and therefore the mean of all three solutions was used in this research project. The baseline period, with respect to which the anomalies are determined was changed to the active period of GRACE within the model period (April 2002 till February 2007).

NDVI

NDVI is the normalized difference vegetation index and is effectively a measure of the 'greenness' of an area. The index thereby gives an indication of the health of the vegetation. The NDVI dataset adds some dynamics to the model, and can possibly correct the E_a-term (See Figure 4) where the land use map is wrong. The NDVI is calculated using the reflectance of red light in the visible spectrum and the reflectance of near-infrared (NIR) light. Vegetation strongly absorbs light in the visible spectrum for photosynthesis, but strongly reflects near-infrared light. The NDVI is calculated according to Equation 1, in which the Red and NIR refer to the reflectance values (between 0 and 1) of light in the NIR and Red spectrum, resulting in an NDVI value between -1 and +1. The higher the index value, the greener an area. An indication for several land cover classes corresponding the certain NDVI ranges is given in Table 2.

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{1}$$

Table 2: NDVI land cover classification

NDVI range	**Land cover class**
-0.28 - 0.015	Water
0.015 - 0.14	Built-up
0.14 - 0.18	Barren land
0.18 - 0.27	Shrub and Grassland
0.27 - 0.36	Sparse Vegetation
0.36 - 0.74	Dense Vegetation

NDVI range indications for several land cover classes according to Akbar et al. (2019).

An NDVI dataset will be used to assess the AET simulation from the model. NDVI data is used because AET cannot be directly observed from satellites. AET estimations are therefore always based on models while NDVI is directly based on RS observations. These AET models, like GLEAM (Miralles et al., 2011), are often based on datasets like precipitation, temperature and soil moisture, and are therefore not independent observations of the datasets already used in this project. They also accumulate the errors and uncertainty associated with each of the datasets. In this study, the proven positive correlation between NDVI and AET (Wang et al., 2007; Cihlar et al., 1991; Seevers & Ottoman, 1994; Islam & Mamun, 2015; Kerr et al., 1989; Szilagyi et al., 1998) will be used. An additional advantage of NDVI data is that the spatial resolution is often higher than the output of evaporation models like GLEAM (Miralles et al., 2011), which means that model output can possibly be directly compared to NDVI observations. The downside of using RS NDVI observations is that these are often contaminated by clouds. Clouds hinder the observation of the reflectance of NIR and Red light when they are present in the specific area observed by the sensor. They lower the NDVI observations and introduce a lot of noise in the timeseries.

The specific dataset used in this study is the NDVI climate data record (CDR) from the National Oceanic and Atmospheric Administration (NOAA) (Vermote, 2019). This is a climate record constructed from a composite of NDVI observations from different satellites, all observed with an Advanced Very High Resolution Radiometer (AVHRR). The product provides daily NDVI observations on a 0.05° resolution, meaning that the model output can directly be compared to NDVI observations. Clouds are filtered out of the observations as good as possible using a dynamic filter that sets NDVI observations more than 0.05 lower than the last observation to a missing value, and does the same for NDVI observations more than 0.15 higher than the last observation. This approach is justified because it is assumed that the 'greenness of an area is cannot go up and down so much on such a short timescale. Also a centered 30-day running mean was applied to the data to filter out the noise, resulting from among other factors the observation angle of the satellite. Water bodies are also set to missing values because the assumed positive correlation between NDVI and AET does not hold for these surfaces. Water bodies namely have high AET, but low NDVI values (See Table 2).

SM

Soil moisture is the amount of water in the upper soil layer. This soil layer is also know as the unsaturated zone or root-zone. More specifically, the root-zone is the zone till the depth that plants are able to extract water from and the unsaturated zone is the zone till the groundwater table. The unsaturated zone is one the most important components of the hydrological system because it distributes the water in the model between evaporation, storage, direct runoff and slower runoff processes. Therefore it is essential to include SM data in the model calibration and evaluation, to assess the performance of the model to simulate this amount of water correctly. The amount of moisture in the soil can be observed from satellites by passive and active remote sensing using microwave radiation. Microwave radiation has a relatively large wavelength, which has as advantage that it penetrates clouds. Soil moisture observations are therefore not hindered by clouds, like NDVI observations, which are based on optical remote sensing techniques.

Microwave radiation is naturally emitted by the moisture in the soil. When this radiation is observed in space, this is called passive remote sensing. When microwave radiation is actively being transmitted and the reflectance of this signal is measured by the satellite, this is called active remote sensing. The sensors that measure microwave radiation need to be very large because of the large wavelength of the signal. This is the main reason that the spatial resolution of the soil moisture signal is generally low. Another disadvantage of remotely sensed soil moisture observations is that these observations are only representative of the top 1-5 centimeters of the soil layer. This representative depth is directly related to the frequency band used to observe the signal. This observed signal is generally converted to a brightness temperature, from which the amount of water in the soil is being derived via for instance the Land Parameter Retrieval Model (de Jeu, 2003).

In this research project the soil moisture dataset used is the ESA CCI SM product (v05.2) (Dorigo et al., 2017; Gruber et al., 2017; Gruber et al., 2019). This product is based on observations of almost all microwave satellite sensors (both active and passive) that are available. It has a daily temporal and a 0.25° spatial resolution and is available in the full model period and area. These SM observations will be converted to Soil Water Index (SWI) values, which are representative for the complete root-zone of the area, using a method explained in subsubsection 3.3.3 and Appendix D. These SWI values can then directly be compared to the simulated amount of water in the soil moisture reservoir (See Figure 4).

Table 3: Overview of the data used in this study

Variable	Product	Spat. res.	Temp. res.	Reference
Static input data				
DEM	GMTED 2010	225 m	Static	(Danielson & Gesch, 2011)
Land use	Globcover 2009	300 m	Static	(Bontemps et al., 2011)
Dynamic input data				
Precipitation	CHIRPS	0.05	Daily	(Funk et al., 2015)
PET	E2O	0.05	Daily	(Sperna Weiland et al., 2015)
Calibration / Evaluation data				
Streamflow	See sect. Streamflow	Point	Daily	(GRDC, 2021; Boyer et al., 2006) (Dembele, Hrachowitz, et al., 2020)
TWSA	GRACE TWSA	1°	Monthly	(Swenson, 2012)
NDVI	NOAA CDR NDVI	0.05°	Daily	(Vermote, 2019)
SM	ESA CCI SM	0.25°	Daily	(Gruber et al., 2019)

Spat. res.: Spatial resolution. Temp. res.: Temporal resolution

2.2 Study Area

The choice for the study area was based on two reasons, namely size and the (in)availability of data. RS data is available at a wide range of spatial resolutions, depending on the type of data and the RS technique used. Because of the very coarse resolution of some RS data products, the basin to which the hydrological model is applied as a case-study needs to be very large. Otherwise it is more suitable to implement a specific dataset in a lumped manner. Using RS data is also justified best in regions with little meteorological information, because, although the quality of RS products may not be as good as of ground-based measurements, RS data is available almost anywhere on Earth. The Volta basin in Western Africa (See Figure 1) is such a large and data-scarce catchment and is therefore chosen as a case-study for this research project. The basin also includes a large reservoir, Lake Volta, for which the developed hydrological model could be used as a real business case for reservoir management applications.

The Volta basin is an enormous basin of over 400.000 km^2 covering large parts of Ghana and Burkina Faso, and smaller parts of Ivory Coast, Togo, Benin and Mali. The most important river sections are the Black Volta, White Volta, the Red Volta (branch of the White Volta) and the Oti. All these river sections in the end drain into Lake Volta, which is the largest man-made reservoir of the world (van Zwieten et al., 2011), formed by the Akosombo dam. The area around Lake Volta is called the Lower Volta. After the Akosombo dam, the Volta drains into the Atlantic Ocean in the Gulf of Guinea. The basin is characterized by a relatively flat topography. Land cover is dominated by shrubland (55%) and cropland (35%), with some forests (7%) and grassland (2%). A small part of the basin is identified as water body (1.4%), which is mainly due to Lake Volta. Bare and urban areas are very small compared to the other land cover types (See Figure 1).

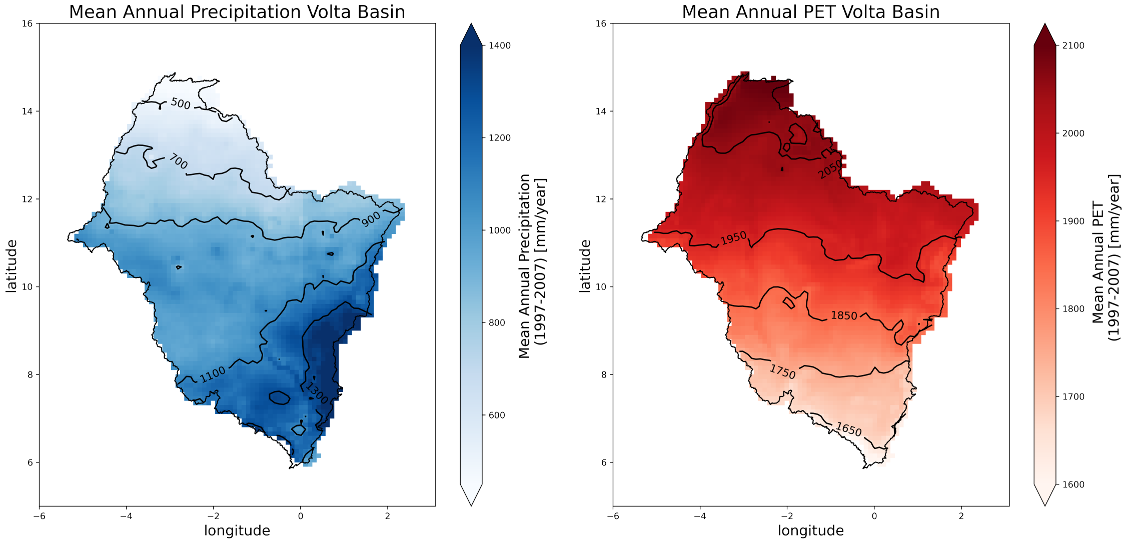

Mean Annual Precipitation Volta Basin
Mean Annual PET Volta Basin
latitude
longitude
Mean Annual Precipitation (1997-2007) [mm/year]
Mean Annual PET (1997-2007) [mm/year]
500
700
900
1100
1300
1650
1750
1850
1950
2050
16
14
12
10
8
6
−6
−4
−2
0
2
1400
1200
1000
800
600
2100
2000
1900
1800
1700
1600

Figure 2: Annual mean spatial pattern of precipitation (left) and PET (right) in the spin-up and model period

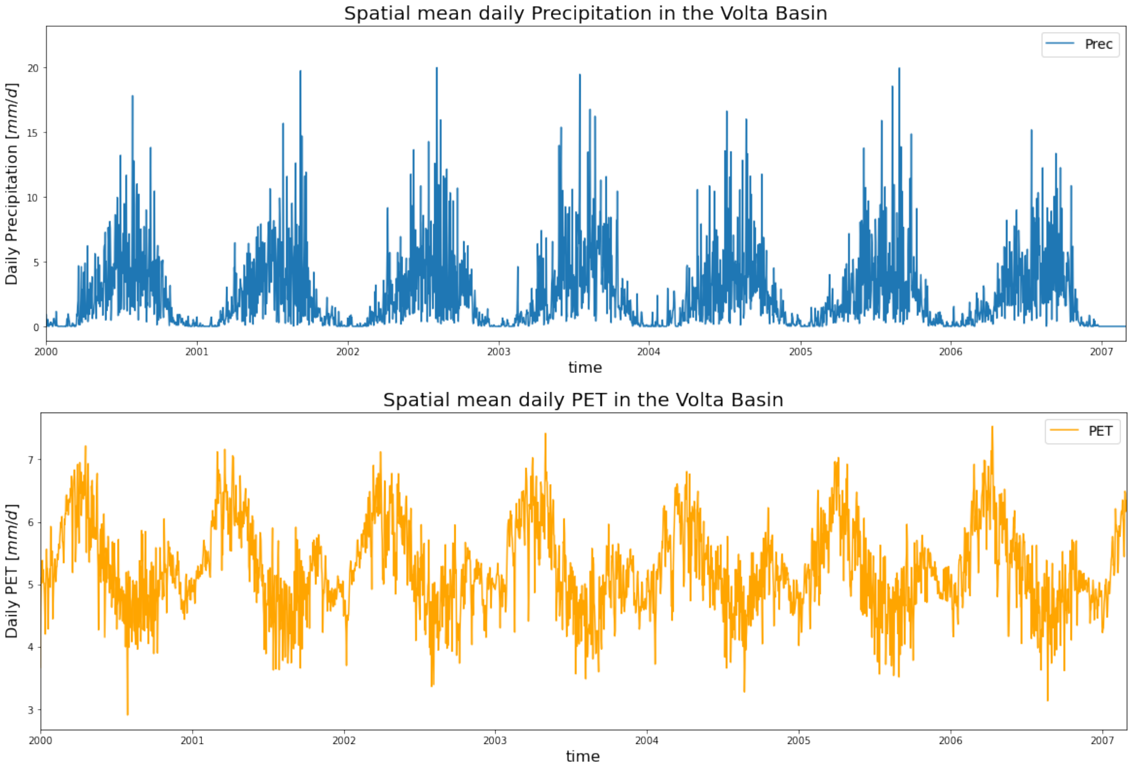


Figure 3: Mean temporal pattern of precipitation (top) and PET (bottom) in the model period

The precipitation and potential reference evaporation is highly variable in the Volta basin in both space and time. In Figure 2, the spatial pattern of the annual mean precipitation and PET is shown. A strong north-south gradient can be observed, with relatively little precipitation (±500 mm) in the northern part of the basin, and high precipitation ($>$ 1300 mm) in the south-eastern corner of the basin. For the PET, this strong gradient also exists but with the most extreme PET values ($>$ 2050 mm) in the north, and the least extreme values (±1700 mm) in the south of the basin. Figure 3 shows the basin-mean temporal pattern of precipitation and PET on a daily time-scale. It can be observed that there is a very strong seasonal precipitation pattern, with a wet season from April to October and a dry season from November till March. In the dry season, almost no precipitation events occur. This seasonal cycle is also present in the PET, but the pattern is again reversed (high PET values in the dry seasons and lower values in the wet season) and the seasonality is also less strong.

Not only the precipitation and PET, but also the RS calibration and evaluation data show interesting spatio-temporal patterns in the Volta basin in the model period. The spatial pattern of the TWSA in the Volta basin is shown for two observations in the year 2004 in Appendix C in Figure 25. The mean temporal spatial pattern of this dataset is not shown here because it contains no information. This is because the mean of an anomaly in a certain period is per definition zero. Instead, the spatial TWSA patterns of May and October 2004 are shown. It is observed that the whole basin has a negative anomaly at the end of the dry season (May) and the whole basin has a positive anomaly at the end of the wet season (October). However, the variation is larger in the south-eastern corner of the basin. The variation in the northern part of the basin is relatively small. This can be explained by the fact that this is also the region with the lowest precipitation. The plot at the top of Figure 27 in Appendix C shows the mean temporal TWSA pattern for the Volta basin. A strong seasonal cycle can be observed with values between -150 to $+150$ mm w.r.t. to the mean terrestrial water storage.

In Figure 26 in Appendix C also the mean spatial NDVI pattern is shown. The presented NDVI pattern is based on the filtered NDVI observations as was explained in subsubsection 2.1.3. The NDVI is highest in the utmost south of the basin, east and west of Lake Volta, while NDVI is lowest in the northern part of the basin. Temporal mean NDVI values indicate very little dense vegetation in the basin, but do indicate a lot of shrub- and grassland and sparse vegetation (Table 2). The plot in the middle of Figure 27 in Appendix C shows the spatial mean temporal NDVI signal of the Volta basin. Generally, two peaks in NDVI values can be observed each year, of which the second is always higher. These peaks can be explained by the fact that farmers in the south of the basin can grow 2 crops per growing season because sufficient precipitation is available, while farmers in north can grow only 1 crop per growing season (Padi, 2018). The difference in the height of the peaks can possibly be explained by the cultivation of different crops in the first and second growing season, and the more gradual rise of mean NDVI values of non-cropland areas, like forest, grassland and shrubland during the wet season.

Figure 26 in Appendix C shows the temporal mean spatial SM pattern of the Volta basin. Generally, SM values are lower in the north of the basin and higher in the south of the basin, but the pattern is not as distinct as for instance the gradient in precipitation. Some spots with higher or lower soil moisture values are also observed. The spatial mean temporal SM pattern shown in Figure 27 in Appendix C does show a very clear seasonal pattern, with the highest SM values at the end of the wet season and the lowest at the end of the dry season.

3 Methodology

This Chapter first describes the wflow hbv model that was used in this project in subsection 3.1. In subsection 3.2 the model implementation and the different scenarios applied to the model are discussed and finally the performance assessment procedure is explained in subsection 3.3.

3.1 Model description: wflow hbv

wflow hbv

The wflow hbv model was used in this project. wflow is a conceptual distributed hydrological modelling framework from the OpenStreams project of Deltares. The framework includes several established hydrological models like sbm, gr4, pcrglobwb, topoflex, and hbv (Schellekens, 2021). The hbv model was originally developed by Sten Bergström in 1972 but a revised version of the model was published in 1997 which is known as the HBV-96 model (Lindström et al., 1997). This version of the model is the now widely known HBV model and was also the basis for the wflow hbv model. The hbv model was chosen because this model is very suitable for calibration, which is the essence of this study. With around 10 free parameters in a lumped version of the model, there is plenty of freedom in the parameter space. This may also be a reason why the hbv model, which was originally developed for Scandinavian conditions, has been applied in various environments all over the world (See Appendix A (Wetterhall, 2014)).

The wflow hbv model is a distributed version of the hbv model, implemented in a PCRaster python framework. Like all conceptual hydrological models, the wflow hbv model is a bucket-type model, that consists of a series of buckets with different inflow and outflow options, of which each corresponds to 1 or more hydrological processes. The model consists of 4 routines, namely the snow routine, the interception routine, the soil routine and the runoff response routine. Since the model is applied in a tropical savanna / semi-arid environment in this study, the snow routine is not used. After the runoff response routine, the water is routed downstream using a kinematic wave approach. An overview of the wflow hbv model is given in Figure 4.

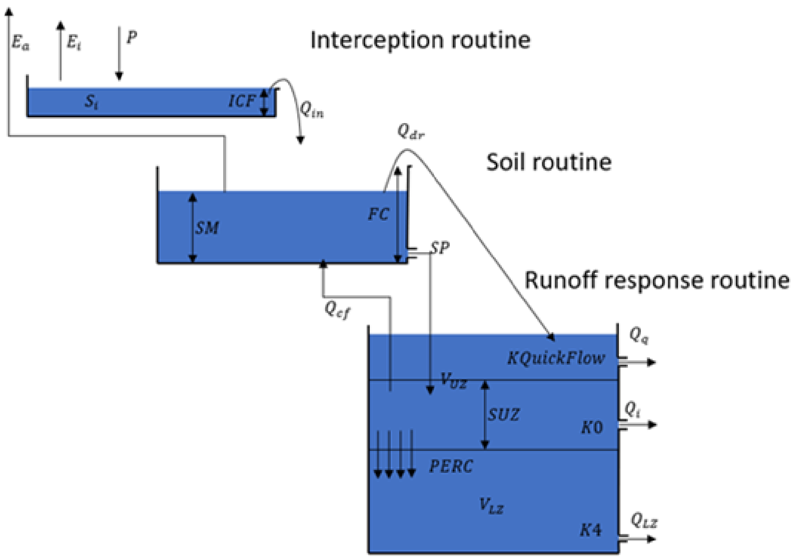


Figure 4: The wflow hbv model with its interception, soil and runoff response routines. See Table 4 for an explanation of the acronyms used in this figure.

Interception, soil and runoff response routines

The first routine of the wflow hbv model is the interception routine. Precipitation P falls in the interception reservoir with capacity ICF. The interception reservoir is a conceptualization of the capacity of leaves, branches and other parts of vegetation to intercept and store the precipitation before it reaches the soil. Water that evaporates from the interception reservoir is called interception evaporation E_i. The maximum amount of water that can evaporate is either limited by the PET or the storage in the interception reservoir S_i. If the water storage in the interception reservoir exceeds its capacity, water will flow directly into to the soil routine (Q_{in}).

The soil routine represents the unsaturated part of the soil, where several important hydrological processes are active. The input for this routine is the overflow (Q_{in}) from the interception routine, which flows directly into the soil moisture reservoir with capacity FC. The amount of water in the reservoir is given by SM. If FC is exceeded, then the water flows immediately to the runoff response routine as direct runoff (Q_{dr}). The net input to the soil moisture reservoir I_{net} is the difference between Q_{in} and Q_{dr}. This net input, together with the seepage parameter β and the fraction of SM over FC determine the seepage through the soil layer SP. Evaporation from the soil routine is called actual evaporation E_a and is dependant on the evaporation factor LP, which determines from which fraction of SM over FC the actual evaporation from the soil is equal to the PET that is left after correcting for interception evaporation. A last small flux into the soil moisture reservoir is the capillary flux Q_{cf}. In reality, this is a flux from below the groundwater table back to the unsaturated zone. In the wflow hbv conceptualisation, the water comes from the runoff response routine and the amount of capillary flux is determined by the maximum capillary flux parameter C_{flux} and the shortage of water in the soil moisture reservoir. An overview of the equations used in the wflow hbv model can be found in Table 7.

The runoff response routine is the last routine of the wflow hbv model and consists of a reservoir divided in an upper and lower zone, UZ and LZ. The upper zone is again divided into two parts. The seepage flux from the soil routine is input for the lower part of the upper zone with capacity SUZ. However, per timestep a maximum amount of water ($PERC$) can percolate downwards into the lower zone. Capillary flow also takes place from this lower part of the upper zone. When the storage in this part, SUZ, is exceeded, water ends up in the upper part of the upper zone. Also direct runoff from the soil routine always flows directly into this part of the runoff response routine. The goal of this 3-layered runoff response routine is to simulate quick flow Q_q from the upper part of the upper zone, interflow Q_i from the lower part of the upper zone, and slow flow Q_{LZ} from the lower zone. Flow generation from these three reservoirs is determined via linear reservoir coefficients, $K_{QuickFlow}$ for Q_q, $K0$ for Q_i and $K4$ for Q_{LZ}. An overview of the free parameters in the routines is given in Table 4.

Kinematic wave module and input correction factors

After the runoff response routine, the water is routed downstream via a kinematic wave module. River cells are defined as cells with a streamorder (Strahler, 1952) of 4 or higher. Other cells are land cells. The parameters for the kinematic wave module consist of the Manning's roughness coefficients for flow over land (N) and in the river (N_{River}). An overview of the parameters of the kinematic wave model is given in Table 5.

The precipitation and potential evapotranspiration input can both be corrected with a linear correction factor, P_{corr} and E_{corr} respectively, if it is plausible that precipitation input is biased. The potential evaporation input is often an estimation for a well-watered reference grass. The PET can be corrected for more specific types of vegetation via the linear $CEVPF$ factor. On wet days, the PET can be restrained by using the exponential precipitation correction factor EPF. Usually, these correction factors are not used. An overview of the input correction factors is given in Table 6. In Table 7, the formulas of the corrections are given.

Table 4: Free parameters in the wflow hbv model

Parameter	Unit	Explanation	Dependant on	Parameter range
ICF	mm	Interception Capacity	LU non-forest	[0.3 - 0.8]
			LU forest	[0.8 - 1.2]
$CEVPF$	-	Factor connecting PET per LU	LU non-forest	[0.9 - 1.1]
			LU forest	[1.1 - 1.75]
FC	mm	Soil reservoir capacity	based on ICF non-forest	[400 - 650]
			based on ICF forest	[350 - 600]
β	-	Seepage parameter	-	[2.0 - 3.5]
LP	-	Evaporation factor	-	[0.25 - 0.55]
Q_{cf}	mm/d	Max capillary flux	-	[0.01 - 2.5]
$PERC$	mm/d	Threshold UZ/LZ	-	[3.5 - 6.5]
SUZ	mm	Quickflow threshold	-	[10.0 - 25.0]
$K4$	d^{-1}	Q_{LZ} reservoir constant	determined per subcatchment	[0.044 - 0.123]
$K0$	d^{-1}	Q_i reservoir constant	-	[0.10 - 0.30]
$K_{QuickFlow}$	d^{-1}	$K_{QuickFlow}$ reservoir constant	-	[0.30 - 0.90]

Note: LU: Land use. The $K4$ and FC parameter values are not determined via calibration. This means that there are 11 calibration parameters, being $ICFnf$ (non-forest), $ICF2_f$ (forest), $CEVPF_{nf}$, $CEVPF_f$, β, LP, Q_{cf}, $PERC$, SUZ, $K0$ and $K_{QuickFlow}$. These parameters values are applied to the whole Volta basin, with the exception of ICF and $CEVPF$ being implemented in a spatially distributed manner based on land use.

Table 5: Parameters in the routing model

Parameter	Unit	Explanation	Dependant on	Parameter value
N	$T/L^{\frac{1}{3}}$	Manning's N value for land	LU + LS wetland	0.055
			LU + LS forest	0.065
			LU + LS nf hillslope	0.028
			LU + LS nf plateau	0.030
N_{River}	$T/L^{\frac{1}{3}}$	Manning's N value for rivers	streamorder 4	0.030
			streamorder 5	0.028
			streamorder 6	0.026

Note: LU: Land use, LS: Landscape, nf: non-forest, SC: subcatchment. The routing parameters are fixed and are not used in the calibration of the model.

Table 6: Correction factors the wflow hbv model

Parameter	Unit	Explanation	Dependant on	Parameter value
P_{corr}	-	Linear P correction factor	-	1
E_{corr}	-	Linear PET correction factor	-	1
EPF	d/mm	wet day PET correction factor	-	0

The correction factors are not used in the implementation of the wflow hbv model in this study.

Table 7: Overview of all formulas of the wflow hbv model

Symbol	Unit	Formula
Input correction formula's		
P	mm/timestep	$P = P_{corr} * P$
PET	mm/timestep	$PET = e^{-EPF*P} * E_{corr} * PET$
PET	mm/timestep	$PET = PET * CEVPF$
Interception routine		
Q_{in}	mm/timestep	$Q_{in} = max(S_i + P - ICF; 0.0)$
S_i	mm	$S_{i+1} = S_i + P - Q_{in}$
E_i	mm/timestep	$E_i = min(S_i; PET)$
PET_{rest}	mm/timestep	$PET_{rest} = PET - E_i$
S_i	mm	$S_{i+1} = S_{i+1} - E_i$
Soil routine		
Q_{dr}	mm/timestep	$Q_{dr} = max(SM + Q_{in} - FC; 0.0)$
I_{net}	mm/timestep	$I_{net} = Q_{in} - Q_{dr}$
SP	mm/timestep	$SP = \left(\frac{SM}{FC}\right)^{\beta} * I_{net}$
SM	mm	$SM_{i+1} = SM_i + I_{net} - SP$
T_m	mm	$T_m = LP * FC$
E_a	mm/timestep	$E_a = \frac{SM}{T_m} * PET_{rest}; SM < T_m$
E_a	mm/timestep	$E_a = PET_{rest}; SM \geq T_m$
SM	mm/timestep	$SM_{i+1} = SM_{i+1} - E_a$
Runoff respone routine		
ΔV_{LZ}	mm	$min(PERC; SP)$
ΔV_{UZ}	mm	$max(0.0; SP - PERC)$
V_{LZ}	mm	$V_{LZ_{i+1}} = V_{LZ_i} + \Delta V_{LZ}$
V_{UZ}	mm	$V_{UZ_{i+1}} = V_{UZ_i} + \Delta V_{UZ}$
Q_{cf}	mm/timestep	$Q_{cf} = C_{flux} * \frac{FC-SM}{FC}$
V_{UZ}	mm/timestep	$V_{UZ_{i+1}} = V_{UZ_{i+1}} - Q_{cf}$
SM_{i+1}	mm	$SM_{i+1} = SM_{i+1} + Q_{cf}$
Q_{LZ}	mm/timestep	$K4 * V_{LZ}$
V_{LZ}	mm	$V_{LZ_{i+1}} = V_{LZ_{i+1}} - Q_{LZ}$
Q_i	mm/timestep	$Q_i = K_i * min(SUZ; V_{UZ})$
Q_q	mm/timestep	$Q_q = K_q * [max(V_{UZ} - SUZ; 0.0) + Q_{dr}]$
Q_{tot}	mm/timestep	$Q_{tot} = Q_{LZ} + Q_i + Q_q$

3.2 Model Implementation & Scenarios

Classifications: Land use classes, Subcatchments and Landscapes

In a fully distributed model, all cells in the catchment do not only have their own stocks and in- and outgoing fluxes, but also their own parameter sets. However, already for a small number of cells within a catchment, this parameterization setup will result in massive equifinality problems. In wflow models, cells can be classified into groups that show hydrologically similar behaviour. This results in a model setup that can be categorized as somewhere in between semi-distributed and fully distributed. This parameterization is generally based on land use, subcatchments and soil. Each parameter of the wflow hbv model can be given a different value for each different combination of these 3 classes. To avoid equifinality problems, it is essential to keep the number of free parameters as low as possible. However, the model should also be flexible enough to simulate the most important hydrological processes active in each cell. The implementation and the parameterization of the wflow hbv model in this project is explained in this section.

Several land use classes are distinguished using a simplified version of the landcover dataset (Bontemps et al., 2011), namely bare areas, grassland, cropland, shrubland, forests, urban areas and water bodies. Several subdivisions within some of these land use classes can also be discriminated, but instead of subdividing the land use classes, it was chosen to aggregate the classes into two groups, namely forests and non-forests. This was done because forests diverge hydrologically the most from the other classes, especially in their interception capacity, PET and depth of the root zone. Therefore, forest is the most important class to separate from the other classes, which are hydrologically much more similar. Shrubland and cropland, which together make up about 90% of the basin are for example not that different in their hydrological behaviour. These 2 classes have similar interception capacities, potential evapotranspiration and root zone depth, although for cropland these features are also dependant on the timing of the growing season(s). The advantage of having only 2 land use groups is that it keeps the model simple and the parameter space as small as possible, while at the same time creating the needed flexibility to simulate the diverse hydrological behaviour of two different systems. See Figure 5a for a map of the land use classification.

The subcatchments are the second classification on which the parameteriation can be based. The 13 subcatchments are defined by the gauges shown in Figure 1 of which 12 are gauges with actual streamflow observations and 1 is the outlet of Lake Volta, the most downstream point of the Volta basin considered is this project. The high number of subcatchments makes a parameterization based on subcatchments undesirable, because it would lead to a large parameter space relatively fast. See Figure 5b for a map of the subcatchments classification.

The landscape classification was adopted from the FLEX-Topo model (Savenije, 2010) and includes the landscapes wetland, hillslope and plateau. The wetland delineation is based on the height above nearest drainage (HAND) (Rennó et al., 2008). If the HAND is lower than a certain threshold, the cell is categorized as wetland. The hillslope delineation is based on the slope. Cells with a slope above a certain threshold are thereby categorized as hillslopes. Areas that are not classified as wetland or slope are classified as plateau. Thresholds of 5.3 m for HAND (Rennó et al., 2008) and 11% for slope were used in this study, see Table 8. The goal of this landscape classification is to simulate hydrological behaviour in these different conditions as good as possible. Wetlands are known for their high water tables and hence their low depth of the unsaturated zone, while hillslopes generate floods and plateaus replenish the groundwater (Savenije, 2010). See Figure 5c for the landscape classification of the Volta basin.

Table 8: Landscape classification thresholds

Landscape	HAND threshold	slope threshold
Wetland	≤ 5.3 m	-
Hillslope	-	$\geq 11\%$
Plateau	> 5.3 m	$< 11\%$

Model Parameterization

The ICF parameter is parameterized based on land use because the vegetation is assumed to be the most determining factor for the interception capacity. Forests are assumed to have a much higher interception capacity then non-forests. Since forest can hold more water compared to non-forests, it is also assumed that the maximum evapotranspiration of forest is much higher than for non-forests. This results in a higher $CEVPF$ parameter for forest than for non-forests. The soil reservoir capacity FC is parameterized based on land use and subcatchments. The value of FC is estimated from the ICF and a water balance approach, which is explained in detail in subsubsection 3.3.3 and Appendix D. It is assumed that there is no capillary flux in plateaus or hillslopes, only in wetlands, because in this landscape the water table is close to the surface. The $K4$ linear reservoir coefficients are determined for each subcatchment separately from baseflow recession curves (See Appendix E). This configuration of the wflow hbv model leads to a total of 11 free parameters. 7 of those parameters are uniformly applied to the whole Volta basin (β, LP, Q_{cf}, $PERC$, SUZ, $K0$ and $K_{QuickFlow}$), and 2 of those parameters are parameterized based on the 2 land use classes ($ICFnf$ (non-forest), $ICF2_f$ (forest), $CEVPF_{nf}$ and $CEVPF_f$). See also Table 4.

The Manning's roughness coefficient parameterization for land cells is based on combinations of landscape and land use classes. There are relatively high roughness coefficient for wetlands and forests, which are generally densely vegetated, and relatively low coefficients for non-forested hillslopes, and non-forested plateaus. The N-values for the rivers are linked to streamorder, with lower roughness values for higher streamorders. The routing parameter values were derived from Chow (1959) and model performance was found to be relatively insensitive to the routing parameter values. Therefore, the routing parameters were kept constant during calibration at the values given in Table 5.

Model Resolution

The spatial and temporal model resolution is limited by the resolution of the input data. The dynamic input data (P and PET) are available in sub-daily, daily and coarser temporal resolutions and both have a spatial resolution of 0.05°. The static input data is not important for the temporal model resolution and the spatial resolutions of the static input data are higher than that of the dynamic input data. The calibration data is generally available at a daily timescale (except for GRACE), but the spatial resolution ranges between 0.05° and 1.0°. This means that the dynamic input data is the limiting factor. The model is set up to run on a daily timescale and at 0.05° resolution. This spatial resolution is most consistent with the dynamic input data and this temporal resolution allows for comparison of the original (non-aggregated) streamflow timeseries and the RS calibration data with the model output. For this comparison, modelled data can always be upscaled but never downscaled. A spatial resolution of 0.05° also makes it possible to compare model output to NDVI observations at the highest resolution possible. The high spatial resolution is also needed to obtain representative classification maps of land use, subcatchments and landscapes. Using a lower spatial resolution results in unrepresentative land use and landscape classifications, because all details that these static maps contain are then flattened out and averaged over larger areas.

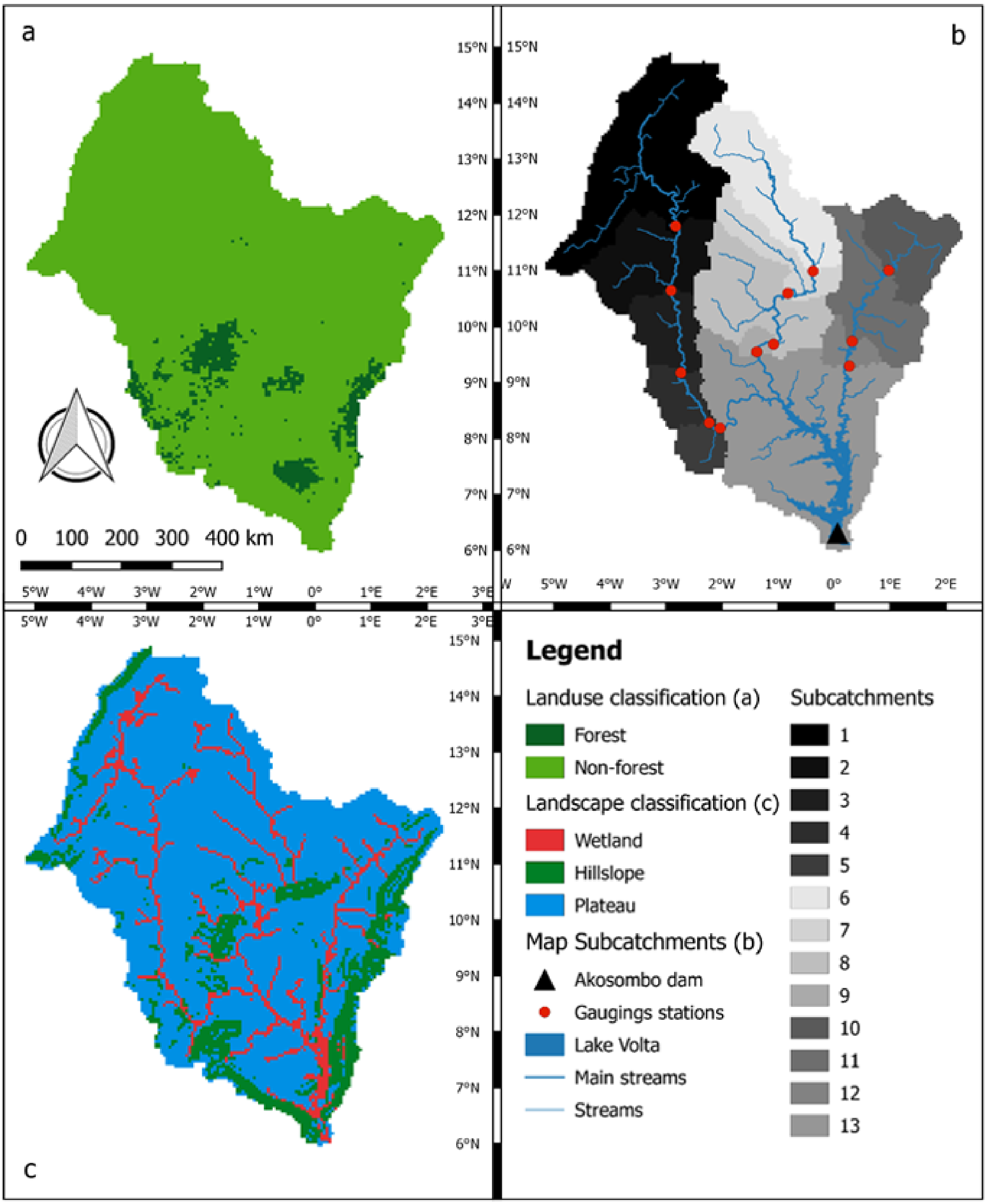


Figure 5: Overview of the (a) land use classification, (b) subcatchment classification, and (c) landscape classification

Model calibration and evaluation periods and catchments

The model period and the modelled area are split-up into two periods and areas to allow for a temporal and spatial model evaluation. The first part of the modelled period is used as spin-up period. This part consists of 3 years, from 1997-01-01 till 1999-12-31. This spin-up period is needed to fill up all the reservoirs and to start each run with different initial conditions, depending on the parameter set. Then follows a calibration period of 4 years, starting at 2000-01-01 and ending at 2003-12-31. The last period is the evaluation period, which starts at 2004-01-01 and ends at 2007-02-28. The performance of the model to simulate Q, the TWSA, AET or SM is assessed separately in the calibration and evaluation period, to examine if a parameter set found in the calibration period also performs well in the evaluation period. This is called the temporal evaluation.

The modelled area is also split-up in a calibration and evaluation area. These areas are formed by two combinations of subcatchments. The subcatchments in the Black Volta (subcatchments 1,2,3,4,5: See Figure 5b) and the Oti (subcatchments 10,11,12: See Figure 5b) are chosen as calibration catchments, while the subcatchments of the White Volta (subcatchments 6,7,8,9: See Figure 5b) are chosen as (spatial) evaluation catchments. The performance of the model to simulate the Q, TWSA, AET or SM is assessed separately in the calibration and evaluation catchments, to examine if a parameter set found in the calibration catchments also performs well in the evaluation catchments. This is called the spatial evaluation. Also the performance of the model to simulate the Q, TWSA, AET and SM is assessed in the evaluation period and evaluation catchments. This is called the spatio-temporal evaluation. An overview of the division in calibration and evaluation periods and catchments described here is given in Table 9 and Figure 6.

Table 9: Overview of the spin-up, calibration and evaluation periods and catchments

Period / Subcatchments	**Calibration catchments (1,2,3,4,5,10,11,12)**	**Evaluation catchments (6,7,8,9)**
Spin-up period 1997-01-01 - 1999-12-31	-	-
Calibration period 2000-01-01 - 2003-12-31	Calibration (optimization)	Spatial evaluation
Evaluation period 2004-01-01 - 2007-02-28	Temporal evaluation	Spatio-temporal evaluation

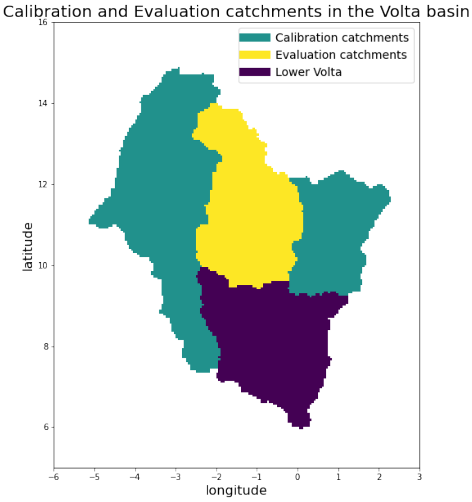


Figure 6: Classification of the Volta basin in calibration and evaluation catchments. The lower Volta was not used because no streamflow data of this area was available.

Calibration Scenarios

Several scenarios are modelled in this study to answer the research question. The first scenario is a baseline scenario using only streamflow timeseries (Q-only) as calibration data. This scenario represents the classical hydrological model, however, already in a distributed configuration. The following scenarios each add one (or more) RS dataset(s) to the calibration procedure, next to streamflow. The results of the baseline scenario will determine which dataset can best be added first and also if it is useful to add each dataset to the calibration. Also multiple RS datasets can be added to the calibration procedure in one and the same scenario. In each of these scenarios the overall model performance is assessed for all 4 sets of calibration and evaluation data, as is described in subsection 3.3, but the model optimization is only performed for a specific combination of calibration and evaluation datasets. In the last scenario, an attempt is made to simulate streamflow using only RS datasets in the calibration (and hence no streamflow data). This scenario is added to see if a good hydrological model performance can be obtained by only using RS data. Thereby, this study contributes to the development hydrological predictions in ungauged basins (Hrachowitz et al., 2013) from space. An overview of the scenarios modelled in this study (selection of the scenarios was done based on the results of the baseline scenario, see subsection 4.1) is given in Table 10.

Table 10: Overview of the calibration scenarios modelled in this study

Calibration Scenarios	**Datasets used for model calibration**			
Scenario name	**Q**	**TWSA**	**NDVI**	**SM**
Scenario 1: Q-only	x			
Scenario 2: Q+SM	x			x
Scenario 3: Q+TWSA	x	x		
Scenario 4: Q+SM+TWSA	x	x		x
Scenario 5: SM+TWSA		x		x

The calibration scenarios given in this table were chosen based on the results of subsection 4.1

3.3 Model assessment: A multivariate approach

In this study, not only the performance of the model in simulating streamflow is assessed, but also the performance of the model in simulating the spatial and temporal patterns of the TWSA, the AET and the SM. The assessment of multiple variables is called a multivariate approach. The model performance on each of the components can be assessed using objective functions, which compare the simulation results to observations and give this comparison a performance score. The more similar the simulation and the observations are, the better the performance score. The objective functions used in this study to assess the performance of the streamflow simulations are discussed in subsubsection 3.3.2, and the objective function used to assess the performance of the model to simulate the spatial and temporal patterns of the RS calibration and evaluation data is discussed in subsubsection 3.3.3. subsubsection 3.3.4 describes how all objective functions are combined into one single objective functions per scenario.

The objective function results can be optimized by calibration. Calibration can be done manually by changing parameter values one by one by hand, but also automatic calibration using calibration algorithms is possible. In a calibration is being searched for the best performing parameter set (within a defined parameter space) for a specific (set of) objective function(s). Many calibration algorithms can be used for model optimization. Some algorithms, like Latin Hypercube Sampling, scan the whole parameter space. And in Monte-Carlo calibration, random sets of parameter values are drawn from the parameter space. Other calibration algorithms reuse the results of former model runs to find global and local optima of the parameter space. One such a calibration algorithm is the Dynamically Dimensioned Search (DDS) Algorithm (Tolson & Shoemaker, 2007), which is used in this study and discussed in subsubsection 3.3.1.

3.3.1 The calibration algorithm: DDS

DDS is an automatic calibration algorithm, especially suitable for calibration of high dimensional models. DDS is able to find the global optimum of a high dimensional parameter space in a relatively low amount of model evaluations. DDS has only one parameter, namely the number of model evaluations and this is also the only stopping criterion. DDS is dynamic in the sense that the total number of model evaluations determines how many parameter set samples DDS uses to determine in which direction in the parameter space it continues its search. A higher number of model evaluations will therefore increase the chance of finding the global optimum. DDS is also dynamic in the sense that the amount of parameter set samples decreases when the algorithm is converging towards an optimum. This means the algorithm changes from a global search method to a more local search method near optima. This process makes DDS a very efficient algorithm. It needs a much lower number of model evaluations than other calibration algorithms like the Shuffled Complex Evolution (SCE) algorithm (Duan et al., 1993) to approach global optima of high-dimensional problems (Tolson & Shoemaker, 2007).

The calibration of a distributed model often is a high dimensional problem, because distributed models tend to have much more parameters than lumped models. In the relatively simple setup of the wflow hbv model in this study (described in subsection 3.2), already 11 free parameters are identified, making the calibration a 11-dimensional problem. To sample the whole parameter space of a 11-dimensional problem, a very large number of model evaluations is needed. However, distributed models also tend to have a much longer run-time than lumped models, because distributed models are essentially a grid of lumped models with different input, parameters, stocks, fluxes, and output in each specific cell and all calculations need to be performed per cell. This makes it undesirable to perform a large number of models evaluations.

Because DDS is a highly efficient optimization algorithm for high-dimensional problems, the algorithm is very suitable for the calibration of distributed hydrological models. DDS is nowadays probably one of the most used calibration algorithms in distributed hydrological model optimization (Becker et al., 2019; Dembele, Hrachowitz, et al., 2020; Huot et al., 2019; Kumar et al., 2013; Rakovec, Kumar, Mai, et al., 2016) and is therefore also chosen as the calibration algorithm to use in this study. According to Tolson & Shoemaker (2007), approximately between 1.000 and 2.000 model evaluations should be sufficient to approach the global optimum of a 11-D calibration problem.

However, the applied number of runs can be much lower than for other known high dimensional problems, such as discussed in Tolson & Shoemaker (2007), because of the fact that the parameters in the model are correlated. This correlation is a consequence of the fact that the parameters together mean to describe a natural system, governed by physical laws. In Appendix D it was already shown that the interception capacity (ICF) and the depth of the unsaturated zone in the soil moisture reservoir (FC) are strongly correlated. The same holds for the other parameters in the model. Thus, the effective parameter space is much smaller than a 11-dimensional problem in which all parameters are completely independent of each other. Therefore, an optimal solution can be found with less runs than one may expect purely based on the total number of parameters.

A test was carried out to find out how many model runs were needed in a calibration run to find a solution with a good balance between result and effort (See Appendix F). It was found that 200 model runs is sufficient to approach an optimum in the parameter space and therefore the DDS algorithm is applied for 200 model runs in every calibration run. A model run is defined here as running the full model period one single time, while assessing all optimization variables for a specific scenario. A calibration run is the full optimization during the 200 model runs. At the end of a calibration run, the overall model performance of the 'best' parameter set of the 200 tested parameter sets is calculated.

The run time of a calibration run of each scenario differs from approximately 8 hours (Q-only) to about a day on the high performance cloud (HPC) machine using a parallel (8-core) implementation of the DDS algorithm. Multiple calibration runs per scenario will give an idea of the uncertainty in the model output and the resulting parameter space. For this project 2 model runs per scenario were done because of time limitations in the project.

3.3.2 Streamflow assessment

The performance of the model to simulate streamflow is assessed using a set of 5 hydrological signatures. The use of a set of hydrological signatures allows the modeler to assess the model performance on different features of the hydrograph. This results in a very complete assessment and in parameter sets that will also perform well in other periods or similar catchments because of a better system representation (Hrachowitz et al., 2014). For consistency, the same objective function, the modified Kling-Gupta Efficiency KGE' (Kling et al., 2012; Gupta et al., 2009) (hereafter referred to as KGE) was used for the assessment of all 5 hydrological signatures. A metric very similar to this objective function will also be used to assess the spatial and temporal pattern simulations of the TWSA, AET and SM to observations.

The Hydrograph

The first hydrological signature that is assessed with KGE is the ordinary hydrograph. The resulting function value is called the KGE_{OH}. The KGE has several advantages over the widely used and established Nash-Suthcliffe Efficiency (NSE) (Nash & Sutcliffe, 1970). NSE can be decomposed into a combination of three components, namely correlation, bias and variability. Problems related to NSE are an underestimation of the variability and large water balance errors and a poor bias estimation in highly variable flow regimes (Santos et al., 2018). 1 minus the Euclidean distance (ED) of a correlation term, a bias term and a variability term makes up the

KGE (Gupta et al., 2009). The modified version of the *KGE* (Kling et al., 2012) is used in this study. The modified *KGE* and its three components are calculated as follows (Kling et al., 2012):

$$KGE = 1 - \sqrt{(r-1)^2 + (\beta - 1)^2 + (\gamma - 1)^2} \tag{2}$$

$$r = corr(Q_{obs}, Q_{sim}) \tag{3}$$

$$\beta = \frac{\mu_{sim}}{\mu_{obs}} \tag{4}$$

$$\gamma = \frac{CV_{sim}}{CV_{obs}} = \frac{\sigma_{sim}/\mu_{sim}}{\sigma_{obs}/\mu_{obs}} \tag{5}$$

In Equation 2 till Equation 5, r is the correlation coefficient between the observed and simulated flow, β is the bias term, given by the ratio of the mean of the simulated and observed flows, and γ is the variability term, given by the ratio of the coefficients of variation of the simulated and observed flows. Q is the flow (m^3/s), μ is the mean of the flow (m^3/s), CV is the coefficient of variation ($-$) and σ is the standard deviation of the flow (m^3/s). *sim* and *obs* denote simulated and observed, respectively.

The modified version of the *KGE* differs from the original version in that not the fraction of the simulated and observed standard deviation was used as the variability ratio, but the ratio of the coefficients of variation. This guarantees that the variability and bias term are not correlated (Kling et al., 2012). The optimal values of all components of the *KGE* and of the *KGE* itself are all 1, and the worst value of *KGE* is $-\infty$. When a model is just as good as taking the mean of the flow as your model, the *KGE* is equal to -0.41 (Knoben et al., 2019).

Baseflow

The assessment of errors in the baseflow are usually underestimated because errors in the baseflow are small compared to errors in peak flows. Therefore, often a transformation is applied on the hydrograph with the goal to emphasize the assessment of low flows. Using the *NSE* criterion, a log-transformation of the flow is the most applied transformation. Using the *KGE* criterion however, a log-transformation of the flow results in problems with zero flow values and numerical instability issues when flow values average around 1. The *KGE* of log-transformed flows does also not remain dimensionless (Santos et al., 2018). Zero flow values are common in some of the streamflow timeseries used in this study and therefore this issue is also important for this project.

A solution to these problems is to use the modified Box-Cox transformation of the flows (Santos et al., 2018), instead of a log-transformation. However, numerical instability issues are still a problem around specific average flow values. The modified Box-Cox transformation does increase the weight of low flows, but does so less than a logarithmic transformation. Still, the increase in weight in low flows is more than other alternatives that tackle the problem of zero flow values, such as a square-root transformation. The Box-Cox transformed flow values $f'_{BC}(Q)$ are given by Equation 6, with the Box-Cox parameter λ being equal to 0.25 (Vázquez et al., 2008). The *KGE* of the Box-Cox transformed flow values is used as the second hydrological signature.

$$f'_{BC}(Q) = \frac{Q^{\lambda} - (0.01\mu_{obs})^{\lambda}}{\lambda} \tag{6}$$

Flow duration curve

The flow duration curve (*FDC*) is a hydrological signature in which the streamflow values are ordered from high to low and plotted on the y-axis against their exceedance probability on the x-axis. The steepness of the *FDC* shows how variable the flow regime is. By comparing the *FDC* of the observed and the simulated hydrographs, the distribution of flow values is assessed. Usually, the log-transformed flow values are used for this signature to not overemphasize the errors in high flow values. In this study, the Box-Cox transformed flow timeseries are used with again the KGE as objective function.

Autocorrelation function
The fourth hydrological signature that is used for the streamflow assessment is the autocorrelation function (ACF). The ACF assesses the internal memory effect of the system. This means that it says something about the influence of a flow value at a specific timestep on flow values in future timesteps. The ACF is constructed by calculating the correlation between the hydrograph shifted from 1 to a specified maximal time lag in time with itself. In this study, a maximal time lag of 100 days was used. The steeper the slope of the ACF, the shorter the memory effect, and vice versa. The KGE is used again as objective function for assessing the similarity between the ACF of the observed and simulated timeseries.

Monthly runoff coefficients
The last hydrological signature used is the timeseries of monthly runoff coefficients in the wet season. Runoff coefficients are given by the total runoff in an area over the total precipitation input in that same area in a given amount of time. It tells the modeller the fraction of water that is leaving the system as streamflow and hence, what fraction is leaving the system as AET or is stored in the system. If the amount of water stored in the system is assumed to be low, the runoff coefficients not only help the modeller to close the water balance, but via the fraction of evaporation over precipitation also help to close the energy balance, and thereby to secure long-term conservation of energy (Hrachowitz & Clark, 2017). The KGE is used again for the assessment of the monthly runoff coefficients of observed and simulated timeseries.

Combing all hydrological signatures and stations
A calibration algorithm can only optimize for one objective function at a time. Therefore, the $KGE's$ of the hydrological signatures need to be combined into one objective function. Ideally, this would be done by taking the ED of the 5 objective functions described in this section, but for consistency reasons explained further in subsubsection 3.3.4, it was chosen to use a weighted average of the objective functions.

The KGE of the original hydrograph, KGE_{OH}, is deemed the most important objective function, and is prioritized over the other objective functions by using a higher weight. The KGE_{OH} is given a weight of 0.32 and the KGE of the Box-Cox transformed flows, KGE_{BC}, the KGE of the flow duration curves, KGE_{FDC}, the KGE of the autocorrelation function, KGE_{ACF}, and the KGE of the monthly runoff coefficient timeseries, KGE_{R_c}, are given a weight of 0.17. This was was done because especially KGE_{ACF} and KGE_{R_c} are based on much less data points than the other hydrological signatures. In this way, the sum of the weights of all objective functions equals 1. See Equation 7 for how the weighted average KGE of a station WA_{KGE_s} is calculated.

$$WA_{KGE_s} = 0.32KGE_{OH} + 0.17(KGE_{BC} + KGE_{FDC} + KGE_{ACF} + KGE_{R_c}) \tag{7}$$

The weighted average of the objective functions is calculated for each discharge station separately in the calibration and evaluation period. The performance score that is passed on to the optimization is the arithmetic mean of the performance scores (WA_{KGE_s}) of the calibration stations in the calibration period (KGE_Q, See Table 9). The arithmetic mean of these scores of the calibration stations in the evaluation period is used for temporal evaluation, while the arithmetic mean of the scores of the evaluation catchments in the calibration period is used for the spatial evaluation. The arithmetic mean of the streamflow performance scores of the evaluation stations in the evaluation period is used for the spatio-temporal evaluation.

3.3.3 TWSA, AET and SM assessment

The spatial pattern efficiency metric E_{SP}

The performance of the model to simulate the spatial pattern of the TWSA, AET and SM as observed by RS data is tested using the spatial pattern efficiency metric E_{SP} (Dembele, Hrachowitz, et al., 2020), see Equation 8. This metric has multiple components and has the same structure as the KGE (Gupta et al., 2009; Kling et al., 2012). The E_{SP} is defined as 1 minus the ED of a correlation term r_S, a variability term γ, and spatial location matching term α. For the correlation term r_S, the Spearman rank-order correlation coefficient is used (Equation 9), in which d is the distance in rank between the modelled and the observed variable. γ is again the variability ratio, which is the same as is defined in Equation 5. And the spatial location matching term α (Equation 10) is defined as 1 minus the root-mean-squared error (E_{RMS}) of the Z-scores of the observed and simulated variable (Z_{obs} and Z_{sim}, respectively). The Z-score is a standardized value that allows for comparison between different distributions. A Z-score tells you how many standard deviations a value is from the mean of the values (Dembele, Hrachowitz, et al., 2020).

The E_{SP} does not look at absolute values of a variable and is not sensitive to the possible bias present in the RS data (there is no bias term). The metric concentrates on the spatial pattern similarity and is therefore an excellent metric to compare simulation results with RS observations, because the spatio-temporal pattern is a strength of RS observations, while the accuracy of the absolute values is not. The optimal value of all components of the E_{SP} and of the E_{SP} itself is equal to 1, while the lowest value of E_{SP} is equal to $-\infty$, just like for the KGE. Applying the method of Knoben et al. (2019) to the E_{SP} results in a score of -0.73 for the mean of the observations as simulation.

$$E_{SP} = 1 - \sqrt{(r_S - 1)^2 + (\gamma - 1)^2 + (\alpha - 1)^2} \tag{8}$$

$$r_S = 1 - \frac{6\sum_1^n d^2}{n(n^2-1)} \tag{9}$$

$$\alpha = 1 - E_{RMS}(Z_{sim}, Z_{obs}) \tag{10}$$

The spatial pattern efficiency metric can also be applied on timeseries to assess the temporal pattern of a simulation. All three terms of the E_{SP} aggregate the gridded information in the spatial application of the metric to single values (variability term and E_{RMS} of Z-scores) or 1D ranks (spearman rank correlation term and Z-scores). Therefore, exactly the same formula (Equation 8) can also be applied on timeseries. The correlation term r_S now ranks observations and simulations of timeseries, the variability term γ was already applied to timeseries in Equation 5 and the spatial (but now temporal) location matching term α assesses the E_{RMS} of the position of the data at each time step in terms of the number of standard deviations from the mean. Hence, simulated values will be forced to take a similar place in the distribution as the observations.

This temporal application of E_{SP} enables the modeller to compare and assess the performance of strongly correlated but not exactly similar simulations and observations (like AET and NDVI), not only spatially but also temporally and has the same advantages as in the spatial application, related to using pattern information instead of absolute values of RS observations. The only downside is that you lose the information that is hidden in the absolute values of the RS observations, but these absolute values are also very uncertain and can therefore not be used one-to-one. To discriminate between the application of the spatial pattern efficiency metric to spatial or temporal patterns, the application to spatial patterns will be called E_{SP}, while the application to temporal patterns will be called E_{TMP} in the rest of this study.

TWSA assessment

The performance of the model to simulate the TWSA can be assessed spatially and temporally. Spatially, this is done using the E_{SP}. This metric assesses the performance of the model to simulate the spatial pattern of TWSA values, on each day in the TWSA calibration and evaluation period. The TWSA calibration period is shorter than the calibration period of the other calibration datasets because observations from GRACE are only available from April 2002. In the first 2 years of observations, also 3 months of data are missing. Therefore, the calibration period for GRACE is defined from April 2002 up to and including February 2005 (32 months with data), and the evaluation period is defined from March 2005 till March 2007 (24 months with data).

The E_{SP} is calculated for both the calibration catchments and the evaluation catchments separately. Since GRACE TWSA observations are spatially very coarse (1° resolution), observations were defined as inside the basin if more than half of the cells cover the basin and outside the basin if less than half of the cells cover the basin. This resulted in a total of 17 cells in the calibration catchments and only 6 cells in the evaluation catchments. The E_{SP} score for the evaluation catchments should be looked at with caution, because the calculation of this score involves taking the mean and the standard deviation of 6 samples, which is generally considered too low to result in a representative value. The simulated TWSA is upscaled to the GRACE TWSA resolution to allow for the calculation of the E_{SP}.

Since every month in the simulation period results in a new E_{SP} value, the temporal mean of the E_{SP} in the calibration period and the calibration catchment is the metric that is added to the calibration procedure and which is optimized using DDS. The temporal mean of the E_{SP} in the calibration catchment and the evaluation period, and the temporal means of the E_{SP} in the evaluation catchment in the calibration and evaluation periods, are used for temporal, spatial and spatio-temporal evaluation respectively.

Not only the spatial pattern of TWSA can be assessed, but also the temporal pattern. The GRACE TWSA observations show a very clear seasonal signal, with rising TWSA values in the wet season, and descending TWSA values in the dry season (See Figure 27 in Appendix C). As is explained earlier in this section, E_{TMP} is used for the assessment of the temporal pattern. Again, upscaling of the modelled data is needed to allow for a fair comparison between simulations and observations. The spatial mean of the $E_{TMP}'s$, based on the calibration period in the calibration catchments is added to the calibration procedure, and is the value which is optimized using DDS. The spatial mean of the $E_{TMP}'s$ in the calibration catchment in the evaluation period, and the spatial means of the $E_{TMP}'s$ in the evaluation catchment in the calibration and evaluation periods are used for temporal, spatial and spatio-temporal evaluation respectively.

AET assessment

As was explained in subsubsection 2.1.3, the simulated AET is assessed using RS observations of NDVI. This assessment is justified because of the strong positive correlation between NDVI and AET (Wang et al., 2007; Cihlar et al., 1991; Seevers & Ottoman, 1994; Islam & Mamun, 2015; Kerr et al., 1989; Szilagyi et al., 1998). Some of these publications suggest even stronger positive correlations between NDVI and PET or cumulative NDVI and cumulative AET (Cihlar et al., 1991; Chen et al., 2019), or between NDVI and a time-lagged AET (Kerr et al., 1989; Szilagyi et al., 1998). Szilagyi et al. (1998) also argues that NDVI is especially a useful indicator for AET in a water limited environment.

Since the purpose of this research is not to find the best relation between NDVI and AET, the filtered (as explained in subsubsection 2.1.3) NDVI observations are directly used in this research for assessment of the AET flux. This simulated AET flux consists of two components, namely interception evaporation and evaporation from the soil moisture reservoir, which represents both soil evaporation and transpiration. NDVI observations are available at the spatial and temporal model resolution, so simulations can directly be compared to the observations. Again, the division in calibration and evaluation catchments and periods is used.

Spatially, the comparison between observed NDVI and AET can easily be made because the E_{SP} metric is insensitive to the unit and the absolute value of the variable. Only the spatial pattern of NDVI and AET is compared and assessed on similarity. The underlying assumption that is made here is that higher NDVI values indicate higher AET. This assumption is justified by the results of the publications mentioned above. The temporal mean of the $E_{SP}'s$ in the calibration period and calibration catchments is added to the calibration procedure, while the temporal mean of the $E_{SP}'s$ in the evaluation period and the calibration catchment, and the temporal means of the $E_{SP}'s$ in the calibration and evaluation periods in the evaluation catchment are used for temporal, spatial and spatio-temporal evaluation respectively.

Also temporally the comparison between the observed NDVI and simulated AET can easily be made because of the use of the E_{TMP} metric. The spatial mean E_{TMP} of the calibration or evaluation catchments in the calibration or evaluation period forms 1 of the 4 final temporal performance scores for this dataset. The spatial mean E_{TMP} of the calibration catchment in the calibration period is added to the calibration procedure and is the value that is optimized for using DDS. The spatial mean E_{TMP} of the calibration catchments in the evaluation period and the spatial mean $E_{TMP}'s$ of the evaluation catchments in the calibration and evaluation periods are used for temporal, spatial and spatio-temporal evaluation respectively.

SM assessment

The assessment of the water present in the soil moisture reservoir against the RS soil moisture observations is the least trivial assessment in this study. This is because RS observations of soil moisture are only representative for the top few centimeters of the unsaturated zone, while the soil moisture reservoir is assumed to represent the complete unsaturated zone, which is generally several hundreds of millimeters deep. The RS surface soil moisture observations are indicated with SSM, while the simulated amount of water in the soil moisture reservoir is indicated with SM in the rest of this study.

However, the amount of water in the topsoil is strongly related to the amount of water in the complete unsaturated zone, via the Soil Water Index (SWI). The SWI can be estimated using the characteristic time length T, which is a measure for how fast water sinks from the topsoil to the deeper parts of the unsaturated zone. Bouaziz et al. (2020) developed a method to estimate T based on relatively simple water balance statistics. The method assumes that "optimal T values are linked to ... catchment-scale accessible water storage capacities in the unsaturated zone" (p. 1) Bouaziz et al. (2020). More specifically, the depth of the unsaturated zone ($S_{u,max}$) is estimated to be equal to the 1-in-20-year occurring soil moisture deficit, which can be found using Gumbel extrapolation of soil moisture deficit timeseries. The optimal T value can be found by identifying the highest Spearman-rank order correlation coefficient between the simulated soil moisture timeseries per catchment using $S_{u,max}$, and the timeseries of the SWI based on the RS observations of soil moisture for a T-value ranging from from 1 till 100 days. The method is explained in more detail in Bouaziz et al. (2020) and is applied for each combination of subcatchments and land use classes in this model. This classification was chosen because especially the land use is important in the estimation of the depth of the unsaturated zone. This is because the interception capacity ICF is the main parameter that determines how much water ends up in the soil routine. Since the discharge is known for each subcatchment (subcatchments are based on gauging stations), the combination of land use and subcatchment classes results in the highest resolution possible for the estimation of T, and finally the depth of the unsaturated zone.

The method of Bouaziz et al. (2020) thus not only estimates the characteristic time length T, which makes a comparison between the simulated and observed soil moisture possible, but also provides an estimation of the depth of the unsaturated zone $S_{u,max}$. In Bouaziz et al. (2020) it is argued that $S_{u,max}$ is only minimally dependant on the interception capacity ICF. However, in Appendix D it is shown that for the Volta basin this dependence is important and should be taken into account. The relation between ICF and $S_{u,max}$ is almost perfectly linear (Appendix D) and therefore a first order relation is used per sub-catchment and land use class in the calibration procedure to estimate $S_{u,max}$ directly from ICF. In this way, calibration of ICF and $S_{u,max}$ reduces to calibration of only ICF, which directly results into 26 realistic values of $S_{u,max}$. The model then runs with these ICF and $S_{u,max}$ values, after which the optimal T values per sub-catchment and land use class are determined. The SWI timeseries per model cell based on RS SSM observations corresponding to those specific T-values are then extracted from a predefined SWI-file containing for each cell in the model the SWI timeseries for 100 characteristic time lengths between 1 and 100 days.

In the final assessment, the simulated water level in the soil moisture reservoir SM is compared to the 'observed' SWI, derived using the method explained in this section. The SWI timeseries based on RS SSM observations are available at a resolution of 0.25°, which is 5 times coarser than the model resolution. Therefore, the simulated SM timeseries are upscaled to the resolution of the observations before the E_{SP} assessment function is applied. Again, the temporal mean of the $E_{SP}'s$ in the calibration period and calibration catchment is added to the calibration procedure, and is the value for which is optimized using DDS. The temporal mean of the $E_{SP}'s$ in the evaluation period and the calibration catchment, and the temporal means of the $E_{SP}'s$ in the calibration and evaluation periods in the evaluation catchment are used for temporal, spatial and spatio-temporal evaluation.

The temporal assessment is again performed using the E_{TMP} function, again on the simulated water level in the soil moisture reservoir compared to the SWI derived using the method explained in this section. The E_{TMP} is calculated for every cell, for the calibration and evaluation period separately. The spatial mean E_{TMP} of the calibration period in the calibration catchments is added to the calibration procedure and is the value which is optimized using DDS. The spatial mean E_{TMP} of the evaluation period in the calibration catchments, and the spatial mean $E_{TMP}'s$ of the evaluation catchments in the calibration and evaluation period are used for temporal, spatial and spatio-temporal evaluation respectively.

3.3.4 Overall model assessment

The overall model assessment is the same for every scenario. The difference between the scenarios is the combination of datasets which is optimized for. In the first scenario (Q-only), the model is optimized for streamflow performance, but is finally assessed on all 4 calibration and evaluation datasets. In the following scenarios, the model is optimized for the performance on both streamflow and one or more RS datasets, but is again finally assessed on all calibration and evaluation datasets. In the fifth scenario the optimization is performed for a set of RS calibration and evaluation datasets only, but the final assessment is again performed based on both RS data and streamflow observations.

In order to optimize for multiple objective functions, the objective function for streamflow, denoted as KGE_Q, and the objective functions for the spatial and temporal patterns of TWSA, AET and SM, denoted as $E_{SP,TWSA}$, $E_{TMP,TWSA}$, $E_{SP,AET}$, $E_{TMP,AET}$, $E_{SP,SM}$ and $E_{TMP,SM}$, need to be combined into 1 single optimization function per scenario. Since only temporal and spatial mean values of these objective functions are used for spatial and temporal aggregation, and all these objective functions are built-up in a very similar way, the final performance scores are very comparable and can be put together into 1 single optimization function using the Euclidean Distance (ED). The resulting optimization functions that were used in this study are given in Equation 11 till Equation 15. Because the RS calibration and evaluation data functions all have a temporal and spatial component, while the streamflow only has a temporal component, this temporal component is counted double in the optimization functions, so that each dataset used in the optimization has the same weight.

$$\begin{aligned}
ED_Q =& KGE_Q - 1 \quad (11)\\
ED_{Q+SM} =& \sqrt{2(KGE_Q - 1)^2 + (E_{SP,SM} - 1)^2 + (E_{TMP,SM} - 1)^2} \quad (12)\\
ED_{Q+TWSA} =& \sqrt{2(KGE_Q - 1)^2 + (E_{SP,TWSA} - 1)^2 + (E_{TMP,TWSA} - 1)^2 +} \quad (13)\\
ED_{Q+SM+TWSA} =& \sqrt{2(KGE_Q - 1)^2 + (E_{SP,SM} - 1)^2 + (E_{TMP,SM} - 1)^2 +}\\
& \overline{(\mathrm{E}_{SP,TWSA} - 1)^2 + (E_{TMP,TWSA} - 1)^2} \quad (14)\\
ED_{SM+TWSA} =& \sqrt{(E_{SP,SM} - 1)^2 + (E_{TMP,SM} - 1)^2 + (E_{SP,TWSA} - 1)^2 +}\\
& \overline{(\mathrm{E}_{TMP,TWSA} - 1)^2} \quad (15)
\end{aligned}$$

The big advantage of using the ED compared to using the mean of a set of objective functions is that the ED is much better able to optimize for all components of a function compared to the mean. The mean option is more likely to optimize for only several components of the optimization function, which then compensate for the other lower performance scores. In the ED calculation, low performance scores of individual components will also result in a lower final optimization function. Thus by using the ED of a set of components as optimization function, the modeller forces the calibration algorithm to optimize for all components together.

However, this can only be done if the individual components of the optimization function are comparable functions. Therefore, it was chosen to use the mean of the KGE of all calibration stations in the calibration period for streamflow, which were based on the (non-arithmetic) mean of the 5 objective functions for streamflow, and the temporal mean E_{SP} and the spatial mean E_{TMP} (within the calibration catchments and calibration period) of the RS calibration and evaluation data, so that the final performance scores of each assessed variable are comparable.

Nevertheless, also this approach has its shortcomings. The first flaw is that although the KGE and E_{SP} (and thus E_{TMP}) are built-up the same way, they are not equivalent. The KGE may be a stricter objective function than the E_{SP}, because the KGE assesses absolute values, while the E_{SP} only assesses spatial (and temporal) patterns via relative values, without for instance using a bias term. This feature of the E_{SP} has multiple advantages, especially for comparison of simulated data with RS observations, as is explained in subsubsection 3.3.3, but it also makes the 2 objective functions used in this study less comparable. However, it is deemed necessary to assess streamflow performance using absolute values, and therefore the here described inconsistency is accepted.

A second shortcoming is that, although the final objective function values for each dataset are very comparable and can therefore be put together in an optimization function using ED, mean components of the final objective function of each dataset are used as input for those function values. The mean of the 5 hydrological signatures is used, and the mean of the stations in the calibration catchments, which allows the model here to compensate a low performance score of one station with a high performance score of another station, or a low performance score for one hydrological signature, with a high performance score for another.

4 Results

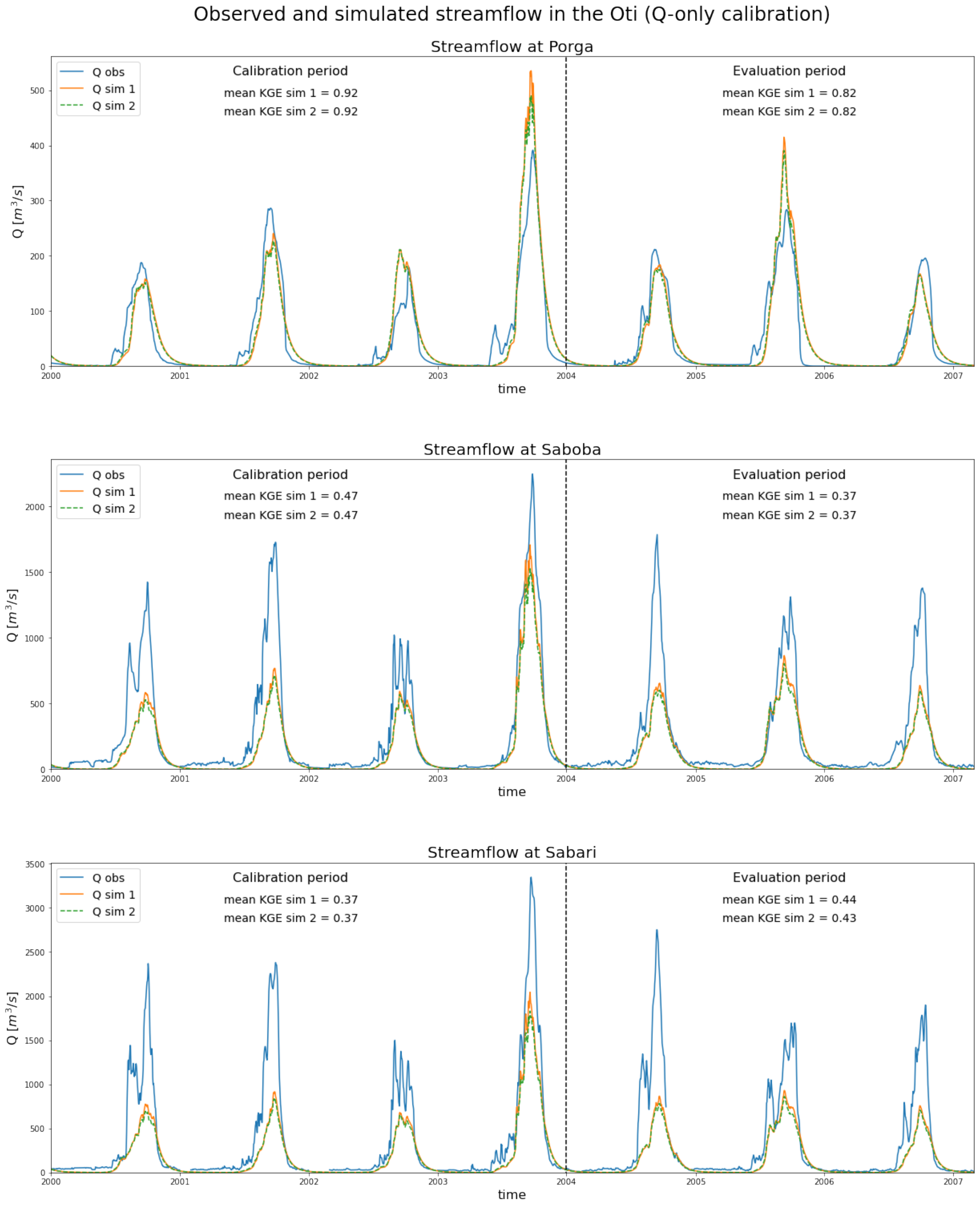
In this chapter the most important results of the model simulations are discussed. 2 calibration runs of 200 model runs were done for every scenario. Only key-figures, important to understand the described results are shown here. All other results can be found in Appendix G till Appendix L. In subsection 4.1, the model results after calibration on Q-only are discussed. Based on these results, it was chosen to leave streamflow data of several stations and NDVI data out of the other scenarios and to add SM data to the calibration in the next scenario. These results are discussed in subsection 4.2. This choice is further elaborated on in section 5. It is also explained here why the calculation of the temporal and spatial pattern performance scores was changed for the TWSA dataset. In subsection 4.3, TWSA observations are added to the calibration on Q-only and in subsection 4.4 the results of adding both RS datasets to the calibration are discussed. Finally, an attempt is made to model streamflow based on only SM and TWSA observations, of which the results are discussed in subsection 4.5. An overview with comparisons of the results of different scenarios is given in subsection 4.6

4.1 Results Scenario 1: Q-only Calibration

In this subsection the results of the calibration on Q only are discussed. These results will be used as a baseline scenario to compare all other scenarios with. The hydrographs for the stations in the Black Volta, White Volta and Oti are shown in Figure 37, Figure 38 in Appendix G and Figure 7, respectively. In Figure 8 (top), the hydrographs for the station Chache in the Black Volta are also shown, together with the timeseries of the spatial mean of the observed and simulated TWSA in the calibration catchments (second from above), the timeseries of the spatial mean of the observed NDVI and simulated AET in the calibration catchments (third from above) and the timeseries of the spatial mean of the surface soil moisture (SSM), thereof derived SWI for both simulations, and the modelled amount of water in the soil moisture reservoir (SM) in the calibration catchments (bottom). In Figure 9, spatial plots of the same observations and simulated model states and fluxes are shown (except Q), temporally averaged over the calibration period. The same 2 figures showing the spatial mean timeseries in the evaluation catchments and the temporal mean spatial plots in the evaluation period are given in Figure 39 and Figure 40 in Appendix G, respectively. The performance scores for each gauging station, the performance scores for each RS dataset in the calibration and evaluation periods and catchments, the determined T-values and the parameter set found by DDS per simulation can also be found in Appendix G.

Streamflow performance

The performance of the model calibrated on Q only is expressed using the mean of the weighted mean KGE values of the 5 streamflow performance functions as described in subsubsection 3.3.2. This value is 0.64 for the calibration stations in the calibration period for both simulation 1 and 2. This score goes down to 0.61/0.60 for the calibration stations in the evaluation period, and to 0.36/0.37 for the evaluation stations in the calibration period (See Table 13 and Table 14 in Appendix G). The streamflow performance scores are also given per station in the hydrographs shown in Figure 37, Figure 38 and in Figure 7 and for Chache, the hydrographs and the corresponding performance scores are als shown in Figure 8.



Observed and simulated streamflow in the Oti (Q-only calibration)

Streamflow at Porga

Streamflow at Saboba

Streamflow at Sabari

Figure 7: Hydrographs of the stations in the Oti based on calibration on Q only. These stations are calibration stations. The mean KGE values given in the plots are mean values of the 5 objective functions used for calibration as was explained in subsubsection 3.3.2.

In most of the simulated hydrographs the general flow regime is simulated quite accurately, resulting in reasonable performance score values. However, often a delay is observed in the timing of the start and end of the yearly flow peak. The streamflow observations generally also have sharper peaks compared to the simulations, although this peaky flow regime is simulated better in the Oti catchments. Another observation that can be made from the hydrographs is that the simulated yearly flow peaks are a bit higher or about the same value in the most upstream subcatchments of the three main branches of the Volta. See for example the stations Lawra in the Black Volta and Porga in the Oti. However, in the more downstream subcatchments, the yearly observed flow peaks are not always reached by the simulations. See for example the stations Bamboi in the Black Volta (note that this timeseries is edited, See section 5) and Sabari in the Oti. For the stations Nawuni and Daboya in the White Volta, this observed pattern is the other way around, with flow peaks more downstream in the Volta branch being better simulated.

A final observation that can be made with respect to the simulated streamflow timeseries is that the hydrographs of simulation 1 and 2 are very close to each other, which indicates that the solution found by DDS after 200 model runs is quite stable. This is also confirmed by the very comparable performance scores. However, some parameter values found by DDS are quite different (See Table 17). A complete overview of the performance scores of all streamflow assessment functions for both simulations can be found in Table 13 and Table 14 in Appendix G.

TWSA performance

In Figure 8 (second from above), timeseries of the observed and simulated spatial mean TWSA in the calibration catchments are plotted with the corresponding (spatial mean) E_{TMP} values for both simulations. It can be observed that the general pattern and the amplitude of the TWSA is very well simulated by the model, but that there are again some issues with the timing of the simulation. Generally, the increase in TWS during the wet season is simulated earlier than it is observed. However, the timing and value of the most extreme positive TWSA is very accurate. The decrease in TWS is then again simulated earlier then it is observed, and this shift in time gets larger during the dry season, with the largest temporal shift at the most extreme negative TWSA of the year. The resulting performance scores given in the plot are also spatial mean values of the calibration catchments. The scores indicate reasonable simulations, with E_{TMP} values of 0.39 and 0.35 in the calibration period, and E_{TMP} values of 0.33 and 0.29 in the evaluation period, for simulation 1 and 2 respectively. Both simulations correspond very well. A similar plot with the results in the evaluation catchments is given in Figure 39 in Appendix G. These results are very comparable to the results for the calibration catchments.

In Figure 9 (top row), spatial plots of the observed and simulated temporal mean TWSA in the calibration period are shown, including the (temporal mean) E_{SP} values for both simulations. It can be observed that both the observations and the simulations are close to zero (the amplitude in the colorbar is relatively small) but that there are some minor differences between the observations and simulations. In the observed spatial pattern, the western part of the catchments is relatively wet, while this is the south-eastern part in the simulated spatial patterns. Both simulated spatial patterns are very similar, indicating a stable solution. The (temporal mean) E_{SP} values are lower than the (spatial mean) E_{TMP} values and even lower than using the mean of the observations as simulation. This is because the ratio of these relatively small standard deviation values (See subsection 5.1) still results in large values in the TWSA assessment calculation. In Figure 40 (top row) in Appendix G, the same plots for the evaluation period are shown. The spatial patterns observed here are inverses of the spatial patterns in Figure 9, so that the complete temporal mean is always equal to 0. The (temporal mean) E_{SP} values are comparable to the values in the calibration period.

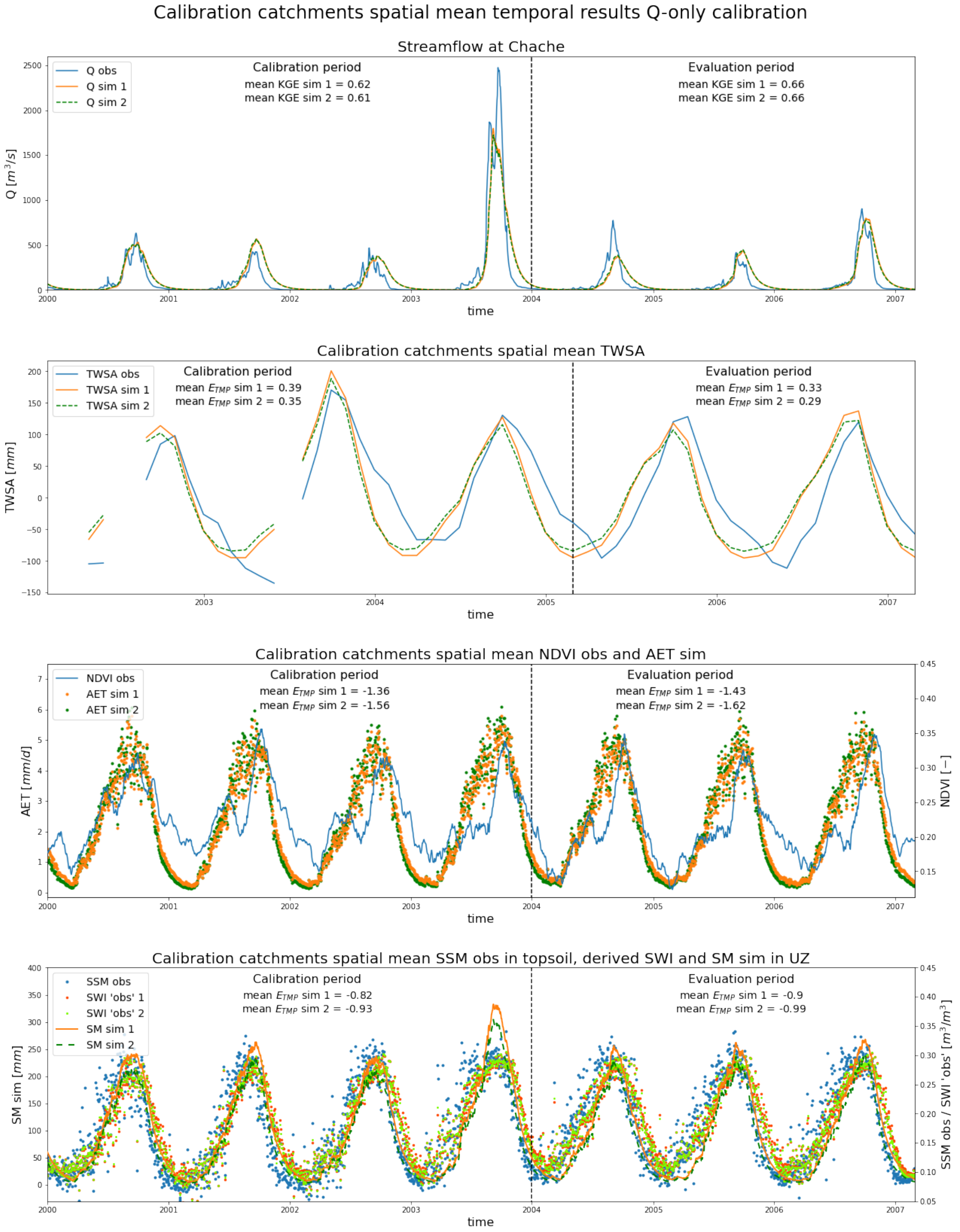


Figure 8: Timeseries of streamflow observations and simulations at Chache (top). Timeseries of the mean TWSA observations and simulations within the calibration catchments (second from above). Timeseries of the mean NDVI observations and mean AET simulations within the calibration catchments (third from above). Timeseries of the mean surface soil moisture observations and the mean simulated amount of water in the soil moisture reservoir (SM) within the calibration catchment (bottom). The same results for the evaluation catchments can be found in Appendix G.

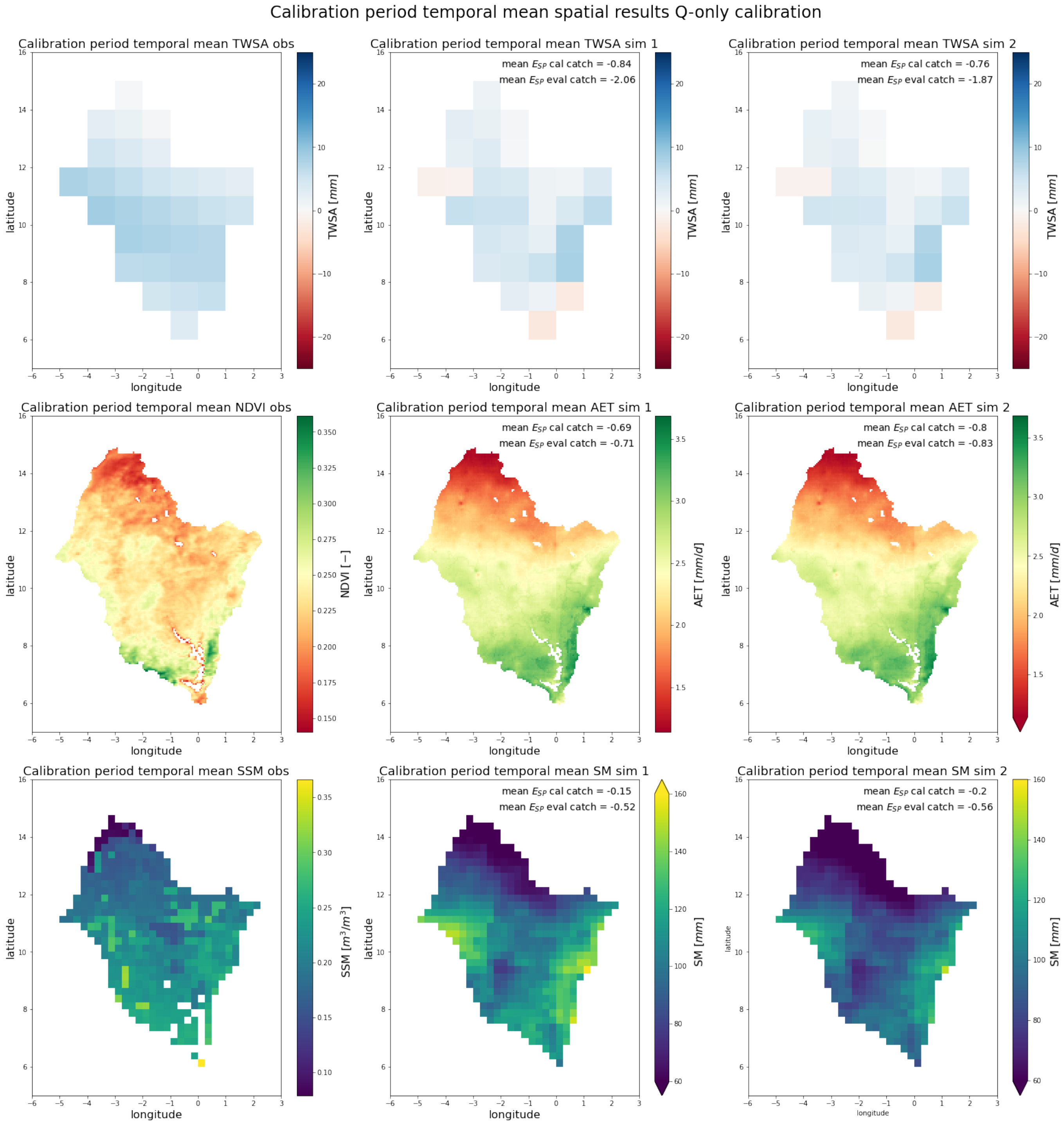

Calibration period temporal mean spatial results Q-only calibration
Calibration period temporal mean TWSA obs
Calibration period temporal mean TWSA sim 1
mean E_{SP} cal catch = -0.84
mean E_{SP} eval catch = -2.06
Calibration period temporal mean TWSA sim 2
mean E_{SP} cal catch = -0.76
mean E_{SP} eval catch = -1.87
TWSA $[mm]$
Calibration period temporal mean NDVI obs
NDVI $[-]$
Calibration period temporal mean AET sim 1
mean E_{SP} cal catch = -0.69
mean E_{SP} eval catch = -0.71
Calibration period temporal mean AET sim 2
mean E_{SP} cal catch = -0.8
mean E_{SP} eval catch = -0.83
AET $[mm/d]$
Calibration period temporal mean SSM obs
SSM $[m^3/m^3]$
Calibration period temporal mean SM sim 1
mean E_{SP} cal catch = -0.15
mean E_{SP} eval catch = -0.52
Calibration period temporal mean SM sim 2
mean E_{SP} cal catch = -0.2
mean E_{SP} eval catch = -0.56
SM $[mm]$
latitude
longitude

Figure 9: Spatial plots of the mean TWSA observations and simulations in the calibration period in the Volta basin (top). Spatial plots of the mean NDVI observations and AET simulations in the calibration period in the Volta basin (middle). Spatial plots of the mean surface soil moisture observations and the mean simulated amount of water in the soil moisture reservoir in the calibration period in the Volta basin (bottom). The same results for the evaluation period can be found in Appendix G.

NDVI vs. AET performance

The results of the comparison between timeseries of spatial mean NDVI observations and AET simulations in the calibration catchments and the spatial plots of temporal mean NDVI observations and AET simulations in the calibration period are shown again in Figure 8 (third from above) and Figure 9 (second from above). The same plots for the evaluation catchments and evaluation period are shown in Figure 39 and Figure 40 in Appendix G.

The spatial mean timeseries of NDVI observations and AET simulations in the calibration and evaluation catchments both confirm the relation between NDVI and AET, discussed in subsubsection 2.1.3 and subsubsection 3.3.3. High NDVI values coincide with high AET values, and low NDVI values coincide with low AET values. However, the path in between the yearly extremes of both timeseries differ a lot. In the NDVI observations, two peaks per year can be distinguished, of which the first one is lower than the second one. In the AET simulations, this first peak cannot be distinguished. Both simulations present again very comparable results, with a slightly lower performance score for the second simulation. In the evaluation catchment, results are very comparable.

The temporal mean spatial plots of NDVI observations and AET simulations in the calibration and evaluation periods also show a clear qualitative relation, because both have a strong north-south gradient, but especially in the middle part of the basin, NDVI observations and AET simulation show a different pattern. The mean spatial pattern looks worse than the mean temporal pattern but the E_{SP}-values shown in Figure 9 (middle row) are higher than the E_{TMP}-values in Figure 8, and the same holds for the evaluation period. Based on these results and the argumentation given in Chapter 5, it was chosen to leave NDVI data out of the calibration scenarios in this study.

SM performance

The results of the comparison between surface soil moisture observations (SSM) and the amount of water in the unsaturated zone (SM) in the calibration catchments and the temporal mean SSM observations and SM simulations in the calibration period are shown again in Figure 8 (bottom) and Figure 9 (bottom) for the SSM observations and both SM simulations. Also the SWI, derived from the SSM observations using the water balance method described in subsubsection 3.3.3 and Appendix D, is plotted in Figure 8, because the SWI is the 'observation' which is compared to the simulated amount of soil moisture in the soil moisture reservoir. The same plots for the evaluation catchments and evaluation period are shown in Figure 39 and Figure 40 in Appendix G.

The timeseries in Figure 8 show that the mean temporal pattern is very well simulated. The timing of the yearly extremes is very similar for the SWI 'observations' and the the SM simulations, and also the timing of in- and decreasing values over the year is accurate. However, a time-lag is observed in both the wet and the dry season. In the wet season, the observed soil moisture values generally increase earlier than the simulation, while in the dry season, the observations decrease later than the simulation. The good mean temporal pattern performance is not directly reflected in high mean E_{TMP} values. The results show a mean E_{TMP} of -0.82 for simulation 1 and -0.93 for simulation 2 in the calibration catchments and calibration period. In the evaluation period this score decreases to -0.9 and -0.99 for simulation 1 and 2, respectively. This is due to low performance scores for the variability ratio component in the E_{TMP} calculation. Results and mean performance scores in the evaluation catchments are comparable.

The spatial plots in Figure 9 show that the simulated spatial patterns of soil moisture slightly diverge from the observed one. The lowest soil moisture values are are found in the north of the basin for both the observations and the simulations, but for the observations this is in the north-west and for the simulation in the north-east corner of the basin. The higher values also match in direction in the basin, but not in the exact locations. The soil moisture simulations are generally much smoother than the SSM observations. This difference will already be smaller when one looks at the SWI ‘observations’. The temporal mean E_{SP} values in the calibration and evaluation catchments are not that bad. Simulation 1 is slightly better than simulation 2 with E_{SP} scores of -0.15 and -0.20 in the calibration catchments, and -0.52 and -0.56 in the evaluation catchments, for simulation 1 and 2 respectively. These performance score values are higher than the values found for the temporal pattern, although the temporal pattern match looks much better. The results in the evaluation period shown in Figure 40 in Appendix G are very comparable. The T-values corresponding to simulation 1 and 2 can also be found in Appendix G in Figure 41.

4.2 Results scenario 2: Q + SM Calibration

The results of the calibration on Q and SM observations are discussed in this subsection. Both datasets have had an equal weight in the optimization function and optimization for SM observations was done w.r.t. both the temporal and the spatial pattern at the same time, also with an equal weight for both patterns (as was explained in subsubsection 3.3.4). The figures that present the results of this calibration scenario are shown in this subsection or in Appendix H. Observed and simulated hydrographs of the Black Volta, White Volta and Oti catchments are given in Figure 42, Figure 43 and Figure 44, respectively. The performance score of every separate streamflow objective function for every separate station can be found in Table 18 and Table 19, and the performance scores for the RS datasets in Table 20 and Table 21. The results of the temporal pattern assessment are presented in Figure 10 for the calibration catchments and Figure 45 for the evaluation catchments. Results of the spatial pattern assessment can be found in Figure 11 and Figure 46 for the calibration and evaluation period, respectively. The T-values and the parameters found for both simulations can be found in Figure 47 and Table 22.

Streamflow performance

The streamflow performance in the calibration catchments and period goes down to a mean KGE of 0.51 (-0.13 w.r.t. to Q-only) for both simulation 1 and 2. In the evaluation period, the streamflow performance goes down to a mean KGE of 0.48 and 0.49 (-0.13 / -0.14 w.r.t. Q-only) for simulation 1 and 2. The streamflow performance in the evaluation catchments is again much lower than in the calibration catchments. The mean KGE for both simulations found is 0.42 in the calibration period and 0.31 in the evaluation period. However, this is 0.05 higher compared to Q-only in the calibration period but 0.04 lower compared to Q-only in the evaluation period. The two simulations in this scenario are almost exactly the same for every hydrograph, which is also confirmed by the very similar performance scores.

Similar observations as in the calibration on Q-only scenario can be made from the hydrographs generated in this scenario. The simulated peaks are generally higher than the observed ones in the most upstream subcatchments, and this effect decreases or turns around in the downstream direction of the three main branches. This effect is especially observed in the Oti, but also in the Black Volta, but not so much in the White Volta, where this effect is reversed. Most simulated peaks are higher than the peaks in the Q-only simulation (See Figure 15 till Figure 17 in subsection 4.6 and Figure 69 till Figure 73 in Appendix L). The peaks are therefore also often much higher than the observed ones, which may be the largest cause of the reduction in streamflow performance. For the performance scores in the evaluation catchments in the calibration period however, the peaks in the hydrographs now often look more accurate than the ones simulated in the Q-only scenario, resulting in a higher performance score than in the Q-only scenario. The general streamflow pattern is very similar to the pattern simulated in the Q-only scenario, so the hydrographs look a bit like they are generated from a multiplication with a factor higher than 1 from the Q-only simulations. The overall streamflow performance per streamflow station may be lower, but for the stations Nawuni, Saboba and Sabari the streamflow performance score actually increases. However, the score decreases (often several tenths) in all other stations.

TWSA performance

The spatially averaged temporal pattern of the TWSA in the calibration catchment results in an E_{TMP} of 0.43 for both simulations in the calibration period, and 0.37 for both simulations in the evaluation period. This is an increase of 0.06 compared to the Q-only scenario in both periods. The E_{TMP} score in the evaluation catchments is 0.45 and 0.46 for simulation 1 and 2 in the calibration period, and 0.39 for both simulations in the evaluation period. This is an increase of 0.10 compared to the Q-only scenario in both periods. The two simulations are almost perfectly similar, just like

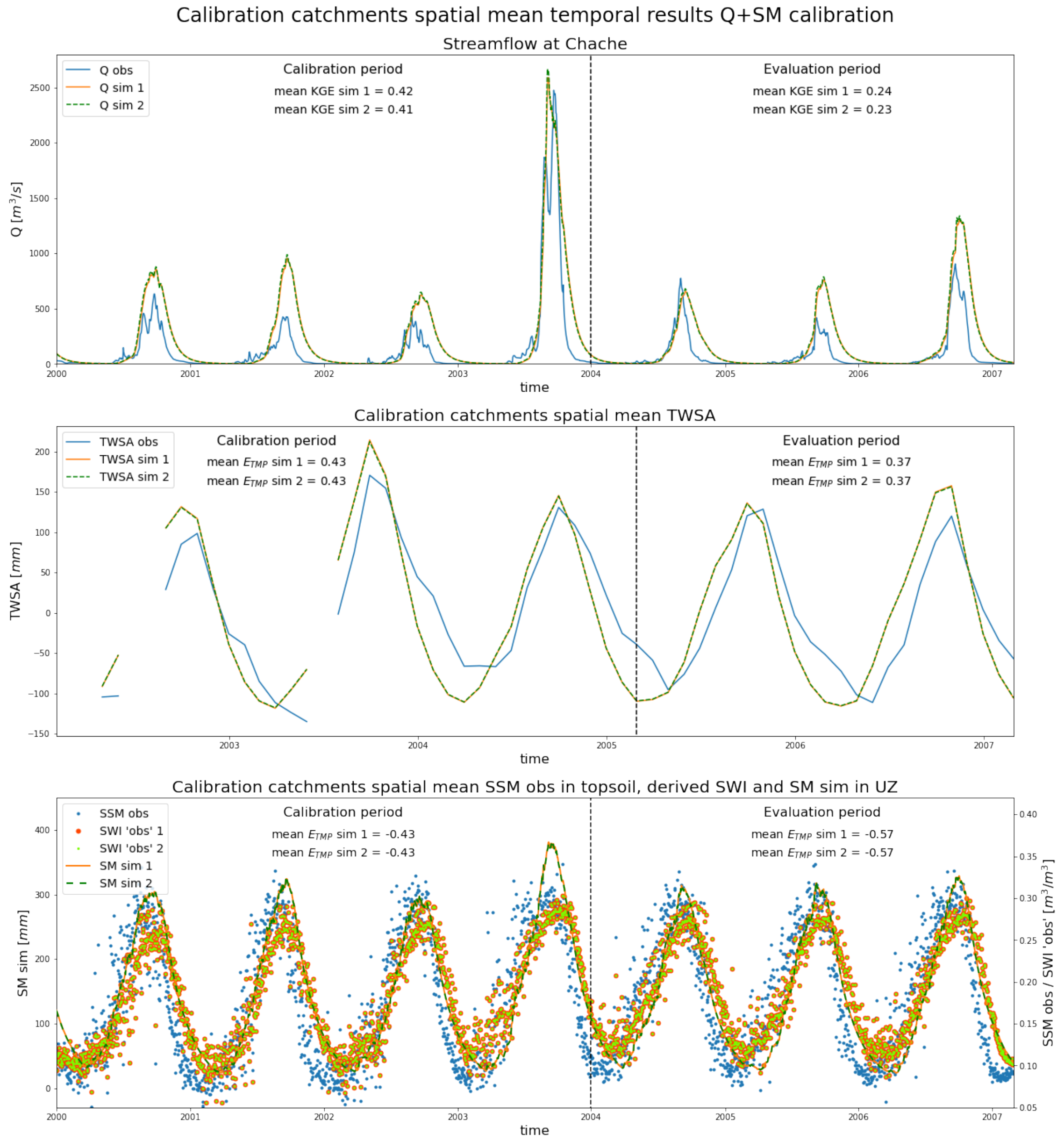


Figure 10: Timeseries of streamflow observations and simulations at Chache (top). Timeseries of the mean TWSA observations and simulations within the calibration catchments (middle). Timeseries of the mean surface soil moisture observations and the mean simulated amount of water in the soil moisture reservoir (SM) within the calibration catchments (bottom).

the streamflow simulations. Note that the E_{TMP} values are higher in the evaluation catchment than in the calibration catchment. As can be seen in Figure 18 in subsection 4.6 and Figure 74 in Appendix L, the amplitude of the TWSA temporal pattern increases compared to the Q-only calibration scenario, and the time difference between simulation and observations decreases a little, which increases the E_{TMP} performance score, in both the calibration and evaluation catchments.

Spatially, the performance values are again much lower than temporally. In the calibration period, the E_{SP} values are -0.95 and -0.94 in the calibration catchments for simulations 1 and 2, and in the evaluation period, these values go down to -1.09 and -1.07. In the evaluation catchments the opposite is observed. E_{SP} values are very low in the calibration period (-2.34 and -2.32 for simulation 1 and 2, respectively), and performance is a little better in the evaluation period (E_{SP} of -1.83 and -1.82 for simulation 1 and 2). In Figure 20 in subsection 4.6 and Figure 78 in Appendix L it can be seen that these performance values are a little lower compared to the Q-only calibration scenario, but that the simulated pattern is very similar.

SM performance

The performance of the soil moisture temporal pattern simulations looks very good but this cannot directly be seen back in the E_{TMP} scores. These are -0.43 for the calibration catchments in the calibration period, -0.57 for the calibration catchments in the evaluation period, -0.46 for the evaluation catchments in the calibration period, and then -0.65 for the evaluation catchments in the evaluation period. Both simulations are so similar that simulation 1 and 2 score exactly the same in all periods and catchments. It is observed that the spatial evaluation (transferring the parameter set to another catchment) is performing better than the temporal evaluation (transferring the parameter set to another period), however the lowest temporal pattern performance value is still found in the spatio-temporal evaluation.

The performance values in general are lower than the values observed for the temporal TWSA pattern because of large variability ratios all over the basin, but especially in the Oti catchments. Therefore these performance scores can also not be compared 1-to-1 with the TWSA performance scores because, as is explained in subsection 5.1, the TWSA variability term is only the ratio of the standard deviations of observations and simulations, while this is the ratio of the coefficients of variation of those two for SM. However, the E_{TMP} values are still much higher in all calibration catchments and periods than in the Q-only calibration scenario (order +0.4), as can be seen clearly in Figure 19 in subsection 4.6 and Figure 77 in Appendix L. The simulated amount of soil moisture in the soil is also higher than for the Q-only scenario, the SWI 'observations' are smoother and with a lower amplitude, and the T-values are much higher.

In the spatial plots of the temporal mean SM simulations it can clearly be seen that the simulated amount of water in the basin is much higher than in the Q-only calibration scenario. The general pattern is however still very much the same. The E_{SP} performance values are again very similar for both simulations, resulting in scores of -0.04 and -0.03 for the calibration catchments in the calibration period, 0.0 for the calibration catchments in the evaluation period, -0.4 for the evaluation catchments in the calibration period, and -0.32 for the evaluation catchments in the evaluation period (See Figure 11). This means that the performance scores in the evaluation catchments are actually higher than the performance scores in the calibration catchments. In the temporal evaluation, performance still decreases. Compared to the Q-only scenario, the performance goes up a little in the calibration period (+0.14 in E_{SP}) and a bit more in the evaluation period (+0.19 and +0.18 in E_{SP}), as can be seen in Figure 21 in subsection 4.6 and Figure 79 in Appendix L.

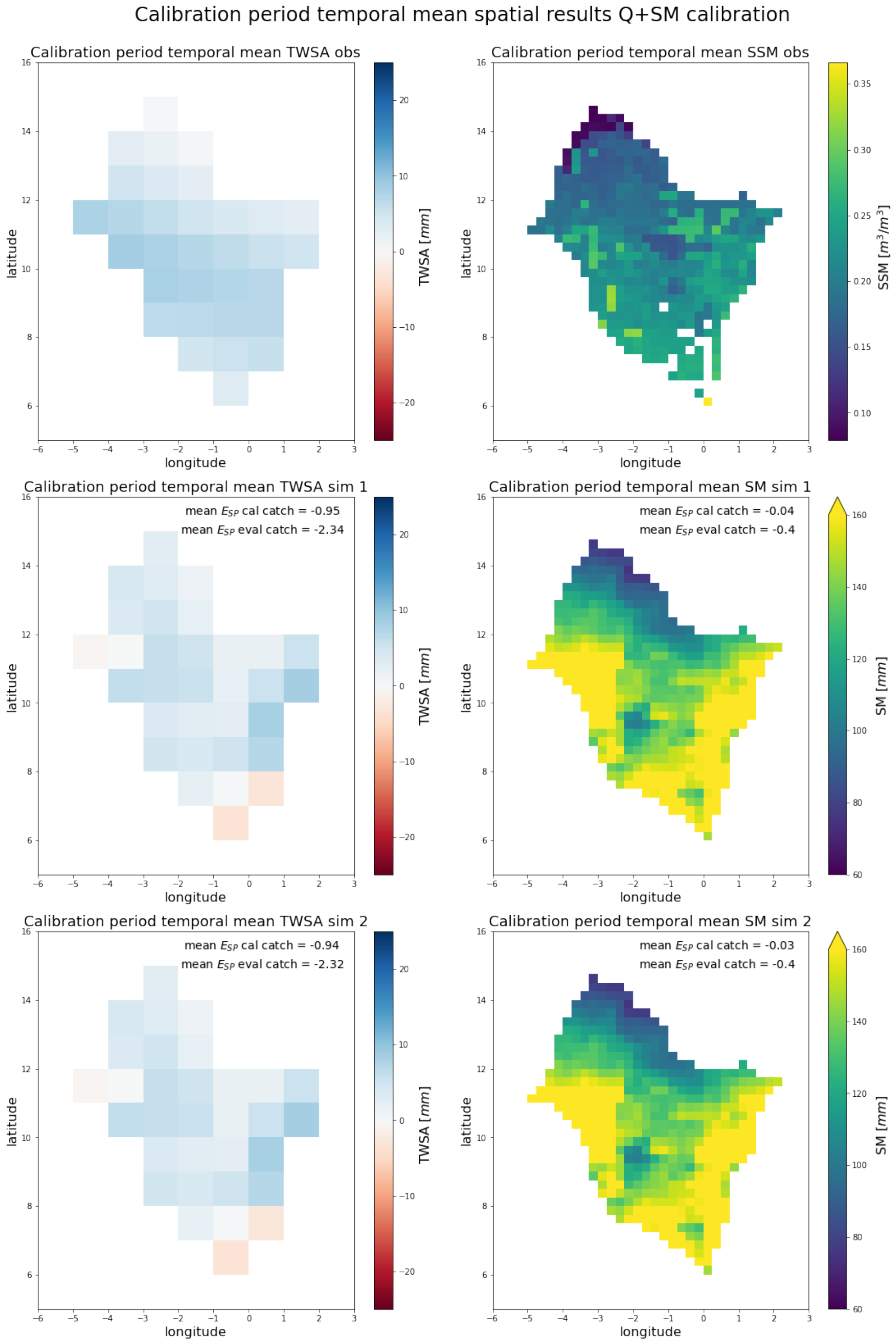

Calibration period temporal mean spatial results Q+SM calibration
Calibration period temporal mean TWSA obs
Calibration period temporal mean SSM obs
Calibration period temporal mean TWSA sim 1
mean E_{SP} cal catch = -0.95
mean E_{SP} eval catch = -2.34
Calibration period temporal mean SM sim 1
mean E_{SP} cal catch = -0.04
mean E_{SP} eval catch = -0.4
Calibration period temporal mean TWSA sim 2
mean E_{SP} cal catch = -0.94
mean E_{SP} eval catch = -2.32
Calibration period temporal mean SM sim 2
mean E_{SP} cal catch = -0.03
mean E_{SP} eval catch = -0.4
latitude
longitude
16
14
12
10
8
6
−6
−5
−4
−3
−2
−1
0
1
2
3
TWSA [mm]
20
10
0
−10
−20
SSM [m^3/m^3]
0.35
0.30
0.25
0.20
0.15
0.10
SM [mm]
160
140
120
100
80
60

Figure 11: Spatial plots of the mean TWSA observations and simulations in the calibration period in the Volta basin (left). Spatial plots of the mean surface soil moisture observations and the mean simulated amount of water in the soil moisture reservoir in the calibration period in the Volta basin (right).

4.3 Results scenario 3: Q + TWSA Calibration

In this subsection the results of the calibration on Q and TWSA observations are discussed. Again, both datasets have had an equal weight in the optimization function and optimization is done for both the temporal and spatial patterns of the TWSA at the same time (subsubsection 3.3.4). All figures with the results of this calibration scenario can be found in Appendix I. Most of the results are very similar to what already was observed in the previous scenarios so the description of this subsection will describe the differences with the other scenarios. The performance scores for the separate and combined streamflow objective functions for all stations for both simulations are given in Table 23 and Table 24, and the performance scores for the RS datasets in all calibration and evaluation cases for both simulations are given in Table 25 and Table 26. The observed and both simulated hydrographs for the stations in the Black Volta, White Volta and Oti are given in Figure 48, Figure 49 and Figure 50 respectively. Temporal patterns of TWSA and SM observations and simulations are given in Figure 51 and Figure 52 for the calibration and evaluation catchments, and their spatial patterns are given in Figure 53 and Figure 54 for the calibration and evaluation period. Lastly, the T-values found over the basin in both simulations in this scenario are given in Figure 55 and the parameters corresponding to the simulations are given in Table 27.

Streamflow performance

The streamflow performance in the calibration and evaluation catchments and periods is slightly different for the 2 simulations. In the calibration period and calibration catchments, the mean KGE is 0.63 for both simulations, but in the temporal evaluation, mean KGE values are 0.59 and 0.57, in the spatial evaluation 0.40 and 0.44 and in the spatio-temporal evaluation 0.38 and 0.42, for simulation 1 and 2 respectively. Mean performance values in the calibration catchments are slightly lower than in the baseline scenario (-0.01 in calibration period, -0.03 in evaluation period), but mean performance values in the evaluation catchments are higher (+0.05 in both the calibration and evaluation catchments).

It can be seen in the hydrographs in Figure 15 till Figure 17 in subsection 4.6 and Figure 69 till Figure 73 in Appendix L, that in general the streamflow simulations are very similar to the ones in the baseline scenario, but that the peaks in this scenario are a bit higher in most of the years. This negatively impacts streamflow performance in the Black Volta and at Porga in the Oti, but positively impacts streamflow performance in the White Volta and at Saboba and Sabari in the Oti. The same change was observed for the calibration on Q+SM, but in this calibration scenario, the effect is much smaller. This can also be seen back in the much closer mean KGE values.

TWSA performance

In this scenario, the resulting spatially averaged temporal patterns of the 2 simulations differ from each other. This results in E_{TMP} values of 0.34 and 0.22 in the calibration catchments and period for simulation 1 and 2, respectively, and E_{TMP} values of 0.29 and 0.15 in the temporal evaluation. The spatial evaluation results in values of 0.32 and 0.18, and the spatio-temporal evaluation in scores of 0.24 and 0.08 for simulation 1 and 2, respectively. Simulation 1 clearly outperforms simulation 2, which is due to the larger amplitude in the temporal TWSA pattern, that better mimics the observed one, in all catchments and periods.

The mean of the two simulations is compared with the mean of the simulations based on Q-only calibration. This comparison shows that in the calibration catchments, the E_{TMP} goes down with 0.09 in both the calibration and evaluation periods. In the evaluation catchments, the E_{TMP} also goes down with 0.09 in the calibration period and with 0.13 in the evaluation period. When comparing the mean baseline results with simulation 1 of this scenario, a smaller reduction in performance values is observed.

Spatially, the 2 simulations are also slightly different, which can not be clearly seen back in the plots, but it can be seen in the performance values. Spatially not simulation 1, but simulation 2 is performing better. In the calibration period and the calibration catchments, the mean E_{SP} values are -0.69 and -0.62 for simulation 1 and 2. In the temporal evaluation, these scores change to -0.7 and -0.56. In the evaluation catchments, the E_{SP} values are lower than in the calibration catchments, with values of -1.79 and -1.49 in the calibration period for simulation 1 and 2, and values of -1.32 and -1.14 in the evaluation period for simulation 1 and 2. This means an increase in performance in simulation 2 in the evaluation period in all catchments, and also for simulation 1 in the evaluation catchments. The spatial evaluation however shows a large decrease in model performance.

It is observed in Figure 20 in subsection 4.6 and Figure 78 in Appendix L that mean E_{SP} score of both simulations goes up with 0.14 in the calibration catchments and period, 0.21 in the temporal evaluation, 0.32 in the spatial evaluation and 0.28 in the spatio-temporal evaluation. So the simulated spatial pattern improves a lot compared to the baseline scenario.

SM performance

Temporally, the two soil moisture simulations also slightly differ from each other. The lowest extremes are very similar but simulation 1 has much higher peaks. The soil moisture values also rise earlier each year in simulation 1 than in simulation 2 and they also drop later in the year, which seems to much better resemble the observed soil moisture pattern than simulation 2. The E_{TMP} scores corresponding to these simulations are -0.95 and -1.2 in the calibration catchments and period and -1.01 and -1.23 in the temporal evaluation, for simulations 1 and 2 respectively. The spatial evaluation values are -0.92 and -1.14 and the spatio-temporal evaluation values are -1.07 and -1.28 for simulation 1 and 2. Simulation 1 outperforms simulation 2 in all catchments and periods.

In Figure 19 in subsection 4.6 and Figure 52 in Appendix L it can be seen that the mean of those 2 simulations is generally lower than the simulations in the baseline scenario in both the calibration and evaluation catchments. This results in lower E_{TMP} scores in all catchments and periods compared to the baseline scenario (order -0.2).

The spatial pattern similarity between observations and the two simulations looks bad because of the relatively low values of the SM simulation. However, only patterns are assessed and since these look more similar, the performance scores are not very low. The simulations differ from each other with simulation 1 outperforming simulation 2. The corresponding E_{SP} values are -0.22 and -0.43 in the calibration catchment and period, and -0.24 and -0.59 in the evaluation period, for simulation 1 and 2 respectively. In the evaluation catchments the E_{SP} values of simulation 1 and 2 are -0.58 and -0.79 in the calibration period and -0.56 and -0.84 in the evaluation period.

The mean spatial SM performance of the baseline scenario is higher than the mean spatial SM performance of this scenario in all calibration catchments and periods. The decrease in E_{SP} in the calibration catchment is -0.15 in the calibration period and -0.18 in the evaluation period. In the evaluation catchment the decrease is -0.14 in the calibration period and -0.19 in the evaluation period. In Figure 21 in subsection 4.6 and Figure 79 in Appendix L it can be seen that the E_{SP} values in this scenario are the lowest of all calibration scenarios.

4.4 Results scenario 4: Q + SM + TWSA Calibration

The results of the calibration on Q, SM and TWSA observations are discussed in this scenario. All datasets have had an equal weight in the optimization function and optimization was done for both the temporal and spatial patterns of SM and TWSA at the same time (subsubsection 3.3.4). The figures with the results of this calibration scenario can be found in Appendix J. The figures are not shown in this section because the simulated hydrographs and patterns are very similar to the results of the other scenarios. The performance scores for the separate and combined streamflow objective functions for all stations for both simulations are given in Table 28 and Table 29, and the performance scores for the RS datasets in all calibration and evaluation cases for both simulations are given in Table 30 and Table 31. The observed and simulated hydrographs in the Black Volta, White Volta and Oti subcatchments are given in Figure 56 in Appendix L, Figure 12 and Figure 57 in Appendix J, respectively. Temporal patterns of TWSA and SM observations and simulations are given in Figure 58 and Figure 59 for the calibration and evaluation catchments, and their spatial patterns are given in Figure 60 and Figure 61 for the calibration and evaluation period. Lastly, the T-values found over the basin in both simulations in this scenario are given in Figure 62 and the parameters corresponding to the simulations are given in Table 32.

Streamflow performance

The two simulations done for this calibration scenario differ only minimally. The mean KGE scores are 0.52 and 0.51 in the calibration catchments and period for simulation 1 and 2. The temporal evaluation results in mean KGE values of 0.44 and 0.43, the spatial evaluation in a value of 0.49 for both simulations, and the spatio-temporal evaluation in a value of 0.31 for both simulations (See Figure 12). This means a decrease w.r.t. to the baseline scenario of 0.12 in the calibration catchments and period, a decrease of 0.17 in the temporal evaluation, an increase of 0.12 in the spatial evaluation and a decrease of 0.04 in the spatio-temporal evaluation.

In Figure 15 till Figure 17 in subsection 4.6 and Figure 69 till Figure 73 in Appendix L it can be seen that the mean simulated hydrographs are almost similar to the ones in the Q+SM calibration scenario, only the highest peaks are just a bit higher. This means the simulated flow is also generally much higher than the simulated flow in the baseline scenario. The streamflow performance values are therefore also very comparable to the Q+SM calibration scenario for all catchments and periods.

TWSA performance

The 2 simulations of this scenario result in very similar spatially averaged temporal TWSA pattern simulations, in both the calibration and evaluation catchments. The E_{TMP} values in the calibration catchments and period is 0.43 for both simulations, the temporal evaluation results in E_{TMP} values of 0.37 for both simulations. The spatial evaluation results in E_{TMP} values of 0.45 and 0.46 for simulations 1 and 2 and the spatio-temporal evaluation results in E_{TMP} values of 0.39 for both simulations. These values are all exactly equal to the ones in the Q+SM calibration scenario. The mean of the 2 simulations of these calibration scenarios are also almost exactly similar, as can be seen in Figure 18 in subsection 4.6 and Figure 74 in Appendix L. Only the amplitude of the anomaly is just a bit smaller compared to the Q+SM calibration scenario.

Spatially, the observed mean TWSA pattern is again very similar to the ones found in other calibration scenarios, but the 2 simulations are slightly different, with simulation 1 outperforming simulation 2. The mean performance scores of both simulations are somewhere in between the values of the other three scenarios discussed in this chapter and can be found in Table 11 at the end of this chapter.

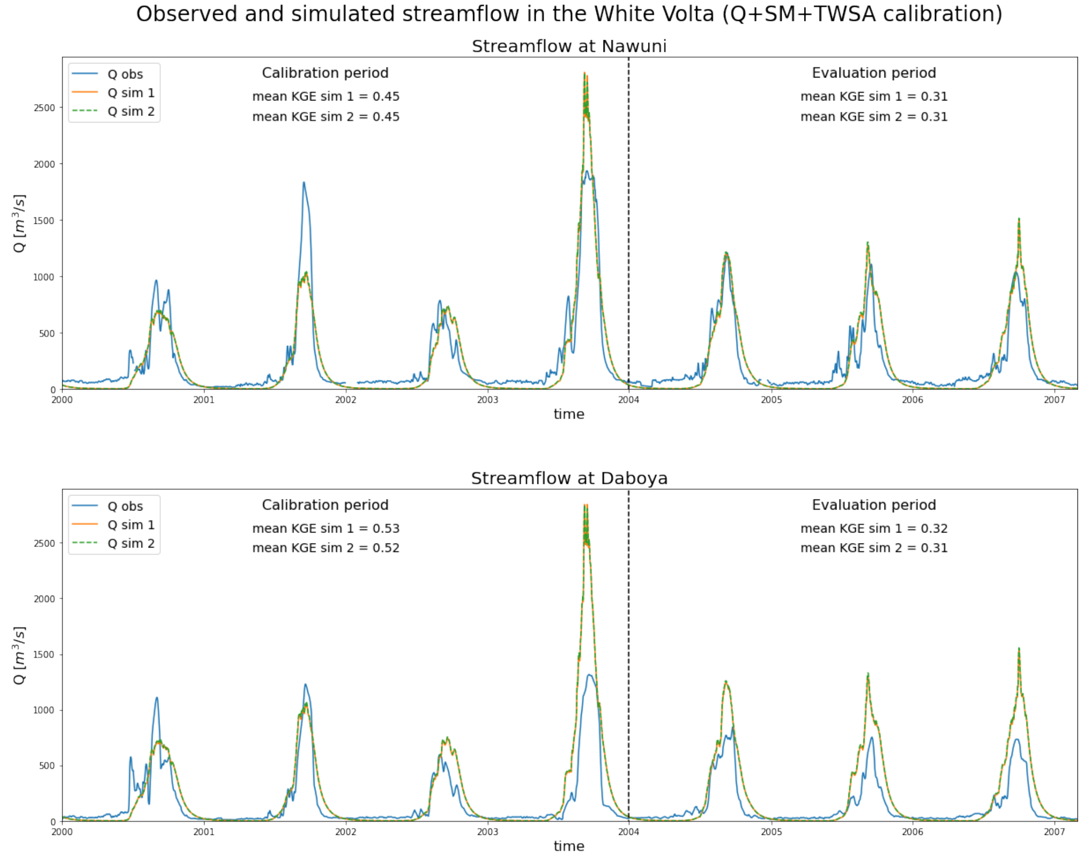


Figure 12: Hydrographs of the stations in the White Volta based on calibration on Q+SM+TWSA. These stations are evaluation stations. The stations Yaragu and Pwalagu were excluded from evaluation because of the high influence of reservoir management on the river flow in this area. The mean KGE values given in the plots are mean values of the 5 objective functions used for calibration as was explained in subsubsection 3.3.2.

SM performance

Temporally, the simulated SM pattern looks again very good but the performance scores are still relatively low because of a high variability term. Simulation 1 slightly outperforms simulation 2 with E_{TMP} values of -0.46 and -0.51 in the calibration catchments and period, and E_{TMP} values of -0.6 and -0.65 for the temporal evaluation. In the spatial evaluation, the E_{TMP} values drop to -0.49 and -0.54, and in the spatio-temporal evaluation to -0.68 and -0.74 for simulation 1 and 2. The mean simulated temporal SM pattern is again very close to the pattern of the Q+SM calibration scenario (See Figure 19 in subsection 4.6 and Figure 77 in Appendix L), but the peaks are just a bit lower. This results in an increase in performance in all catchments and periods w.r.t. the baseline and the Q+TWSA calibration scenarios, but a decrease compared to the Q+SM calibration scenario.

Spatially, the results are again very similar to the results of the Q+SM calibration scenario, as can be seen in Table 11, but the Q+SM scenario performs slightly better. The soil moisture values are higher than the baseline scenario but the observed pattern is very similar. Again the evaluation catchments slightly outperform the calibration catchments. The simulations are also more similar for the spatial pattern than for the temporal one. Mean E_{SP} values differ maximally 0.01 from each other.

4.5 Results scenario 5: SM + TWSA Calibration

The results of the calibration on only RS observations (SM and TWSA) are discussed in this subsection. Both datasets have had an equal weight in the optimization function and optimization was done for both the temporal and the spatial patterns of SM and TWSA at the same time (subsubsection 3.3.4). The figures with the results of this calibration scenario can be found in Appendix K. Some hydrographs are shown in this section but the mean spatial and temporal patterns of RS data were again very similar to the simulated patterns in other scenarios and can therefore be found in the Appendix. The performance scores for the separate and combined streamflow objective functions for all stations for both simulations are given in Table 33 and Table 34, and the performance scores of the RS datasets in all calibration and evaluation cases for both simulations are given in Table 35 and Table 36. The observed and simulated hydrographs in the Black Volta, White Volta and Oti subcatchments are given in Figure 13, Figure 63 in Appendix K and Figure 14, respectively. Temporal patterns of TWSA and SM observations and simulations are given in Figure 64 and Figure 65 for the calibration and evaluation catchments, and their spatial patterns are given in Figure 66 and Figure 67 for the calibration and evaluation period. Lastly, the T-values found over the basin in both simulations in this scenario are given in Figure 68 and the parameters corresponding to the simulations are given in Table 37.

Streamflow performance

The hydrographs of the two simulations performed in this scenario are very similar, but the performance scores differ a lot per subcatchment. The mean KGE's found in the calibration period and calibration catchments are 0.02 and 0.01 for both simulations, which are the lowest values found in all calibration scenarios. In the temporal evaluation the mean KGE's decrease to -0.26 and -0.27, but in the spatial evaluation the mean KGE's increase to 0.28 for both simulations. The spatio-temporal evaluation results in mean KGE's of 0.04 for both simulations. It is observed that all the subcatchments in the Black Volta have a relatively low performance in the calibration case (all mean KGE's < 0, See Figure 13), while the subcatchments in the Oti have a relatively high performance (all mean KGE's > 0, See Figure 14), with the stations Saboba and Sabari actually having the best model performance of all scenarios (See Figure 17 in subsection 4.6 and Figure 73 in Appendix L), in both the calibration and evaluation period. The mean KGE's of the subcatchments in the White Volta are somewhere in between those of the Black Volta and the Oti.

In Figure 15 till Figure 17 in subsection 4.6 and Figure 69 till Figure 73 in Appendix L, it can be seen that calibration on SM and TWSA generally results in this highest flow values compared to all other calibration scenarios. This results in the lowest mean KGE values of all scenarios because often the simulated flow is too high. Only for Saboba and Sabari, where simulated flow values were too low in all other calibration scenarios, an increase in streamflow performance was observed compared to the other scenarios.

TWSA performance

Like in the hydrographs, the two mean temporal pattern simulations are very similar in both the calibration and evaluation catchments. All E_{TMP} values found for both simulations are exactly equal. These E_{TMP} values are 0.45 in the calibration catchments and period, 0.41 in the temporal evaluation, 0.45 in the spatial evaluation and 0.40 in the spatio-temporal evaluation. Comparing the mean temporal TWSA pattern of this calibration scenario to the other scenarios (See Figure 18 in subsection 4.6 and Figure 74 in Appendix L), it can be seen that the mean of the simulations of this scenario is very close to the mean of the simulations of the Q+SM calibration scenario. Also the performance scores differ only minimally.

Spatially, the simulated patterns are again very similar. Both simulations also have the same E_{SP} scores, being -0.06 in the calibration catchments and period, -0.02 in the temporal evaluation,

-0.42 in the spatial evaluation, and -0.35 in the spatio-temporal evaluation. So, the evaluation catchments outperform the calibration catchments, just like for streamflow. The simulated mean spatial TWSA pattern is similar to the mean spatial TWSA patterns in all other calibration scenarios, for both the calibration and evaluation period, as can be seen in Figure 20 in subsection 4.6 and Figure 78 in Appendix L. For the mean spatial TWSA pattern, this scenario is outperformed by the Q+TWSA calibration scenario, but is very close to baseline scenario.

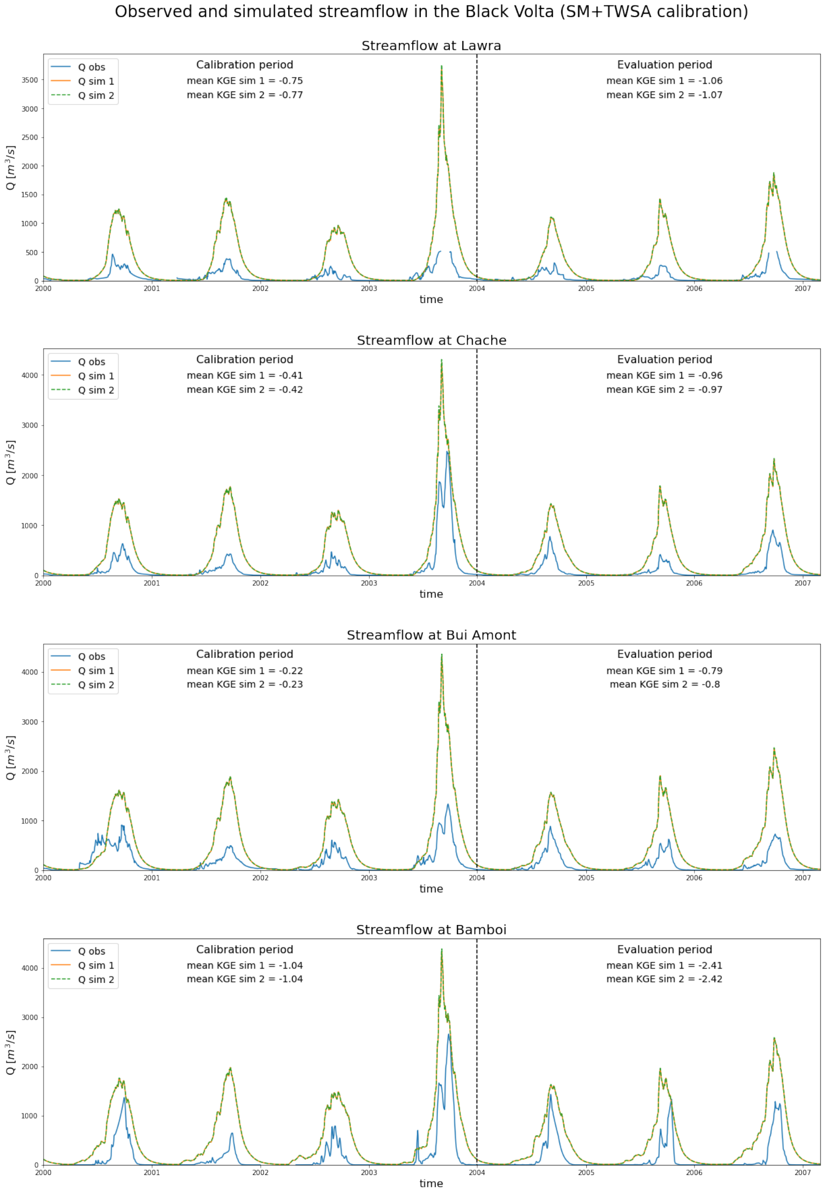


Figure 13: Hydrographs of the stations in the Black Volta based on calibration on SM+TWSA. These stations are calibration stations. The stations Boromo and Bamboi were excluded from calibration because of low data quality at Boromo, and far too high flow observations at Bamboi. The streamflow observations at Bamboi shown in this plot are scaled by catchment area to match the flow at the upstream station Bui Amont. The mean KGE values given in the plots are mean values of the 5 objective functions used for calibration as was explained in subsubsection 3.3.2.

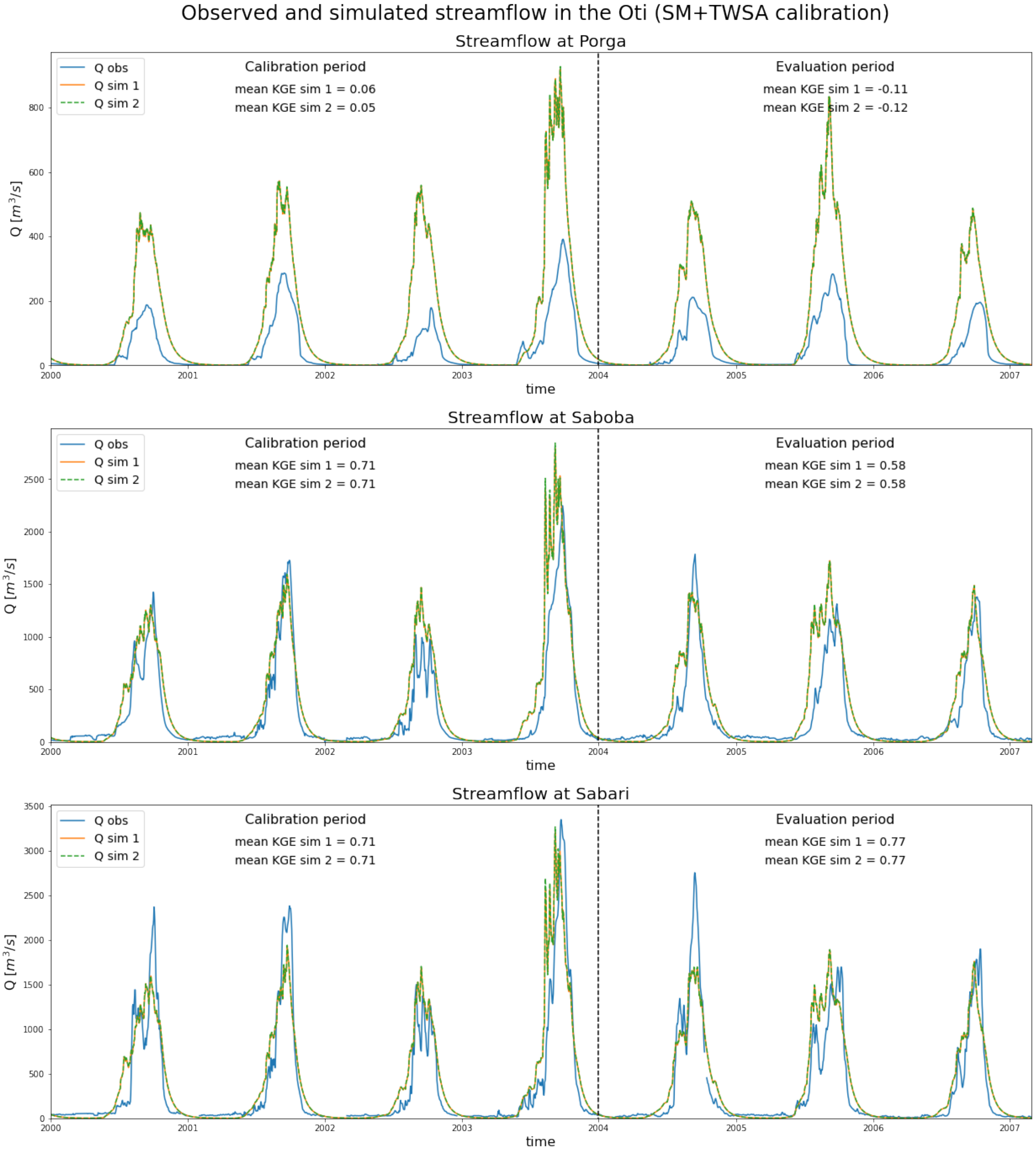


Figure 14: Hydrographs of the stations in the Oti based on calibration on SM+TWSA. These stations are calibration stations. The mean KGE values given in the plots are mean values of the 5 objective functions used for calibration as was explained in subsubsection 3.3.2.

SM performance

The simulations of the mean temporal SM pattern are also very similar but do not have exactly the same E_{TMP} score. In the calibration catchments and calibration period the E_{TMP} scores are -0.45 and -0.43 for simulation 1 and 2, for the temporal evaluation these scores are -0.59 and -0.57. The spatial evaluation results in E_{TMP} scores of -0.46 and -0.45 for simulation 1 and 2, and the spatio-temporal evaluation in values of -0.66 and -0.64. So, simulation 2 slightly outperforms simulation 1. Compared to other scenarios (Figure 19 in subsection 4.6 and Figure 77 in Appendix L), the mean simulated SM pattern is very close to the one simulated in the Q+SM calibration scenario. Also the performance scores are very close to each other, in all 4 the calibration and evaluation catchments and periods. This also means that the mean simulated temporal SM pattern is also better in this scenario than all other scenarios, except for the Q+SM case.

Spatially, both mean spatial pattern SM simulations are again very similar to the results of other simulations and to each other, resulting in also exactly the same E_{SP} values for both simulations. These values are -0.06 in the calibration case, -0.02 in the temporal evaluation, -0.42 in the spatial evaluation, and -0.35 in the spatio-temporal evaluation. This means that the evaluation catchments again outperform the calibration catchments. Compared to the other scenarios (Figure 21 in subsection 4.6 and Figure 79 in Appendix L), the simulated SM values are neither very high nor very low. The resulting E_{SP} values are very comparable to the other scenarios in which SM data is involved in the calibration, but the Q+SM and Q+SM+TWSA scenarios just outperform the SM+TWSA calibration scenario.

4.6 Overview Results

In this subsection, an overview of the results presented in this chapter is given, and the results of the different calibration scenarios are compared. Some key figures and tables are shown in this subsection, but additional figures and tables that compare the results of the different scenarios are given in Appendix L. A complete overview of the most important performance scores is given in Table 11.

Streamflow performance

In Table 11 the mean KGE values in the calibration and evaluation catchments and periods are given for every calibration scenario. It is observed that no single calibration scenario outperforms the baseline scenario in the calibration catchments for streamflow. In the evaluation catchments however, the streamflow performance of the baseline scenario in the calibration period is outperformed by the calibration on Q+SM, Q+TWSA and Q+SM+TWSA scenarios. In the evaluation period, the baseline scenario is outperformed bu the Q+TWSA scenario. A last observation that can be made is that the streamflow performance of the calibration scenarios Q+SM and Q+SM+TWSA are very similar, and that also the performance between the baseline and Q+TWSA scenario do not differ a lot. The performance of the scenario in which only RS data was used in the calibration (SM+TWSA) is the lowest of all scenarios.

Table 11: Overview of the performance scores for Q, TWSA and SM for all scenarios.

Overview Results		**Scenarios**				
Variable	**catch / period**	**Q-only**	**Q+SM**	**Q+TWSA**	**Q+SM+TWSA**	**SM+TWSA**
Q **mean** KGE	cal, cal	0.64	0.51	0.63	0.52	0.02
	cal, val	0.61	0.42	0.58	0.44	-0.27
	val, cal	0.37	0.49	0.42	0.49	0.28
	val, val	0.35	0.31	0.40	0.31	-0.04
TWSA (temp) **mean** E_{TMP}	cal, cal	0.37	0.43	0.28	0.45	0.45
	cal, val	0.31	0.37	0.22	0.40	0.41
	val, cal	0.36	0.46	0.25	0.47	0.45
	val, val	0.29	0.39	0.16	0.41	0.40
TWSA (spat) **mean** E_{SP}	cal, cal	-0.80	-0.95	-0.66	-0.88	-0.79
	cal, val	-0.84	-1.08	-0.63	-1.01	-0.81
	val, cal	-1.96	-2.33	-1.64	-2.23	-2.05
	val, val	-1.51	-1.82	-1.23	-1.76	-1.49
SM (temp) **mean** E_{TMP}	cal, cal	-0.87	-0.43	-1.08	-0.48	-0.44
	cal, val	-0.94	-0.57	-1.12	-0.63	-0.58
	val, cal	-0.86	-0.46	-1.03	-0.52	-0.46
	val, val	-0.99	-0.65	-1.17	-0.71	-0.65
SM (spat) **mean** E_{SP}	cal, cal	-0.18	-0.04	-0.33	-0.04	-0.06
	cal, val	-0.19	-0.00	-0.41	-0.01	-0.02
	val, cal	-0.54	-0.40	-0.69	-0.42	-0.42
	val, val	-0.50	-0.32	-0.70	-0.34	-0.35
OMP ED_{tot}	cal, cal	2.93	2.69	3.07	2.67	2.76
	cal, val	3.01	2.87	3.13	2.84	3.09
	val, cal	3.88	3.93	3.81	3.88	3.73
	val, val	3.62	3.60	3.65	3.58	3.50

For streamflow the mean WA_{KGE} of all calibration or evaluation stations were used. For TWSA, the temporal mean E_{SP} in the calibration and evaluation period and catchments, and the spatial mean E_{TMP} in the calibration and evaluation catchments were used. For SM the temporal mean E_{SP} in the calibration and evaluation period and catchments, and the spatial mean E_{TMP} in the calibration and evaluation catchments were used. OMP: Overall Model Performance using the ED of every score given here, See subsubsection 3.3.4.

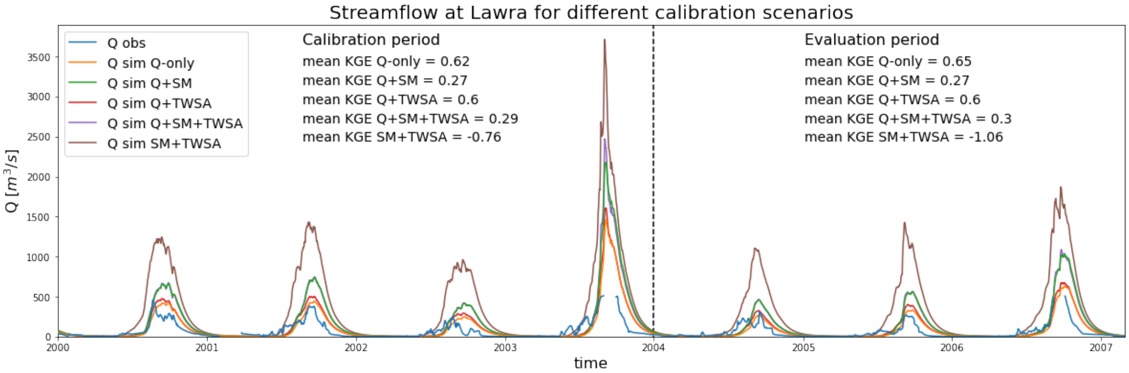


Figure 15: Hydrographs of the observed and simulated streamflow timeseries for all calibration scenarios at Lawra (Black Volta). Lawra is a calibration station. For each calibration scenario, the mean of the 2 simulations is shown. The mean KGE values given in the plots are mean values of the 5 objective functions used for calibration/evaluation as was explained in subsubsection 3.3.2.

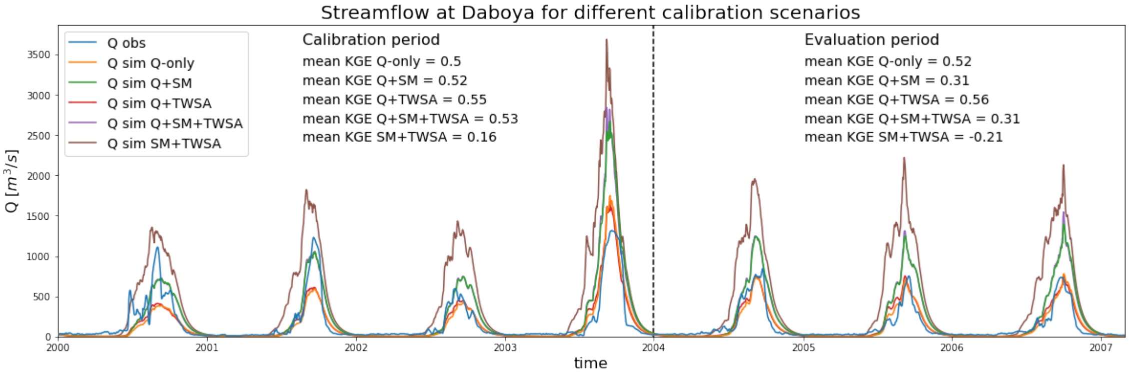


Figure 16: Hydrographs of the observed and simulated streamflow timeseries for all calibration scenarios at Daboya (White Volta). Daboya is an evaluation station. For each calibration scenario, the mean of the 2 simulations is shown. The mean KGE values given in the plots are mean values of the 5 objective functions used for calibration/evaluation as was explained in subsubsection 3.3.2.

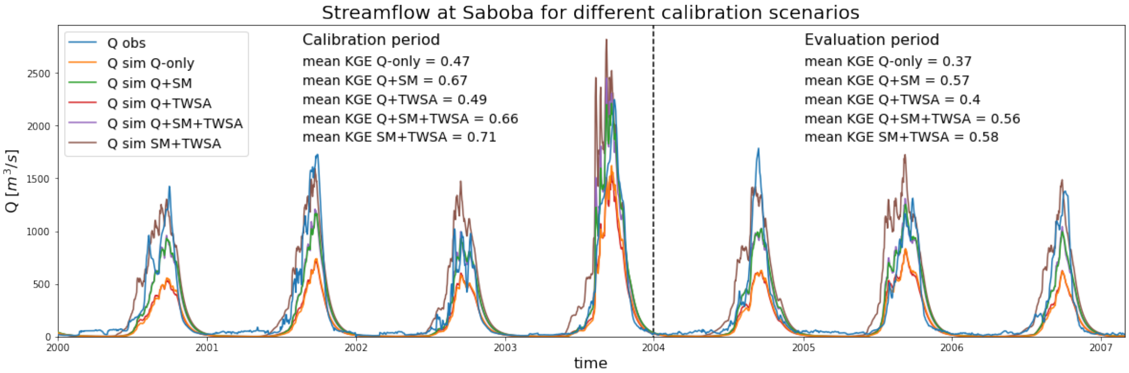


Figure 17: Hydrographs of the observed and simulated streamflow timeseries for all calibration scenarios at Saboba (Oti). Saboba is a calibration station. For each calibration scenario, the mean of the 2 simulations is shown. The mean KGE values given in the plots are mean values of the 5 objective functions used for calibration/evaluation as was explained in subsubsection 3.3.2.

In Figure 15 till Figure 17 the mean hydrographs of the 2 simulations performed for each scenario are shown in one plot for the stations Lawra (Black Volta), Daboya (White Volta) and Saboba (Oti). Similar plots for the other stations can be found in Figure 69 till Figure 73 in Appendix L. It is observed that the simulated flow in the SM+TWSA scenario is highest in all scenarios. The flow pattern is very similar for all scenarios. After SM+TWSA, the simulations of the Q+SM and Q+SM+TWSA are highest, and very similar in all hydrographs. The lowest simulated flow is observed in the Q+TWSA and Q-only calibration cases. The hydrographs in these calibration cases are also very similar. These similarities can also be seen back in the performance values. In general, the streamflow performance is highest for the calibration on Q-only, as is the case at Lawra (Figure 15). However, this is not always the case, as can be seen at Daobya (Figure 16), where higher streamflow performance values are found in other scenarios, and especially at Saboba (Figure 17), where the streamflow performance is highest for the SM+TWSA scenario.

Adding SM data to the optimization function often reduces streamflow performance. Sometimes this reduction is large (at Lawra, Figure 15), sometimes small (at Daboya, Figure 16) and sometimes the streamflow performance increases (at Saboba, Figure 17). The addition of TWSA data to the optimization function has little effect on the streamflow performance, while leaving out Q data has large effects, mostly negative, but sometimes positive.

TWSA performance (temporally)

The observed and simulated mean temporal TWSA patterns for all calibration scenarios are given in Figure 18 for the calibration catchments and in Figure 74 in Appendix L for the evaluation catchments. It is observed that the mean simulated temporal TWSA patterns of all calibration scenarios are very similar. The timing of the yearly extremes is almost always in the same month in all scenarios. However, the absolute values of these extremes differ, as do the slopes of the yearly rise and fall in TWS. The lowest extremes are simulated by the Q+TWSA scenario, and the highest by the Q+SM scenario. The best E_{TMP} scores are observed when SM data is included in the calibration, either with Q or TWSA data or without. When only TWSA data is included, the E_{TMP} actually decreases, because then also the spatial pattern of TWSA, which has a much lower performance score, need to be optimized. The same results apply to the temporal TWSA pattern in the evaluation catchments.

Overall, the Q+SM+TWSA scenario has the highest temporal TWSA performance in most catchments and periods. However, the performance scores of the SM+TWSA and also the Q+SM scenarios are very close. The baseline scenario comes next and the Q+TWSA scenario leads to the lowest E_{TMP} values.

TWSA performance (spatially)

The temporally averaged spatial patterns of the observed and simulated TWSA for all calibration scenarios are given in Figure 20 for the calibration period and in Figure 78 in Appendix L for the evaluation period. It can be observed that the patterns are inverses of each other, because the temporal mean over both periods is set to 0. The E_{SP} values are low in all scenarios. All simulated patterns are very similar, only the magnitude of the absolute values differs per scenario, however, only minimally. This is observed in both the calibration and evaluation period. The patterns do not resemble the general observed spatial pattern.

Calibration on TWSA only really makes a difference in the Q+TWSA scenario, but performance scores are still low. The worst performance is found in the Q+SM scenario. The baseline and SM+TWSA scenario perform almost equally, and the Q+SM+TWSA scenario is in between those two and the Q+SM scenario.

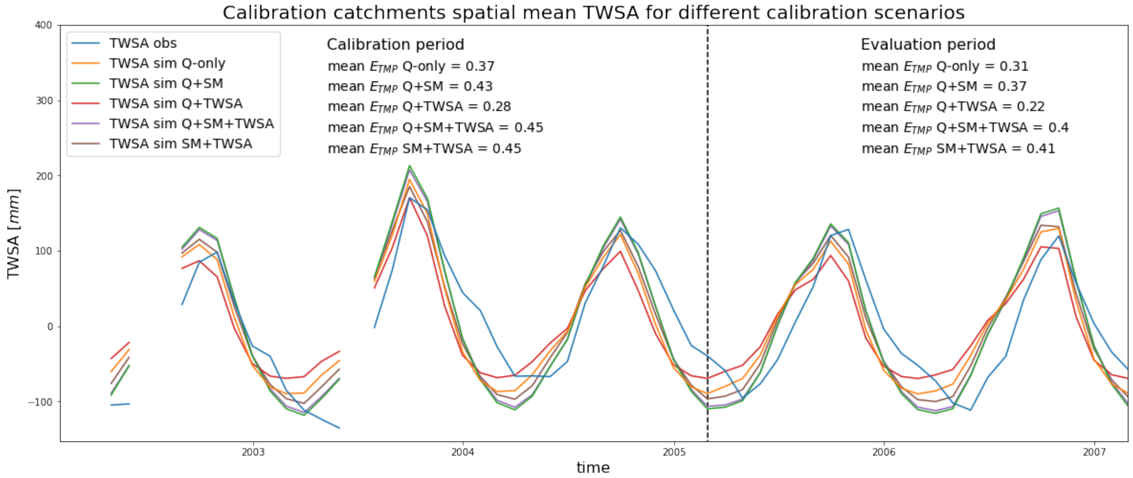


Figure 18: Timeseries of the mean TWSA observations and simulations within the calibration catchments for all calibration scenarios. For each calibration scenario, the mean of the 2 simulations is shown.

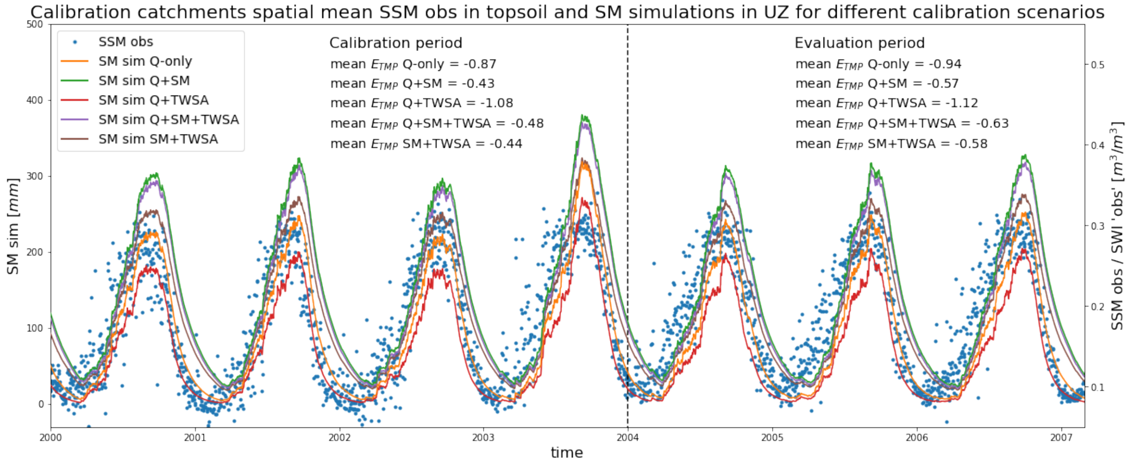


Figure 19: Timeseries of the mean surface soil moisture observations (SSM) and the simulated mean amount of water in the soil moisture reservoir (SM) within the calibration catchments for all calibration scenarios. For each calibration scenario, the mean of the 2 simulations is shown.

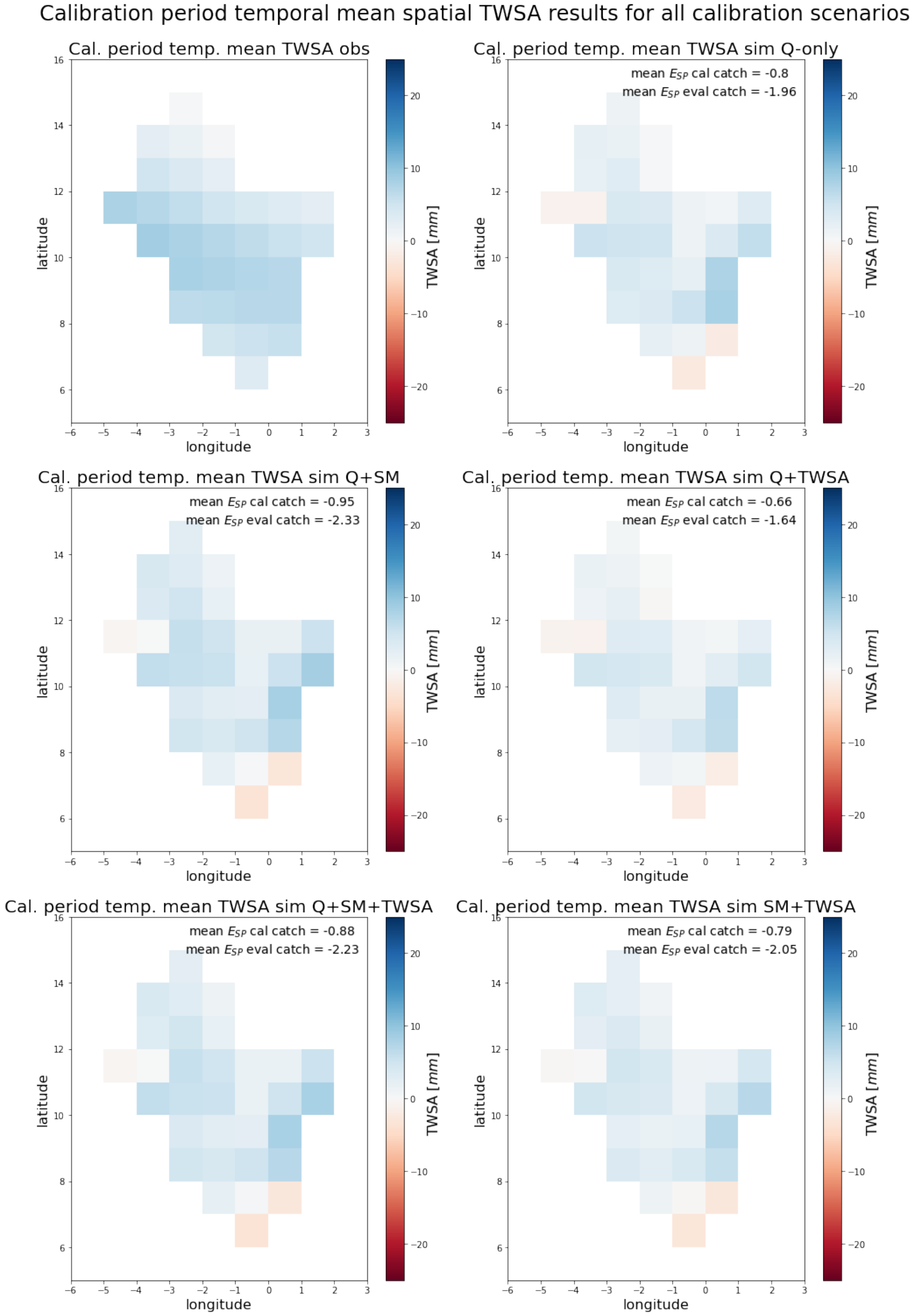

Calibration period temporal mean spatial TWSA results for all calibration scenarios
Cal. period temp. mean TWSA obs
Cal. period temp. mean TWSA sim Q-only
mean E_{SP} cal catch = -0.8
mean E_{SP} eval catch = -1.96
Cal. period temp. mean TWSA sim Q+SM
mean E_{SP} cal catch = -0.95
mean E_{SP} eval catch = -2.33
Cal. period temp. mean TWSA sim Q+TWSA
mean E_{SP} cal catch = -0.66
mean E_{SP} eval catch = -1.64
Cal. period temp. mean TWSA sim Q+SM+TWSA
mean E_{SP} cal catch = -0.88
mean E_{SP} eval catch = -2.23
Cal. period temp. mean TWSA sim SM+TWSA
mean E_{SP} cal catch = -0.79
mean E_{SP} eval catch = -2.05
latitude
longitude
TWSA [mm]
16
14
12
10
8
6
−6
−5
−4
−3
−2
−1
0
1
2
3
20
10
0
−10
−20

Figure 20: Spatial plots of the mean TWSA observations and simulations in the calibration period for all calibration scenarios. For each calibration scenario, the mean of the 2 simulations is shown.

SM performance (temporally)

The observed and simulated temporal SM patterns from all calibration scenarios are given in Figure 19 for the calibration catchments, and in Figure 77 in Appendix L for the evaluation catchments. The corresponding SWI 'observations' can be found in Figure 75 and Figure 76 in Appendix L, respectively. These SWI 'observations' are derived from SSM observations using the T-values shown in Figure 80 in Appendix L.

Again, the patterns of all SM simulations are very similar to each other. The largest difference is in the absolute values of the yearly extremes and timing of the lowest yearly extreme value. The highest peaks are observed in the simulations of the Q+SM and Q+SM+TWSA scenario. Then follows the SM+TWSA scenario an then the baseline scenario and the Q+TWSA scenario. The best performance in terms of E_{TMP} value is found in the Q+SM scenario, buth the SM+TWSA and Q+SM+TWSA also have relatively high E_{TMP} values. The baseline scenario comes next and the Q+TWSA scenario performs the worst. In the evaluation catchments, the performance scores of the Q+SM and the SM+TWSA calibration scenarios are exactly equal, although the simulations differ, especially at the peak SM values. The other calibration scenarios show similar results as in the calibration catchments.

The resulting T-values found in the calibration of each scenario shows that there are large differences between the subcatchments, but also between the scenarios (See Figure 80). Generally, T-values are higher in the north of the basin, and lower in the south of the basin. The highest T-values are found in cases where SM data is included in the calibration, and the lowest values are found in the Q-only and especially the Q+TWSA calibration scenario. This also explains the SWI values shown in Figure 75 and Figure 76 in Appendix L. The SWI values derived in the Q-only and Q+TWSA calibration cases are closer to the SSM observations, while the SWI timeseries of the other calibration cases are more influenced by the high T-values.

SM performance (spatially)

The temporally averaged spatial SM patterns are given in Figure 21 and Figure 79 in Appendix L. The absolute values of the spatial SM patterns differ a lot per scenario, but the pattern of all simulations is very similar. However, these patterns do not resemble the SSM observations as good as the pattern looks temporally. The lowest simulated SM values are found in the Q+TWSA scenario and the Q-only scenario, and the highest in the Q+SM scenario. This was alreayd observed in the temporal patterns. The highest E_{SP} performance scores are found in the scenario in which SM data is included in the calibration, with the Q+SM scenario just outperforming the Q+SM+TWSA and SM+TWSA scenario, and the lowest E_{SP} values are found in the Q+TWSA scenario. The baseline scenario is in between those two groups. This holds for all calibration catchments and periods.

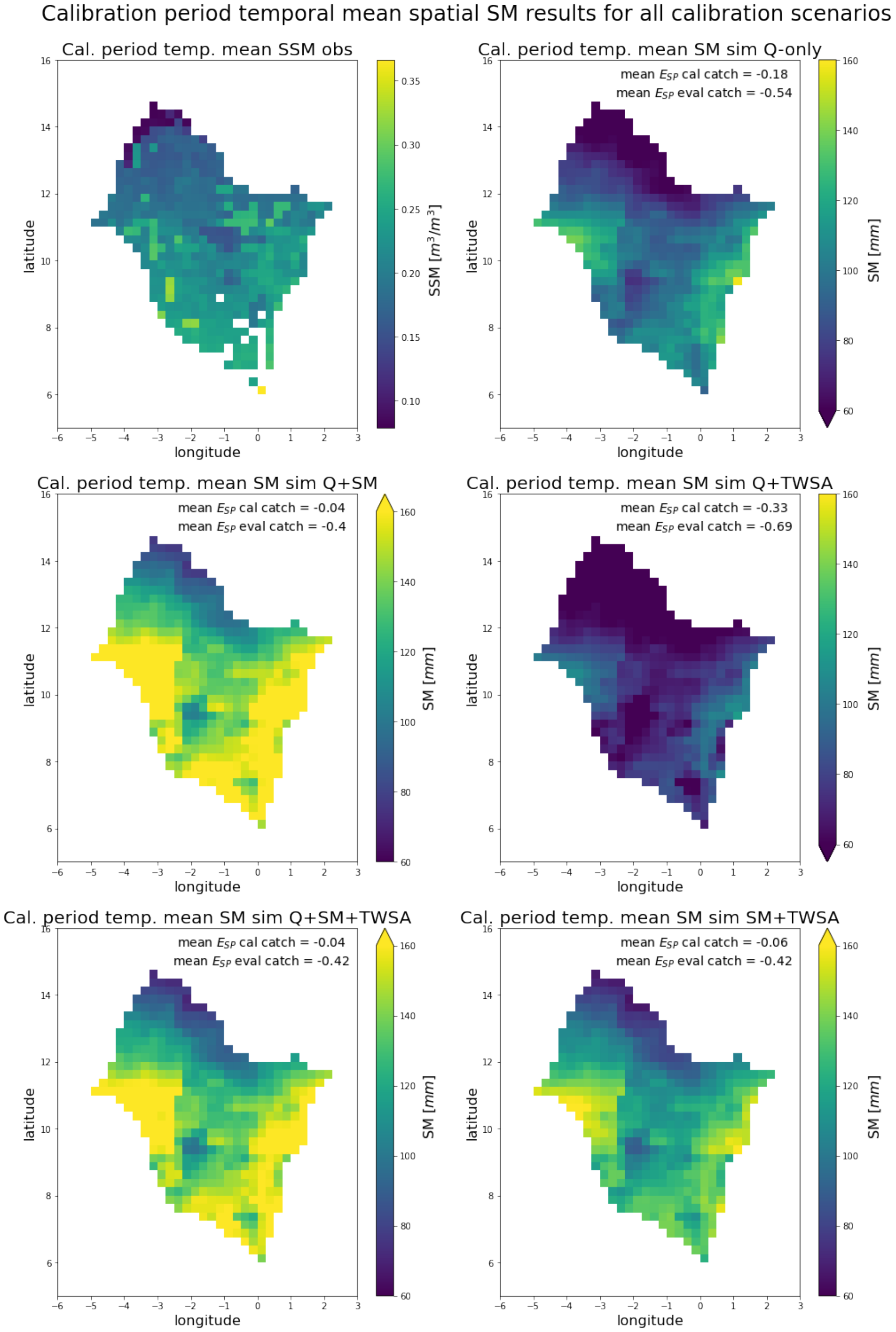

Calibration period temporal mean spatial SM results for all calibration scenarios
Cal. period temp. mean SSM obs
Cal. period temp. mean SM sim Q-only
mean E_{SP} cal catch = -0.18
mean E_{SP} eval catch = -0.54
Cal. period temp. mean SM sim Q+SM
mean E_{SP} cal catch = -0.04
mean E_{SP} eval catch = -0.4
Cal. period temp. mean SM sim Q+TWSA
mean E_{SP} cal catch = -0.33
mean E_{SP} eval catch = -0.69
Cal. period temp. mean SM sim Q+SM+TWSA
mean E_{SP} cal catch = -0.04
mean E_{SP} eval catch = -0.42
Cal. period temp. mean SM sim SM+TWSA
mean E_{SP} cal catch = -0.06
mean E_{SP} eval catch = -0.42
latitude
longitude
SSM [m^3/m^3]
SM [mm]
0.10
0.15
0.20
0.25
0.30
0.35
60
80
100
120
140
160
6
8
10
12
14
16
−6
−5
−4
−3
−2
−1
0
1
2
3

Figure 21: Spatial plots of the mean surface soil moisture observations (SSM) and the simulated amount of water in the soil moisture reservoir (SM) in the calibration period for all calibration scenarios. For each calibration scenario, the mean of the 2 simulations is shown.

5 Discussion

This chapter is divided into four separate parts. First some changes to the methodology explained in Chapter 3 are discussed in subsection 5.1, based on the first simulation results. Then the interpretation of the results of the baseline scenario and the other calibration scenarios is given in subsection 5.2, based on Chapter 4. In this section also possible and likely causes will be given to explain why the results are as they are and possibilities for improvement of the methodology and the results will be brought up. The model results will also be compared to 2 studies using similar methods for hydrological model calibration. In subsection 5.3 is commented on the methodology used in this study. This is done because the data, methodology and results presented in this study should be interpreted taking into account the assumptions, the uncertainties and errors corresponding to each of those matters. The uncertainty in the data and the assumptions taken in the methodology were already extensively discussed in Chapter 2 and Chapter 3, in order to provide the reader with the necessary prior knowledge to interpret the results. However, several important issues still need to be stated here. The discussion ends with a general message about what was learnt from this study and how this can be used in hydrological modelling.

5.1 Changes to the Methodology

Based on the first results of the Q-only calibration, several changes were made to the methodology described in Chapter 3, which are described in this section. First, it is explained why several changes were necessary to the original set of streamflow observations used for model calibration and evaluation, and then some adjustments and conclusions for the different RS datasets are described.

Adjustments in the used streamflow observations
The hydrographs in Figure 37 and Figure 38 in Appendix G and Figure 7 in subsection 4.1 show that the model is not able to simulate streamflow in all subcatchments equally accurate. This can be partly due to the model structure, the used parameters and the model input data, but can also be due to low quality streamflow observations, or river flows being heavily influenced by human actions, like (upstream) reservoir management. Based on these considerations it was chosen to leave some of the streamflow timeseries out of the calibration end evaluation procedure.

Boromo (subcatchment 1) is the most upstream gauging station in the Black Volta and the data from this station was not used for model calibration and evaluation. This choice was made because the amount of days with observations was with hindsight found to be insufficient (Table 1) and because the data shows flow peaks each year before the wet season (See Figure 37), while in this time of the year almost no rain is falling in this subcatchment. These flow observations are therefore either due to human actions or are simply false observations. Either way, the observations can not be used for calibration and evaluation of a model that simulates the natural river flow and are therefore not used in the rest of this study.

Bamboi (subcatchment 5) is the most downstream gauging station in the Black Volta and also the data from this gauging station is excluded from the model calibration and evaluation procedure. This is done because the data at Bamboi indicate much higher (almost double) river flow values compared to the gauging station Bui Amont (subcatchment 4) which is located only a relatively small distance upstream (See Figure 1). This increase in river flow is not likely (physically not possible) to occur and therefore the data from this station is not used. However, the streamflow observations are scaled to match the fraction of subcatchment area with subcatchment 4, so that the pattern of streamflow observations can be used an an extra model evaluation. These scaled observations are also shown in Figure 37 in Appendix G.

Yaragu and Pwalagu are the 2 most upstream gauging stations in the White Volta and the observations from these stations are excluded from the model evaluation. This has not only to do with the low amount of observations in the model period (See Table 1), but even more so with the influence of human actions such as the management of large reservoirs in the upstream part of the White Volta. These reservoirs can clearly be observed from the NDVI observations in Figure 9 in subsection 4.1 (middle left plot), of which large waterbodies are filtered out. Reservoir management results in high peak flows and a baseflow (See Figure 38 in Appendix G, top and second from above) which can both not be simulated with this model setup and forcing data.

This means that model calibration is done based on the timeseries of the gauging stations Lawra (subcatchment 2), Chache (subcatchment 3) and Bui Amont (subcatchment 4) in the Black Volta, and the timeseries of the gauging stations Porga (subcatchment 10), Saboba (subcatchment 11) and Sabari (subcatchment 12) in the Oti. The spatial model evaluation is done using the timeseries of the gauging stations Nawuni (subcatchment 8) and Daboya (subcatchment 9) in the White Volta. This setup for model calibration and evaluation was used for all calibration scenarios, including the baseline scenario.

Adjustments in the used RS data

In the calculation of the E_{SP} and E_{TMP} performance scores for the TWSA, a small adaptation to the procedure described in subsubsection 3.3.3 was made, because the calculation of the ratio of the coefficients of variation of observations and simulations was found to be mathematically unstable. Because the observations are anomalies w.r.t. a certain baseline period, the (temporal) mean of the observations is per definition zero. Since the baseline period used in this study is the calibration and evaluation period of GRACE in the model period (2002 - 2007), and the E_{SP} and E_{TMP} are calculated for the calibration and evaluation period separately, the temporal (but often also the spatial) mean of the observations is always very close to zero. This results in really high coefficients of variation and therefore this term in the calculation of E_{SP} and E_{TMP} was replaced with just the ratio of the standard deviations of the simulations and observations, like in the non-modified KGE (Gupta et al., 2009).

For the second RS dataset holds, that the relation between NDVI observations and AET simulations can mathematically be made because of the use of E_{SP}, which only assesses the spatial (or temporal) pattern of observations and simulations. This makes it possible to compare correlated variables, but it doesn't cancel out their physical differences. NDVI says something about the reflectance of specific wavelengths of radiation in the visible spectrum and observing it is hindered by clouds. AET is greatly influenced by the presence of vegetation, but also by the time of the year, the local climate and the weather, which is largely influenced by the presence of clouds.

Next to these physical differences between NDVI and AET, there is also large uncertainty in the datasets used. The NDVI observations are contaminated with a lot of cloud observations and contain a lot of noise. An attempt was made to filter out these problems as was described in subsubsection 2.1.3, but it is unclear whether this approach was effective. Because of the physical differences between NDVI and AET and the uncertainty in the observations, it was chosen to not use NDVI data in other calibration scenarios. The errors this would introduce in the model would be passed on to other model variables and this would introduce errors to other model variables and hence decrease the performance scores of the model results.

For the last RS dataset holds that the relation between the SWI derived from SSM observations and a derived T-value, and the simulated amount of water in the soil moisture reservoir seems very strong. This can not directly be seen back in the E_{TMP} values, although these are much better than for the comparison between NDVI and AET. The physical relation between SSM and simulated SM is also much stronger, especially after transformation to SWI, which also physically makes sense to do. The good correlation does come back in reasonable temporal mean E_{SP} values, especially in the calibration catchments.

Because of the good results of the soil moisture simulations in subsection 4.1, it was chosen to first add soil moisture data to the calibration in scenario 2 to see if the spatial and temporal patterns of soil moisture simulations can be improved with respect to the observations. In the third scenario, TWSA observations were added to the streamflow observations. As was explained before in this section, NDVI observations were left out for the rest of this study. A fourth scenario added both TWSA and SSM observations to the calibration on streamflow, and in a last scenario an attempt was done to simulate streamflow using only TWSA and SSM observations in the calibration.

5.2 Interpretation of the Results

In this subsection, first the results for streamflow and the RS datasets will be discussed based on the Q-only calibration. Then the results of the other calibration scenarios will be discussed, including reasons for their differences and similarities, and possibilities to improve the methodology and the results. Also the changes in parameter values will be translated directly into hydrological processes and a comparison of the results from this study and 2 other studies will be made, as verification of the results and interpretation presented here.

Interpretation of the Q-only calibration results

First, the streamflow simulations of the baseline scenario are discussed. It was found in subsection 4.1 that the average mean KGE value for the different hydrological signatures of all stations was relatively high. The reason that this satisfactory mean KGE value was found can have multiple causes. The first one is good input data, but as is shown in Appendix B, precipitation input may not be that accurate compared to point observations, although the total precipitation over several years is very accurate. A second reason is a good model setup and parameterization, but it was already explained in Chapter 3 that the parameterization is kept relatively simple and that only 1 model conceptualisation was used. A more likely reason for the relatively good performance is the use of 5 different hydrological signatures in the model calibration, meaning that the model is forced to simulate different aspects of the hydrograph as correctly as possible within this setup.

However, also a large spread in mean KGE values was observed per station with a pattern of flow values being higher compared to the observations in most of the upstream catchments. This effect is very likely the cause of the relatively simple model setup and parameterization (only 1 parameter set was used for the whole Volta basin!). This means that the calibration focuses on finding the average best performance for all stations at once. This approach could be improved a lot by calibrating every subcatchment (Figure 5) separately, so that a different parameter set per subcatchment can be found that best fits the dominant hydrological processes in that specific subcatchment. The Volta basin is simply too large to be described by one parameter set. Using multiple parameter sets would probably result in more comparable streamflow performance scores for every station, with also comparable flaws and parameter values monotonically in- or decreasing in the downstream direction . However, calibrating every subcatchment separately hugely increases the complexity of the model and the computational time needed for this calibration. Model assessment for every subcatchment also makes it more difficult and much more work to assess the simulated internal stocks and fluxes with RS data. Other options would be to use distributed parameter maps, which can be seen as the extended version of using parameter sets per subcatchment, or change the model conceptualisation.

In subsection 4.1 it was observed that the basin mean temporal TWSA pattern is simulated very accurately in the Q-only scenario in both the calibration and evaluation catchments (See Figure 8 and Figure 39). The most likely cause for this match is that the water balance is forced to be zero in the wflow hbv model, meaning that for a good calibration on streamflow, the storage in the model is also forced to be simulated accurately. The TWSA is also a bulk term, allowing different reservoirs (for different hydrological processes) in the hbv model to compensate for each other, and then also the basin-mean pattern is shown in this report. However, the assessed values given in Chapter 4 and Appendix G are means of the temporal pattern of every cell in the calibration or evaluation catchment, and do agree with the accurately simulated pattern.

The spatial pattern of TWSA values is simulated much less accurately. This can most likely be attributed to the simulations, but also partly to the observations. It was found that it is very difficult to simulate any spatial pattern using a simple parameterization with only 11 free parameters, as is implemented in this study, for the whole Volta basin. Again using different parameter sets per subcatchment, using distributed parameter maps or using a more complex model conceptualization would probably increase the spatial pattern performance of the TWSA. However, also the filter approach applied to the GRACE observations (See subsection 5.3) could introduce errors in the observations and hence a lower E_{SP} value.

The Q-only calibration results for the temporal and spatial AET patterns assessed with NDVI data were already discussed in subsection 5.1. It was concluded that the physical relation between the observed and simulated variables is too weak to calibrate a hydrological model on. NDVI data may be used only to check the timing and location of the most extreme values, but not to calibrate a full temporal or spatial evaporation pattern as output of a hydrological model on. A method may be needed that relates NDVI data much better to evaporation using other hydrological observations, but as far as is known to the author, such a method does not exist yet.

The simulated temporal pattern of soil moisture looks very good, as was observed in subsection 4.1. This good fit agrees with the fit of the temporal pattern of the TWSA, which is dominated by the temporal variability of the soil moisture reservoir. The strong seasonal pattern present in the Volta basin can clearly be observed and this is also directly the result of the input data with which the model is forced. The E_{TMP} values indicate that the mean SM performance is relatively low. This is due to a very high variability term. The two axis in the temporal plots like Figure 8 mask that the relative variability of the soil moisture observations and simulations compared to their mean values are actually not as similar as they look in the plots.

The accurate simulation of the basin mean temporal pattern is probably also caused by showing the basin mean, allowing errors in different parts of the basin to compensate for each other. This would also explain the poor spatial pattern representation shown in Figure 9 in subsection 4.1 and Figure 40 in Appendix G. However, the E_{SP} values show much better spatial than temporal SM pattern performance, which can also partly be caused by the much lower spatial than temporal variability. In this study the temporal mean TWSA value was used as benchmark for the anomaly values of the TWS spatio-temporal pattern. It was found that it is more correct to use the spatio-temporal mean TWSA value in the model period. However, the differences are small.

Interpretation of the results of the other calibration scenarios

The general change in all results of the different calibration scenarios compared to the baseline scenario is that the simulated pattern is very similar, but that the magnitude of the simulated variables is not. Timing and location of peaks and minima in streamflow values (See Figure 15 till Figure 17 in subsection 4.6 and Figure 69 till Figure 73 in Appendix L), TWSA values (See Figure 18 and Figure 20 in subsection 4.6 and Figure 74 and Figure 78 in Appendix L) and SM values (See Figure 19 and Figure 21 in subsection 4.6 and Figure 77 and Figure 79 in Appendix L) are all very similar for all different calibration scenarios, even for the non-Q calibration scenario. This also causes the differences between the performance scores of the RS datasets to be small.

Reasons for these small differences could again be sought in the relatively simple model setup, using only 1 parameter set of 11 free parameters to model the whole Volta basin. A trade-off effect is observed where for instance a good streamflow performance in one station is alternated with a much worse streamflow performance in another station. For streamflow this is probably due to the use of only 1 parameter set. The streamflow performance is best at the stations Saboba and Sabari for the SM+TWSA calibration scenario. This can mostly be attributed to the use of only 1 parameter set, and not only to the inclusion of RS data in the calibration. In the SM+TWSA calibration scenario, the magnitude of the flow values goes up and this benefits the streamflow performance at these specific stations. Using 1 parameter set per subcatchment is expected to give much more similar streamflow performance scores per station and to result in much more similar (e.g. higher values but same pattern for SM+TWSA) shifts in the hydrograph.

For the temporal and spatial patterns of the RS datasets, the use of 1 parameter set per subcatchments is probably not sufficient. Especially to improve the spatial patterns, the model complexity need to be increased, by using more complex model conceptualisations and multiple and larger model classifications of land use and for instance soil data. The overall best model performance in the calibration catchments is obtained using the Q+SM+TWSA calibration scenario (See Table 11), but the best model performance in the evaluation catchments in now found in the SM+TWSA calibration scenario. It makes sense that the best model performance in the calibration catchments was found in the Q+SM+TWSA scenario, because this scenario optimizes specifically for this score. However, it is harder to explain why the overall model performance in the evaluation catchments is best for the SM+TWSA scenario. This probably has to do with the relatively good streamflow performance in this scenario in the evaluation catchments, while also the TWSA and SM temporal and spatial performance scores are among the highest of all scenarios. The Q+TWSA calibration scenario is generally outperformed by all other calibration scenarios for the overall model performance. However, it should be noted again that the differences between the scenarios are small. It can also be concluded that adding RS data to the calibration does not improve model transferability to other catchments, for this specific model setup, because the spatial evaluation results in comparable overall model performance scores, and is only higher in the non-Q calibration scenario, which was explained earlier in this paragraph.

It is interesting to also discuss how the magnitude of the flow values can be improved for the non-Q calibration scenario. As is stated before, the general pattern of peaks and baseflow is right, but the absolute flow values are often wrong. Possibilities to improve this calibration can be sought in also assessing absolute values of temporal and spatial patterns of RS datasets, like TWSA and SM. This is already very well possible for the TWSA dataset within this model setup, but this is more difficult for the SM dataset, as the units of the observed and simulated variables are different. If this issue can be solved, than the KGE of the hydrograph of and possibly also of other hydrological signatures (See subsubsection 3.3.2) can be applied to the SM and TWSA dataset, which may improve streamflow performance scores. Also the inclusion of other datasets, like GLEAM or WaPOR for evaporation (95 % of the water in the Volta basin is evaporated (van de Giesen et al., 2001)), or water level data of Lake Volta from altimetry observations, may be very beneficial for a non-Q calibration. It should be noted however that errors within these datasets (and in the simulations) are passed on the last outgoing flux of the model, being streamflow. Because streamflow is the sum of all hydrological processes upstream of the river outlet (Beven et al., 1999), all these hydrological processes need to be simulated correctly for streamflow simulations to be accurate. This way of calibration is therefore difficult, but not impossible.

Discussion of the parameter values

The calibrated parameter values per model run for every calibration scenario can be found in Table 38. It is observed that some parameters hit the boundary of the range they were given based on the Q-only calibration results and that most simulations result in comparable parameter values for every calibration run within a specific calibration scenario, but that this is not always the case, indicating convergence to different (local) optima in the parameter space.

These parameter values can also directly be translated into hydrological differences between the scenarios. It is observed that generally, the ICF_{nf} parameter is higher when TWSA data is included. This means more interception evaporation from non-forest areas and also a smaller soil moisture reservoir capacity, leading to more evaporation. Also the $CEVPF$ parameters are higher for the Q+TWSA calibration scenario, leading to higher evaporation and hence less runoff. It is also observed that the inclusion of SM data to the calibration leads to lower $CEVPF$ parameter values, translating into less evaporation. This is counterbalanced by a higher LP for the calibration scenarios that include SM data, which leads to lower evaporation. The β parameter is lowest in the SM+TWSA calibration scenario. This leads to less evaporation, and more runoff because of faster seepage through the soil. The inclusion of TWSA data also leads to a lower SUZ value and thus earlier runoff, however, the SUZ threshold is almost never exceeded by the model in this scenario due to large evaporation values. The Q_{cf} parameter differs a lot for every calibration scenario, but this is mostly due to the relatively small contribution of this parameter to the model output. Large differences in this parameter value result in only very minimal changes in the results.

Comparison of the results to other studies

The results of this study will be compared with two similar studies to confirm if the results of this study agree with the results of other studies performed in the same niche of hydrological modelling. The first study used is Dembele, Hrachowitz, et al. (2020), in which a similar calibration approach was applied to the Volta basin using another model, and which was also consulted a lot in setting up this research project. The second study that was used is Hulsman et al. (2021), in which the authors also used a similar calibration setup, but applied this to the Luanga basin in Zambia, and also compared different model conceptualizations.

The KGE values found in this study in the Q-only calibration scenario are on average 0.66 for the calibration catchments in the calibration period. The KGE values of separate stations in this case range from 0.49 to 0.90 (most extreme values in sim 2, See Table 13 and Table 14 in Appendix G), which is very comparable to the values found in Figure 2 of Dembele, Hrachowitz, et al. (2020), with only the lowest KGE value found in this study for this case being about 0.15 points higher than in Dembele, Hrachowitz, et al. (2020). This means the baseline scenario is comparable to the one in Dembele, Hrachowitz, et al. (2020) and presents a reasonable estimation of the streamflow values in the Volta basin. In the study of Hulsman et al. (2021), the Nash-Sutcliffe efficiency (NSE) was used to assess streamflow performance so results are not comparable to the results of this study.

Both the studies of Dembele, Hrachowitz, et al. (2020) and Hulsman et al. (2021) showed very good representations of the temporal pattern of the TWSA, as was also observed in this study. In Dembele, Hrachowitz, et al. (2020), the GRACE TWSA dataset was implemented in a basin-average way, but in Hulsman et al. (2021), also the spatial pattern was assessed using the spatial efficiency metric (SPAEF) (Koch et al., 2018) (not spatial pattern efficiency metric E_{SP}, as was used in this study). Also here it was found that the spatial pattern performance scores of the TWSA were much lower than the temporal pattern performance scores, for all model conceptualisations (only 2 model calibration scenarios were used in this study), and that the spatial pattern performance scores were below the mean benchmark for almost every model conceptualisation. This benchmark is only outperformed by the Q+TWSA calibration scenario in this study.

The comparison of the evaporation performance of the three studies is more difficult because three different datasets were used in the studies, namely GLEAM in Dembele, Hrachowitz, et al. (2020), WaPOR in Hulsman et al. (2021) and NDVI in this study. It is clear however, that the performance scores for the temporal and spatial pattern of evaporation were much higher in the studies used for comparison than in this study. The temporal pattern is simulated very well in Dembele, Hrachowitz, et al. (2020) and also performance scores well above the average-benchmark are obtained in Hulsman et al. (2021). This was not the case in this study, also due to the use of NDVI data as assessment dataset, as was already explained in subsection 5.1. The spatial pattern performance scores for evaporation were lower than the temporal ones in both studies used for comparison. In this study, that observations was the other way around. The spatial pattern of evaporation seems to match (on the eye) the simulated pattern in Figure 7 of Dembele, Hrachowitz, et al. (2020) very well, so possibly the spatial pattern performance score of this study, assessed using the GLEAM dataset will be comparable to the values found in Dembele, Hrachowitz, et al. (2020).

In Hulsman et al. (2021), no SM data was used for model calibration or evaluation, but in Dembele, Hrachowitz, et al. (2020) the same dataset (ESA CCI) was used in the same basin but with another model. The temporal pattern representations were very good in both studies, only in Dembele, Hrachowitz, et al. (2020) the comparison of soil moisture observations and simulations was easier because the model used has a topsoil layer, which allows the modeller to directly compare simulated soil moisture output to soil moisture observations, which was not the case in this study. The E_{SP} values were in most calibration cases just below 0, and sometimes just above, which is very similar to what is observed in Figure 21. Only in this study no single calibration scenario exceeds an E_{SP} of 0. The lowest performance was obtained for the Q-only scenario in Dembele, Hrachowitz, et al. (2020), while this was the Q+TWSA calibration scenario in this study. However, this was not a calibration scenario in Dembele, Hrachowitz, et al. (2020). The highest soil moisture spatial pattern performance scores were achieved when SM data was included in the calibration in this study, while excluding soil moisture data in Dembele, Hrachowitz, et al. (2020) resulted in the lowest E_{SP} values for soil moisture.

Summarizing this comparison, it is observed that the basin mean temporal patterns of different RS datasets are simulated well by different models. Especially the temporal soil moisture pattern and the TWSA pattern can be simulated well and the resulting performance scores are very comparable, also for streamflow. The representation of spatial patterns is however more challenging, not only for TWSA but also for evaporation and SM. Assessment of simulated evaporation seems to give the best results using GLEAM data, and not only calibration of different (RS) datasets can improve the simulation of different hydrological stocks and fluxes, but also the use of a different model conceptualisations.

5.3 Discussion of the Data and Methodology

In this subsection, uncertainty in the used datasets, the technical problems encountered in this project and assumptions, motivations and improvements for the model setup are presented.

Uncertainty in data

The first datatype that is discussed here is streamflow. Streamflow is generally observed at gauging stations using Q-H relations, also know as rating curves. These Q-H relations need to be determined for each location separately in order to derive streamflow observations from simple water level measurements. The relations need to be updated or at least checked after a certain time but this is not always done. Q-H relations are also very uncertain for high flows, because observations at peak flows are seldom available. Therefore the streamflow observations may include errors related to this Q-H relation, especially the high flow observations.

A second issue with the streamflow observations is that they include non-natural processes such as effects of reservoir management. This is also one of the reasons why the observations at gauging stations Yaragu and Pwalagu in the White Volta were excluded from calibration and evaluation. However, it should also be mentioned that similar issues may be present in other streamflow timeseries, but just less obvious, and that these non-natural effects are also passed on to downstream gauging stations. In the case of Yaragu and Pwalagu, the reservoir management effects are passed on to (and can also be observed in) the timeseries of the stations Nawuni and Daboya. Calibrating on these observations may therefore introduce simulation errors, because the model than tries to simulate a non-natural effects using natural processes. This issue could partly be solved by using the wflow reservoir module for the most important reservoirs to model their effect.

The GRACE TWSA data used in this study suffers from several issues that need to be taken into account. Most of these issues are described in Landerer & Swenson (2012). GRACE TWSA observations are filtered to reduce the noise in the signal. However, by filtering the data one also loses part of the signal. A similar trade-off between data resolution and accuracy is described. A last important issue with GRACE TWSA observations is the 'leakage' of water from neighbouring cells. As can also be seen in the spatial plots of TWSA observations in Figure 9 in subsection 4.1 and Figure 40 in Appendix G, there is a strong relation between the TWSA value in a certain cell and the TWSA value in neighbouring cells. This is due to this 'leakage' effect and filtering of the data and this should be considered when using GRACE data.

The RS techniques used to do observations of NDVI and SSM data is already discussed in subsubsection 2.1.3 and the physical relation between NDVI and AET and SSM and SM is already discussed in subsection 5.1. Next to uncertainty in the NDVI observations due to clouds and SSM observations due to vegetation, also the physical relationships between NDVI and AET on the one hand, and SSM and SM on the other hand, are important. A last point that need to be made in this thesis is about the importance of (the availability of) accurate data for research projects. Without data it is impossible to carry out studies like these and therefore to do research on hydrological modelling. Collecting and making available accurate hydrological and meteorological observations should be promoted, although the developments in the field of remote sensing may offer solutions to more and more data problems.

Technical problems

Several technical problems were encountered in this research project. These technical problems often took a lot of time and computational and programming skills to solve. Since most hydrologists are not programmers, the technical issues encountered in this project are described in this section to learn from them, so that other students and researchers working in similar topics will be able to prevent or solve similar issues in their projects.

The first issue that will be discussed is the preparation of the data needed in this project. This was also the first step in building the model which took a lot of time. A lot of datasets were used in this study and all needed to be freely available and useable in the wflow hbv model framework. For this project, this often meant downloading huge datasets, clipping the data to the model area and converting the data to the right format (in this case PCRaster). All these steps needed to be taken for every dataset, because data is often available in either NetCDF or GeoTIFF format, but wflow is a PCRaster based model. This data inconsistency is often a time-consuming problem for (starting) hydrologists. PCRaster is also an old format that is used less and less (wflow also switched to Julia during this project), and documentation about how to produce the required input data (like PCRaster mapstacks) is not always available. Also the wflow documentation of the hbv model is incomplete. The only option to really understand how the model works is to look into the source-code, which is openly accessible, but not directly understandable for every project engineer.

The second large technical issue that was encountered in this project has to do with the automatic calibration setup. The DDS implementation of the python package spotpy (Houska et al., 2015) was used for automatic calibration of the model. Spotpy is a package that can be used for numerical optimization problems and includes some optimization algorithms often used in hydrological modelling. The DDS algorithm is the latest addition to the package and as far as the author knows, no documentation is available within the package about this specific algorithm. This made implementing DDS challenging and time-consuming, especially for a hydrologist not trained in object oriented programming. Parallization of the wflow hbv model within the spotpy DDS framework was needed because of the run-time of the model (and the model assessment scripts). For DDS, no documentation was available but it was found (by Niels Drost) that there was an option to run 1 model separately one every core available at the used computer, of which DDS was then able to use the results. Implementing this 'parallelization' was also time-consuming because wflow was not developed with this option in mind. A last problem within the automatic calibration setup was to choose the right number of model runs per calibration run. This issue was solved as is described in Appendix F. It should be noted however, that this method only works under the assumption that the parameter range given to DDS is both reasonable and small.

The third technical problem encountered in this project was the run-time of the model assessment scripts which is strongly related to the size of the output data. For streamflow, this was not an issue since output data consists of 12 timeseries which had to be assessed on several hydrological signatures (See subsubsection 3.3.2). However, the output files for the TWSA, AET and SM were several GB's in size (per model run!), because the output is on the model resolution (0.05 ° / daily). Part of this issue was solved by rescaling the model output to the resolution of the RS datasets (See Table 3), but for the AET and NDVI comparison, this made no difference. Also the determination of the T-value in the scenarios that used calibration with SM included using large files, scripts with long run-times and storing large output files. The assessment of all these model variables could therefore easily increase the model run-time from less than 10 minutes (calibration on Q-only) to over an hour (calibration on Q+TWSA+NDVI+SM), and was hence the determining factor for the run-time of the calibration runs.

A last point related to the technical issues that need to be made in this section is the use of the eWaterCycle environment. The author was allowed to use this platform, which is still under development, for this project and this made it possible to set up the the model calibration scenarios using a parallelized version of DDS. This would have been much more difficult without this platform. It was also possible to migrate the complete model setup to the Cartesius supercomputer, which was eventually not needed for this project, but which could be very useful in similar studies, especially when using more complex parameterizations and model setups. The eWaterCycle platform focuses on solving all of the technical problems encountered in this project, because these were used as input for the further development of the platform. Several people working on eWaterCycle also contributed a lot to this project, especially on the technical issues described in this section. eWaterCycle wants to make hydrologists able to use and compare different models, in different programming languages, and to make working with hydrological models easier. This project shows that this platform has great potential, because there is a lot to gain in the field of computational hydrology, especially in distributed hydrological modelling, but that there is also still a lot of work to be done. Recent developments however already show a great step forwards in the user friendliness of working with distributed hydrological models within eWaterCycle.

Model setup

Next to uncertainty in the data also some very important comments need to be made related to the setup of the model. The first one is that this study mainly focuses on optimizing the performance of a certain model for one or multiple datasets, but this is not the whole story of hydrological modelling. Building a model also includes choosing the right conceptualization or schematisation (Fenicia et al., 2008; Gharari et al., 2014) that contains the most important hydrological processes, as was done in Hulsman et al. (2021). The latter may be even more important to accurately describe the hydrological system because without including the most important natural processes, optimization of the model makes no sense. In this study the wflow hbv model was applied to the Volta basin, but possibly a better model structure exists that results in a better performance for the calibration and evaluation data. The goal of this study however was not to build the best hydrological model for the Volta basin but was related to the effect of including RS observations in the calibration and evaluation procedure on overall model performance.

In this study one single parameter set was used to model the hydrology of the entire Volta basin. This was done in order to prevent overparameterization of the model and to be able to accurately determine the best parameter set for each scenario. Model results can probably be improved by using separate parameter sets per subcatchment, or even distributed parameter maps, which is also possible in the wflow model setup. However, determining different parameter values for each cell may be difficult and methods to do so are still under development. The use of distributed forcing data is already a large step forwards compared to classical lumped hydrological models, as is underlined by the results of this study.

Next the to the model conceptualization and the model parameterization, also improvements can be made in the model classification in different land use, subcatchments and landscape or soil classes. The addition of soil classes and the extension of the number of land use classes may improve not only temporal but also spatial pattern representation. This is of course also strongly related to the model conceptualisation and parameterization.

A last point related to the model setup that should be made is that all results of this study are only valid for this specific model setup. The model setup in this context includes the hydrological model itself, the specific parameterization used in this study, all input data and the model calibration procedure. Changes to this model setup may alter the results found in this study. However, it is believed that the general conclusions drawn from this study are also valid for other DHM's and model setups, as they are also confirmed by other studies (Dembele, Hrachowitz, et al., 2020; Hulsman et al., 2021).

5.4 What did we learn?

In this last section of the discussion, the general lessons that can be taken from this study are identified. These lessons are based on the data, methodology, results and discussion, all provided in the previous chapters of this report. Several pure hydrological lessons can be learnt from this study, but also computationally, some insights were gained.

The first and most important point lesson is that the hydrological model setup in this study was able to simulate the temporal patterns of not only streamflow, but also of TWSA and SM quite accurately, but that there was a mismatch between the observed and simulated spatial patterns. Calibrating the same model setup using different combinations of streamflow and RS datasets did not change the simulated spatial pattern. This can only be due to the model setup being not flexible enough to allow for this spatial pattern representation. A point that was made several times in the discussion is that only 1 parameter set was used to model the hydrology of the whole Volta basin. This approach led to a trade-off effect for streamflow simulations for different stations. A parameter set may provide good results in one subcatchment, but much less good results in another. So, the model was also not able to simulate the spatial pattern for streamflow.

A possibility to cancel out this trade-off effect is to use 1 parameter set per subcatchment, and then separately calibrate all different subcatchments. However, this would hugely increase the model complexity and the amount of calibration work the modeller has to do. The extended version of this change in model setup would be to use distributed parameter maps, in which every cell has its own parameter set, but equifinality and computation power are problems that may impede the determination of so many parameter values. Other possibilities to increase the model flexibility and therefore allowing the model to more accurately simulate spatial patterns of hydrological variables are using other (more complex, e.g. including more hydrological processes) model conceptualisations and classifications of land use and soil.

Another lessen learnt from this study is that without sufficient model flexibility, the automatic calibration of a hydrological model using algorithms such as the dynamically dimensioned search algorithm, which was used in this study, only finetunes the free parameters and will not improve model performance very much compared to model calibration by hand. This is of course also strongly related to the model used. This lesson is also true for the calibration on multiple RS datasets, although the automation in this case does save the modeller a lot of calibration work. Because of the insufficient model flexibility, the calibration on RS datasets did not result in completely different temporal and spatial patterns of hydrological variables, but only the magnitude of these variables was changed. It is therefore again recommended to increase model flexibility when applying a calibration approach as is presented in this thesis.

The questions remains whether it is worth it to add all this RS data to the calibration of a hydrological model. In this specific case, the calibration on TWSA and SM was very similar for all scenarios, and could better have been used for model evaluation only. However, considering a much more complex model, which includes many more hydrological processes, land use and soil classifications and therefore parameters (think order 100+, which could be reasonable for the whole Volta basin), the parameter space gets so large that the model cannot be calibrated by only constraining it to streamflow, because the modeller has to deal with massive equifinality problems. At some point the model gets so complex, that additional data is needed to constrain the parameter space and then, it is definitely worth to include RS datasets in the calibration.

A last hydrological insight that was gained in this study is that it is not possible to directly use NDVI data to assess the simulated evaporation flux. Although it can be used to provide a rough estimate to when and where evaporation should be high or low, the physical relation between the two is too weak and cannot be used to calibrate a hydrological model on. Research into developing methods that provide evaporation estimates based on NDVI data may be needed, although there are probably easier methods available to determine evaporation with RS data.

Not only hydrological insights, but also computational insights were gained in this study. The most important one found is that the run-time of a model may not be the most important factor in determining the time needed for (automatic) calibration. When applying RS data in the calibration or evaluation of a large distributed hydrological model, it was found that the post-processing (storing the output and assessing its performance) may take much longer than the actual time needed to run the model. It is very important to consider this when choosing a model resolution. It may be a lot of work to built your model again on a different resolution, which may be needed to decrease the computational time needed for post-processing the model output.

To decrease computation time, it is recommended to really investigate what model resolution is actually needed to achieve a satisfactory model performance, instead of what best fits the forcing data used. It is also recommended to use models that are programmed in up-to-date fast programming languages (for wflow this is now Julia), and not in outdated languages like PCRaster. It is also important to store model output at the resolution at which it is needed, and not at higher resolutions. Parallelization may be an option to speed up a calibration run, but this could also be a lot of work. To decrease the time needed for model calibration using algorithms like DDS, it may be good to investigate after how many runs the model performance does not increase significantly anymore, like was done in Appendix F.

6 Conclusions & Recommendations

Conclusions

The results from this study indicate that a quite good average streamflow performance, and a good representation of the temporal patterns of TWSA and SM can be achieved, using a relatively simple model setup and an extensive calibration procedure on Q-only. However, the differences in streamflow performance between the different gauging stations are large and also the spatial pattern representation of TWSA and SM was found to be inadequate. Adding TWSA and SM data to the calibration procedure did not change the spatial and temporal patterns of streamflow, TWSA and SM, but did change the magnitude of the different variables. Generally, that meant a decrease in streamflow performance, which was larger for the addition of SM data to the calibration than for the addition of TWSA data, but an increase in temporal and spatial pattern representation, although the differences are relatively small. So, a trade-off effect was observed. It was also concluded that the physical relation between NDVI and AET was not strong enough to be used 1-to-1 for model calibration, even though only pattern information, and no absolute values were used. The method used to translate SSM observations in representative values for the complete unsaturated zone using a determined T-value did provide good results.

It was also found that for this specific model setup, the calibration using DDS and a set of hydrological signatures and spatial and temporal patterns was able to increase the overall model performance only minimally, most likely due to a too low level of model flexibility. Thus, to answer the research question stated in the introduction of this report; The inclusion of the spatio-temporal patterns of RS data products in the calibration of a distributed conceptual hydrological model only minimally improves the overall model performance in the calibration catchments, and does not improve model performance in the evaluation catchments. For a different model setup however, the results may be different.

Recommendations

Based on the results of this study in the field of RS data usage in the calibration of DHM's, several recommendations can be made for future research. The most important recommendation is to repeat this study with a more complex model setup. This could mean including more land use classes and introduce soil classes, comparing several, possibly more extensive, model conceptualisations, but also using 1 parameter set per subcatchment or even using distributed parameter maps. This increases model flexibility and freedom and it is hypothesized that this may have a positive effect on the space given to the model for temporal and spatial pattern optimization, and that extra observations (such as the RS datasets used in this study) at some point in model flexibility become essential to constrain the parameter space. Another more computational hydrological recommendation is to do research on what model resolution is actually needed to improve model performance and to assess model output with observations, because this is the key factor in a lot of computational issues encountered in this project.

References

Akbar, T., Hassan, Q., Ishaq, S., Batool, M., Butt, H., & Jabbar, H. (2019). Investigative spatial distribution and modelling of existing and future urban land changes and its impact on urbanization and economy. *Remote Sensing*, *11*, 105. doi: https://doi.org/10.3390/rs11020105

Becker, R., Koppa, A., Schulz, S., Usman, M., Beek, T. A. D., & Schueth, C. (2019). Spatially distributed model calibration of a highly managed hydrological system using remote sensing-derived ET data. *Journal of Hydrology*, *577*. doi: https://doi.org/10.1016/j.jhydrol.2019.123944

Bergström, S. (1992). The hbv-model—its structure and applications. *SMHI Reports RH No. 4, Norrköping.*

Beven, K. (1993). Prophecy, reality and uncertainty in distributed hydrological modeling. *Advances in Water Resources*, *16*(1), 41-51. doi: https://doi.org/10.1016/0309-1708(93)90028-e

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, *320*(1), 18 - 36. doi: https://doi.org/10.1016/j.jhydrol.2005.07.007

Beven, K., Karsten, S., & Franks, S. (1999). Functional similarity in hydrological modelling at the landscape scale. *J. Feyen und K. Wiyo, Modelling of transport processes in soils, Wageningen Press, Wageningen, Netherlands*, 725 - 735.

Bontemps, S., Defourny, P., Bogaert, E., Arino, O., Kalogirou, V., & Perez, J. (2011). *GLOBCOVER 2009 — products description and validation report.* Retrieved from https://due.esrin.esa.int/files/GLOBCOVER2009_Validation_Report_2.2.pdf

Bouaziz, L. J. E., Fenicia, F., Thirel, G., de Boer-Euser, T., Buitink, J., Brauer, C. C., ... Hrachowitz, M. (2021). Behind the scenes of streamflow model performance. *Hydrology and Earth System Sciences*, *25*(2), 1069–1095. doi: https://doi.org/10.5194/hess-25-1069-2021

Bouaziz, L. J. E., Steele-Dunne, S. C., Schellekens, J., Weerts, A. H., Stam, J., Sprokkereef, E., ... Hrachowitz, M. (2020). Improved understanding of the link between catchment-scale vegetation accessible storage and satellite-derived soil water index. *Water Resources Research*, *56*(3). doi: https://doi.org/10.1029/2019wr026365

Boyer, J. F., Dieulin, C., Rouche, N., Cres, A., Servat, E., Paturel, J. E., & Mahe, G. (2006). SIEREM: an environmental information system for water resources. In Demuth, S., Gustard, A., Planos, E., Scatena, F. and Servat, E. (Ed.), *Climate Variability and Change - Hydrological Impacts* (Vol. 308, p. 19). (5th FRIEND World Conference, Havana, Cuba, Nov., 2006)

Chen, M., Parton, W. J., Hartman, M. D., Del Grosso, S. J., Smith, W. K., Knapp, A. K., ... Gao, W. (2019). Assessing precipitation, evapotranspiration, and NDVI as controls of U.S. Great Plains plant production. *Ecosphere*, *10*(10). doi: https://doi.org/10.1002/ecs2.2889

Chow, T. (1959). *Open-Channel Hydraulics.*

Cihlar, J., St.-Laurent, L., & Dyer, J. (1991). Relation between the normalized difference vegetation index and ecological variables. *Remote Sensing of Environment*, *35*(2), 279-298. doi: https://doi.org/10.1016/0034-4257(91)90018-2

Danielson, J., & Gesch, D. (2011). Global multi-resolution terrain elevation data 2010 (GMTED2010). *U.S. Geological Survey*, 26. doi: https://doi.org/10.3133/ofr201110738

de Jeu, R. (2003). Retrieval of Land Surface Parameters using Passive Microwave Remote Sensing. *Vrije Universiteit, Amsterdam.*

Dembele, M., Hrachowitz, M., Savenije, H. H. G., Mariethoz, G., & Schaefli, B. (2020). Improving the predictive skill of a distributed hydrological model by calibration on spatial patterns with multiple satellite data sets. *Water Resources Research*, *56*(1). doi: https://doi.org/10.1029/2019wr026085

Dembele, M., Oriani, F., Tumbulto, J., Mariethoz, G., & Schaefli, B. (2019). Gap-filling of daily streamflow time series using direct sampling in various hydroclimatic settings. *Journal of Hydrology*, *569*, 573-586. doi: https://doi.org/10.1016/j.jhydrol.2018.11.076

Dembele, M., & Zwart, S. J. (2016). Evaluation and comparison of satellite-based rainfall products in Burkina Faso, West Africa. *International Journal of Remote Sensing*, *37*(17), 3995-4014. doi: https://doi.org/10.1080/01431161.2016.1207258

Dembele, M., Zwart, S. J., Ceperley, N., Salvadore, E., Mariethoz, G., & Schaefli, B. (2020). Potential of satellite and reanalysis evaporation datasets for hydrological modelling under various model calibration strategies. *Advances in Water Resources*, *143*. doi: https://doi.org/10.1016/j.advwatres.2020.103667

Demirel, M. C., Koch, J., Mendiguren, G., & Stisen, S. (2018). Spatial Pattern Oriented Multicriteria Sensitivity Analysis of a Distributed Hydrologic Model. *Water*, *10*(9). doi: https://doi.org/10.3390/w10091188

Demirel, M. C., Mai, J., Mendiguren, G., Koch, J., Samaniego, L., & Stisen, S. (2018). Combining satellite data and appropriate objective functions for improved spatial pattern performance of a distributed hydrologic model. *Hydrology and Earth System Sciences*, *22*(2), 1299-1315. doi: https://doi.org/10.5194/hess-22-1299-2018

Demirel, M. C., Ozen, A., Orta, S., Toker, E., Demir, H. K., Ekmekcioglu, O., ... Booij, M. J. (2019). Additional Value of Using Satellite-Based Soil Moisture and Two Sources of Groundwater Data for Hydrological Model Calibration. *Water*, *11*(10). doi: https://doi.org/10.3390/w11102083

Dezetter, A., & Ruelland, D. (2012). Parameterization based on NOAA-AVHRR NDVI to improve conceptual rainfall-runoff modelling in a large West African catchment. In *Remote Sensing and Hydrology* (Vol. 352, p. 221). IAHS.

Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., ... Lecomte, P. (2017). ESA CCI Soil Moisture for improved Earth system understanding: State-of-the art and future directions. *Remote Sensing of Environment*, *203*, 185-215. doi: https://doi.org/10.1016/j.rse.2017.07.001

Duan, Q., Gupta, V., & Sorooshian, S. (1993). A Shuffled Complex Evolution Approach for Effective and Efficient Global Minimization. *Journal of Optimization Theory and Applications*, *76*, 501-521. doi: https://doi.org/10.1007/BF00939380

Estebanez Camarena, M., van de Giesen, N., ten Veldhuis, M.-C., & de Vries, S. (2020, May). RainRunner - Machine Learning and Earth observation for reliable rainfall information in West Africa. In *EGU General Assembly Conference Abstracts* (p. 22073).

Fenicia, F., Savenije, H. H. G., Matgen, P., & Pfister, L. (2008). Understanding catchment behavior through stepwise model concept improvement. *Water Resources Research*, *44*(1). doi: https://doi.org/10.1029/2006WR005563

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., ... Michaelsen, J. (2015). The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes [Journal Article]. *Scientific Data*, *2*(1), 150066. doi: https://doi.org/10.1038/sdata.2015.66

Gharari, S., Hrachowitz, M., Fenicia, F., Gao, H., & Savenije, H. H. G. (2014). Using expert knowledge to increase realism in environmental system models can dramatically reduce the need for calibration. *Hydrology and Earth System Sciences*, *18*(12), 4839-4859. doi: https://doi.org/10.5194/hess-18-4839-2014

GRDC. (2021). *The global runoff data centre.* (56068, Koblenz, Germany)

Gruber, A., Dorigo, W. A., Crow, W., & Wagner, W. (2017). Triple Collocation-Based Merging of Satellite Soil Moisture Retrievals. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(12), 6780-6792. doi: https://doi.org/10.1109/TGRS.2017.2734070

Gruber, A., Scanlon, T., van der Schalie, R., Wagner, W., & Dorigo, W. (2019). Evolution of the ESA CCI Soil Moisture climate data records and their underlying merging methodology. *Earth System Science Data*, *11*(2), 717–739. doi: https://doi.org/10.5194/essd-11-717-2019

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1), 80-91. doi: https://doi.org/10.1016/j.jhydrol.2009.08.003

Hargreaves, G. H., & Samani, Z. H. (1985). Reference Crop Evapotranspiration from Temperature. *Applied Engineering in Agriculture*, *1*(2), 96-99. doi: https://doi.org/10.13031/2013.26773

Houska, T., Kraft, P., Chamorro-Chavez, A., & Breuer, L. (2015). Spotting model parameters using a ready-made python package. *PLOS ONE*, *10*(12), 1-22. doi: https://doi.org/10.1371/journal.pone.0145180

Hrachowitz, M., & Clark, M. P. (2017). HESS Opinions: The complementary merits of competing modelling philosophies in hydrology. *Hydrology and Earth System Sciences*, *21*(8), 3953-3973. doi: https://doi.org/10.5194/hess-21-3953-2017

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., ... Gascuel-Odoux, C. (2014). Process consistency in models: The importance of system signatures, expert knowledge, and process complexity. *Water Resources Research*, *50*(9), 7445-7469. doi: https://doi.org/10.1002/2014wr015484

Hrachowitz, M., Savenije, H. H. G., Bloschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., ... Cudennec, C. (2013). A decade of Predictions in Ungauged Basins (PUB)a review. *Hydrological Sciences Journal - Journal Des Sciences Hydrologiques*, *58*(6), 1198-1255. doi: https://doi.org/10.1080/02626667.2013.803183

Hulsman, P., Savenije, H. H. G., & Hrachowitz, M. (2021). Learning from satellite observations: increased understanding of catchment processes through stepwise model improvement. *Hydrology and Earth System Sciences*, *25*(2), 957–982. doi: https://doi.org/10.5194/hess-25-957-2021

Hulsman, P., Winsemius, H. C., Michailovsky, C. I., Savenije, H. H. G., & Hrachowitz, M. (2020). Using altimetry observations combined with GRACE to select parameter sets of a hydrological model in a data-scarce region. *Hydrology and Earth System Sciences*, *24*(6), 3331–3359. doi: https://doi.org/10.5194/hess-24-3331-2020

Huot, P.-L., Poulin, A., Audet, C., & Alarie, S. (2019). A hybrid optimization approach for efficient calibration of computationally intensive hydrological models. *Hydrological Sciences Journal - Journal des Sciences Hydrologiques*, *64*(10), 1204-1222. doi: https://doi.org/10.1080/02626667.2019.1624922

Islam, M. M., & Mamun, M. M. I. (2015). Variations of NDVI and Its Association with Rainfall and Evapotranspiration over Bangladesh. *Rajshahi University Journal of Science & Engineering*, *43*, 21-28. doi: https://doi.org/10.3329/rujse.v43i0.26160

Kerr, Y. H., Imbernon, J., Dedieu, G., Hautecoeur, O., Lagouarde, J. P., & Seguin, B. (1989). NOAA AVHRR and its uses for rainfall and evapotranspiration monitoring. *International Journal of Remote Sensing*, *10*(4-5), 847-854. doi: https://doi.org/10.1080/01431168908903925

Kirchner, J. (2006). Getting the Right Answers for the Right Reasons: Linking Measurements, Analyses, and Models to Advance the Science of Hydrology. *Water Resources Research*, *42*. doi: https://doi.org/10.1029/2005WR004362

Kling, H., Fuchs, M., & Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, *424-425*, 264-277. doi: https://doi.org/10.1016/j.jhydrol.2012.01.011

Knoben, W., Freer, J., & Woods, R. (2019, 07). Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores. *Hydrology and Earth System Sciences Discussions*, 1-7. doi: https://doi.org/10.5194/hess-2019-327

Koch, J., Demirel, M. C., & Stisen, S. (2018). The SPAtial EFficiency metric (SPAEF): multiple-component evaluation of spatial patterns for optimization of hydrological models. *Geoscientific Model Development*, *11*(5), 1873-1886. doi: https://doi.org/10.5194/gmd-11-1873-2018

Koch, J., Mendiguren, G., Mariethoz, G., & Stisen, S. (2017). Spatial Sensitivity Analysis of Simulated Land Surface Patterns in a Catchment Model Using a Set of Innovative Spatial Performance Metrics. *Journal of Hydrometeorology*, *18*(4), 1121-1142. doi: https://doi.org/10.1175/jhm-d-16-0148.1

Kumar, R., Samaniego, L., & Attinger, S. (2013). Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations. *Water Resources Research*, *49*(1), 360-379. doi: https://doi.org/10.1029/2012wr012195

Kummu, M., de Moel, H., Ward, P. J., & Varis, O. (2011). How Close Do We Live to Water? A Global Analysis of Population Distance to Freshwater Bodies. *Plos One*, *6*(6), 1-13. doi: https://doi.org/10.1371/journal.pone.0020578

Landerer, F. W., & Swenson, S. C. (2012). Accuracy of scaled GRACE terrestrial water storage estimates. *Water Resources Research*, *48*(4). doi: https://doi.org/10.1029/2011WR011453

Li, K., Coe, M., Ramankutty, N., & Jong, R. D. (2007). Modeling the hydrological impact of land-use change in West Africa. *Journal of Hydrology*, *337*(3), 258 - 268. doi: https://doi.org/10.1016/j.jhydrol.2007.01.038

Lindström, G., Johansson, B., Persson, M., Gardelin, M., & Bergström, S. (1997). Development and test of the distributed hbv-96 hydrological model. *Journal of hydrology*, *201*(1-4), 272–288.

Makkink, G. F. (1957). Ekzameno de la formula de Penman. *Netherlands J. Agric. Sci.*, *5*, 290-305.

Mendiguren, G., Koch, J., & Stisen, S. (2017). Spatial pattern evaluation of a calibrated national hydrological model - a remote-sensing-based diagnostic approach. *Hydrology and Earth System Sciences*, *21*(12), 5987-6005. doi: https://doi.org/10.5194/hess-21-5987-2017

Miralles, D. G., Holmes, T. R. H., De Jeu, R. A. M., Gash, J. H., Meesters, A. G. C. A., & Dolman, A. J. (2011). Global land-surface evaporation estimated from satellite-based observations. *Hydrology and Earth System Sciences*, *15*(2), 453–469. doi: https://doi.org/10.5194/hess-15-453-2011

Monteith, J. L. (1965). Evaporation and Environment. *Symp. Soc. Exp. Biol.*, *19*, 205-34.

Nash, J., & Sutcliffe, J. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, *10*(3), 282-290. doi: https://doi.org/10.1016/0022-1694(70)90255-6

Nijzink, R. C., Almeida, S., Pechlivanidis, I. G., Capell, R., Gustafssons, D., Arheimer, B., ... Hrachowitz, M. (2018). Constraining conceptual hydrological models with multiple information sources. *Water Resources Research*, *54*(10), 8332-8362. doi: https://doi.org/10.1029/2017wr021895

Padi, M. (2018). Food Crop Farming and the Climate in Southern Ghana. *Acta Scientific Agriculture*, *2*(7), 23-26.

Priestley, C. H. B., & Taylor, R. J. (1972). On the assessment of surface heat flux and evaporation using large scale parameters. *Monthly Weather Rev.*, *100*, 81-92.

Rakovec, O., Kumar, R., Attinger, S., & Samaniego, L. (2016). Improving the realism of hydrologic model functioning through multivariate parameter estimation. *Water Resources Research*, *52*(10), 7779-7792. doi: https://doi.org/10.1002/2016wr019430

Rakovec, O., Kumar, R., Mai, J., Cuntz, M., Thober, S., Zink, M., ... Samaniego, L. (2016). Multiscale and Multivariate Evaluation of Water Fluxes and States over European River Basins. *Journal of Hydrometeorology*, *17*(1), 287-307. doi: https://doi.org/10.1175/jhm-d-15-0054.1

Ramirez, J. A. (2012). *CIEV22 Basic Hydrology* (No. 7 Homework). Retrieved from `https://www.engr.colostate.edu/~ramirez/ce_old/classes/cive322-Ramirez/Homework_7_16_Sln-1.pdf`

Rennó, C., Nobre, A., Cuartas, L., Soares, J., Hodnett, M., Tomasella, J., & Waterloo, M. (2008). HAND, a new terrain descriptor using SRTM-DEM: Mapping terra-firme rainforest environments in Amazonia. *Remote Sensing of Environment*, *112*, 3469-3481. doi: https://doi.org/10.1016/j.rse.2008.03.018

Sakumura, C., Bettadpur, S., & Bruinsma, S. (2014). Ensemble prediction and intercomparison analysis of GRACE time-variable gravity field models. *Geophysical Research Letters*, *41*(5), 1389-1397. doi: https://doi.org/10.1002/2013GL058632

Santos, L., Thirel, G., & Perrin, C. (2018). Technical note: Pitfalls in using log-transformed flows within the KGE criterion. *Hydrology and Earth System Sciences*, *22*(8), 4583–4591. doi: https://doi.org/10.5194/hess-22-4583-2018

Savenije, H. H. G. (2001). Equifinality, a blessing in disguise? *Hydrological Processes*, *15*(14), 2835-2838. doi: https://doi.org/https://doi.org/10.1002/hyp.494

Savenije, H. H. G. (2004). The importance of interception and why we should delete the term evapotranspiration from our vocabulary. *Hydrological Processes*, *18*(8), 1507-1511. doi: https://doi.org/10.1002/hyp.5563

Savenije, H. H. G. (2010). HESS Opinions "Topography driven conceptual modelling (FLEX-Topo)". *Hydrology and Earth System Sciences*, *14*(12), 2681-2692. doi: https://doi.org/10.5194/hess-14-2681-2010

Schellekens, J. (2021). wflow documentation. *Deltares*. Retrieved 2021-04-06, from `https://wflow.readthedocs.io/en/latest/index.html`

Seevers, P. M., & Ottoman, R. W. (1994). Evapotranspiration estimation using a normalized difference vegetation index transformation of satellite data. *Hydrological Sciences Journal*, *39*(4), 333-345. doi: https://doi.org/10.1080/02626669409492754

Sperna Weiland, F., Lopez, P., van Dijk, A., & Schellekens, J. (2015). Global high-resolution reference potential evaporation. *21st International Congress on Modelling and Simulation.*

Stisen, S., Koch, J., Sonnenborg, T. O., Refsgaard, J. C., Bircher, S., Ringgaard, R., & Jensen, K. H. (2018). Moving beyond run-off calibration-multivariable optimization of a surface-subsurface-atmosphere model. *Hydrological Processes*, *32*(17), 2654-2668. doi: https://doi.org/10.1002/hyp.13177

Stisen, S., McCabe, M. F., Refsgaard, J. C., Lerer, S., & Butts, M. B. (2011). Model parameter analysis using remotely sensed pattern information in a multi-constraint framework. *Journal of Hydrology*, *409*(1-2), 337-349. doi: https://doi.org/10.1016/j.jhydrol.2011.08.030

Stisen, S., & Sandholt, I. (2010). Evaluation of remote-sensing-based rainfall products through predictive capability in hydrological runoff modelling. *Hydrological Processes*, *24*(7), 879-891. doi: https://doi.org/10.1002/hyp.7529

Strahler, A. N. (1952, 11). Hypsometric (area-altitude) analysis of erosional topography. *GSA Bulletin*, *63*(11), 1117-1142. doi: https://doi.org/10.1130/0016-7606(1952)63[1117:HAAOET]2.0.CO;2

Swenson, S. (2012). GRACE monthly land water mass grids NETCDF RELEASE 5.0. Ver. 5.0. *PO.DAAC, CA, USA.*. doi: https://doi.org/10.5067/TELND-NC005

Swenson, S., & Wahr, J. (2006). Post-processing removal of correlated errors in GRACE data. *Geophysical Research Letters*, *33*(8). doi: https://doi.org/10.1029/2005GL025285

Szilagyi, J., Rundquist, D., Gosselin, D., & Parlange, M. (1998, 05). NDVI Relationship to Monthly Evaporation. *Geophysical Research Letters*, *25*. doi: https://doi.org/10.1029/98GL01176

Tangdamrongsub, N., Steele-Dunne, S. C., Gunter, B. C., Ditmar, P. G., & Weerts, A. H. (2015). Data assimilation of GRACE terrestrial water storage estimates into a regional hydrological model of the Rhine River basin. *Hydrology and Earth System Sciences*, *19*(4), 2079-2100. doi: https://doi.org/10.5194/hess-19-2079-2015

Tolson, B., & Shoemaker, C. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, *43*. doi: https://doi.org/10.1029/2005WR004723

van de Giesen, N., Andreini, M., Edig, A., & Vlek, P. (2001). Competition for water resources of the volta basin. *IAHS-AISH Publication*, *268*.

van Zwieten, P., Béné, C., Kolding, J., Brummett, R., & Valbo-Jørgensen, J. (2011). *Review of tropical reservoirs and their fisheries – The cases of Lake Nasser, Lake Volta and Indo-Gangetic Basin reservoirs.* (Tech. Rep. No. 557). Viale delle Terme di Caracalla, 00153 Roma RM, Italy: FAO Fisheries and Aquaculture.

Vermote, E. (2019). Noaa cdr program. *NOAA Climate Data Record (CDR) of AVHRR Normalized Difference Vegetation Index (NDVI), Version 5*, *NOAA National Centers for Environmental Information*. doi: https://doi.org/10.7289/V5ZG6QH9

Vázquez, R. F., Willems, P., & Feyen, J. (2008). Improving the predictions of a MIKE SHE catchment-scale application by using a multi-criteria approach. *Hydrological Processes*, *22*(13), 2159-2179. doi: https://doi.org/10.1002/hyp.6815

Wang, K., Wang, P., Li, Z., Cribb, M., & Sparrow, M. (2007). A simple method to estimate actual evapotranspiration from a combination of net radiation, vegetation index, and temperature. *Journal of Geophysical Research*, *112*. doi: https://doi.org/10.1029/2006JD008351

Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., & Viterbo, P. (2014). The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data. *Water Resources Research*, *50*(9), 7505-7514. doi: https://doi.org/10.1002/2014WR015638

Wetterhall, F. (2014). HBV – the most famous hydrological model of all? An interview with its father: Sten Bergström. *Hepex*. Retrieved 2021-02-05, from `https://hepex.inrae.fr/the-hbv-model-40-years-and-counting/`

Zink, M., Mai, J., Cuntz, M., & Samaniego, L. (2018). Conditioning a hydrologic model using patterns of remotely sensed land surface temperature. *Water Resources Research*, *54*(4), 2976-2998. doi: https://doi.org/10.1002/2017wr021346

Appendices

A Brief history and applications of the hbv model

The earliest version of the hbv model was developed by Sten Bergström in 1972 (Wetterhall, 2014) at the Swedish Meteorological and Hydrological Institute (SMHI). Nowadays, it is one of the most used hydrological models in the world. The model is used in 95 countries and in more than 1400 publications (Wetterhall, 2014). One of the first use cases of the model was to make forecasts of water inflow for the hydropower industry, also a possible application of the model developed in this study. According to Sten Bergström, the success of the hbv model lies in its balance between model complexity and simplicity. In Figure 22, all the countries in the world where the hbv model is applied according to a list composed by Sten Bergström himself are shown.

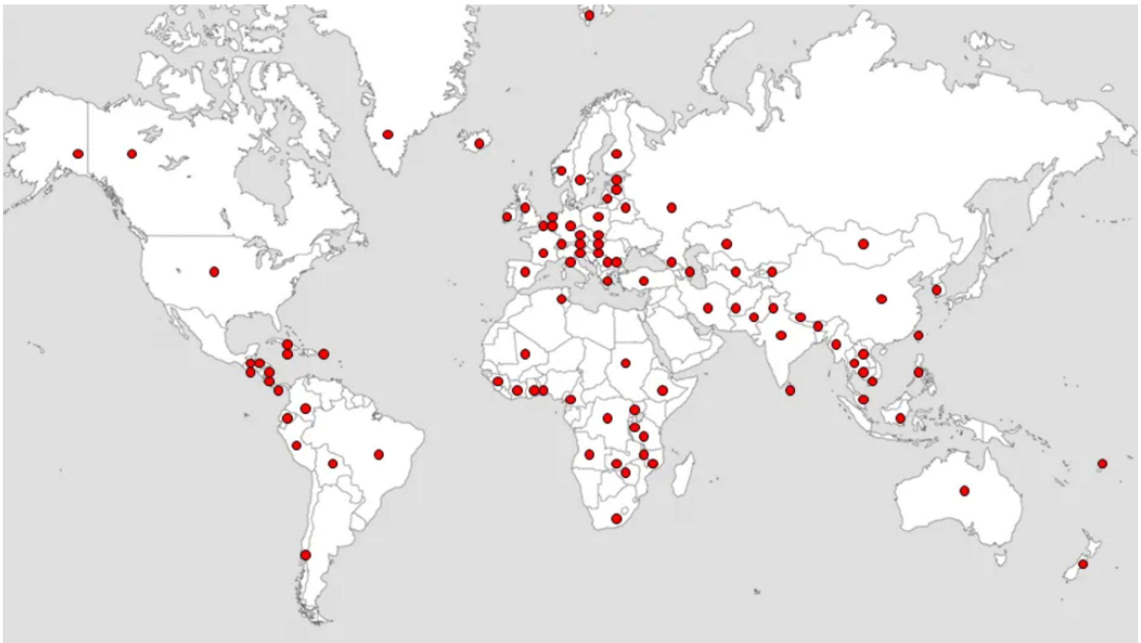


Figure 22: Map of countries where the hbv model is applied, made by the father of hbv, Sten Bergström. Taken from Wetterhall (2014).

B CHIRPS precipitation comparison with rain gauge data

In this appendix, the CHIRPS precipitation observations are compared to rain gauge observations in a point-to-pixel analysis. Rain gauge data in northern Ghana that was used in Estebanez Camarena et al. (2020) was made available by Mónica Estebanez Camarena for this analysis. The locations of the rain gauges are shown in Figure 23. It can be seen that the rain gauges are not spread evenly across the basin. However, since this was the only precipitation data which was available for the model period, the analysis was carried out with this spatially biased data.

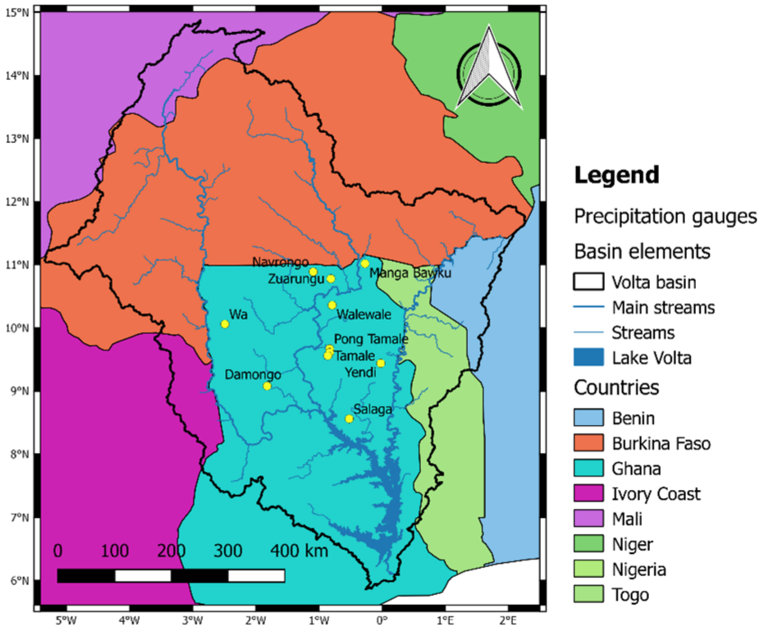


Figure 23: Map showing the locations of the precipitation gauges used in this analysis. (The station Savelugu is in between the stations Pong Tamale and Tamale).

The point-to-pixel analysis was carried out on a daily basis, since this is also the resolution on which the data is used in the model. The results of this analysis are shown in Figure 24 for every precipitation station. It can be seen that the observed best linear fit (green line) is below the ideal relation (orange line), in which the CHIRPS and rain gauge observations are exactly equal, for every station. This means that the rain gauges generally observe higher precipitation than the CHIRPS dataset does. Other results from the comparison of the data are that on 15% of the days, CHIRPS records rainfall while the rain gauges do not measure any rainfall, and on 7% of the days, the rain gauges record rainfall while CHIRPS does not record any at all. CHIRPS measures precipitation on 29% of the days in the model period, and the rain gauges do so on 22% of the days. The mean daily rainfall (on days on which rainfall is observed by the specific dataset) is 9.3 mm for CHIRPS, and 12.7 mm for the rain gauges. The fraction of the total amount of rain observed in the model period by CHIRPS over the total amount of rain observed in the model period by the rain gauges (which ideally would be 1) is between 0.9 and 1.07 for every station, with a mean value of 0.99. So the total amount of precipitation observed by the gauges and the CHIRPS dataset in the full model period is very similar, but the timing and the amount of precipitation on a daily basis is not. This difference also comes back in the R^2-correlation value for every station, which is shown in Figure 24. All R^2-values are between 0.09 and 0.19, with a mean value of 0.12.

The results from this analysis are confirmed by Estebanez Camarena et al. (2020) and can partly be explained by the different observation techniques used. Rain gauges measure precipitation on a point-scale and are therefore only representative for a very small area. Observations from satellites however, are performed on a much larger scale, and the precipitation observations are put in a grid as output to be used in models like the one developed in this project. Therefore, the precipitation observed in an area is averaged over the complete grid cell. This means that precipitation amounts can expected to be lower in RS datasets, and that more precipitation events will be recorded compared to rain gauge data. Exactly these 2 differences are observed for the CHIRPS and rain gauge data in this analysis. RS datasets are often also calibrated to perform well for rainfall amounts, and therefore the rainfall amount over longer periods (such as the modelled period of more than 7 years) are very comparable in this analysis. However, still on 7% of the days the rain gauges observe rainfall while CHIRPS completely misses out on these events, and the correlation on a daily basis is poor. This is a general problem for RS precipitation datasets in West-Africa (Estebanez Camarena et al., 2020). It is shown in this analysis that the CHIRPS dataset can be used because the total amount of precipitation input is comparable to station data, but the data should be used with caution, especially on a daily timescale.

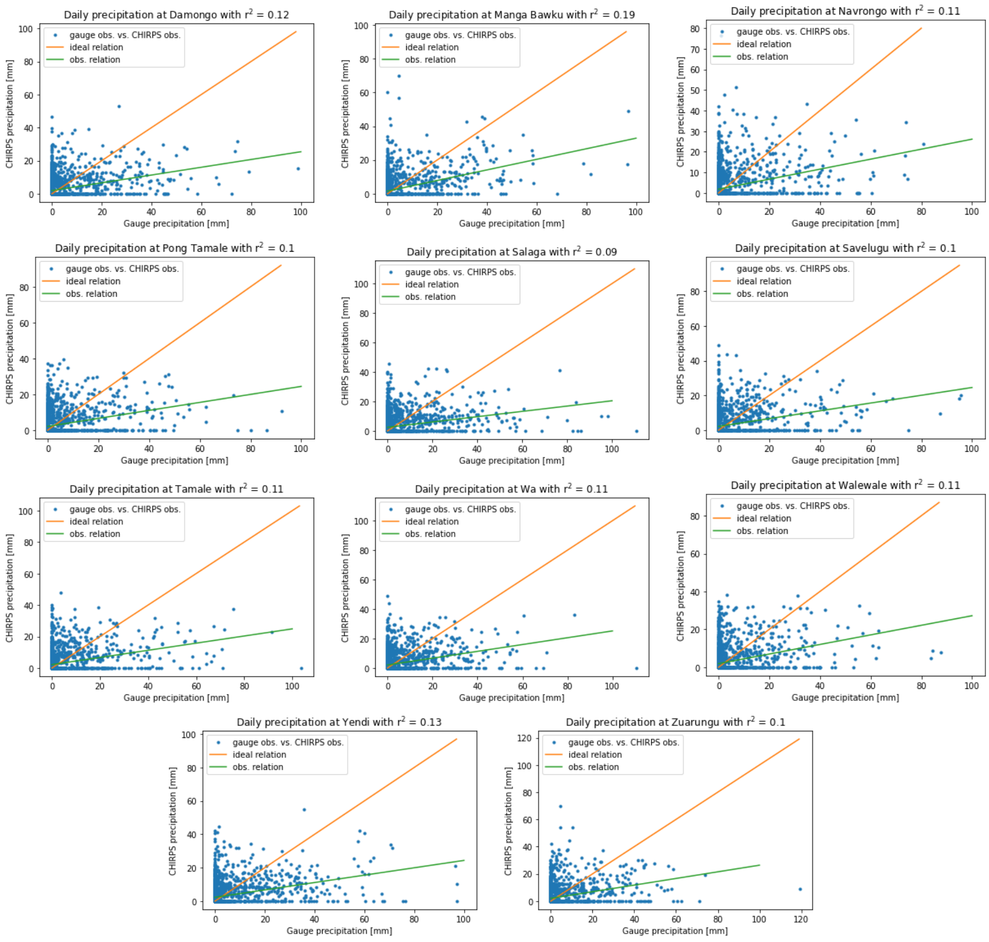


Figure 24: Point-to-pixel precipitation analysis plots of all precipitations gauge and CHIRPS observations.

C Spatio-temporal patterns of RS data

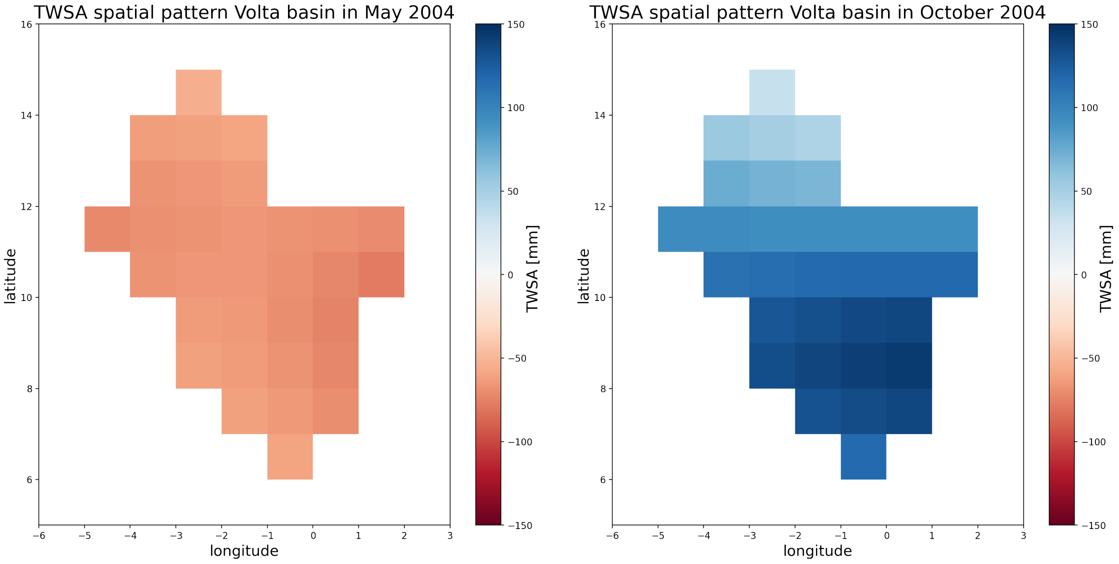


Figure 25: Spatial pattern of TWSA in the Volta basin in May 2004 (left) and October 2004 (right)

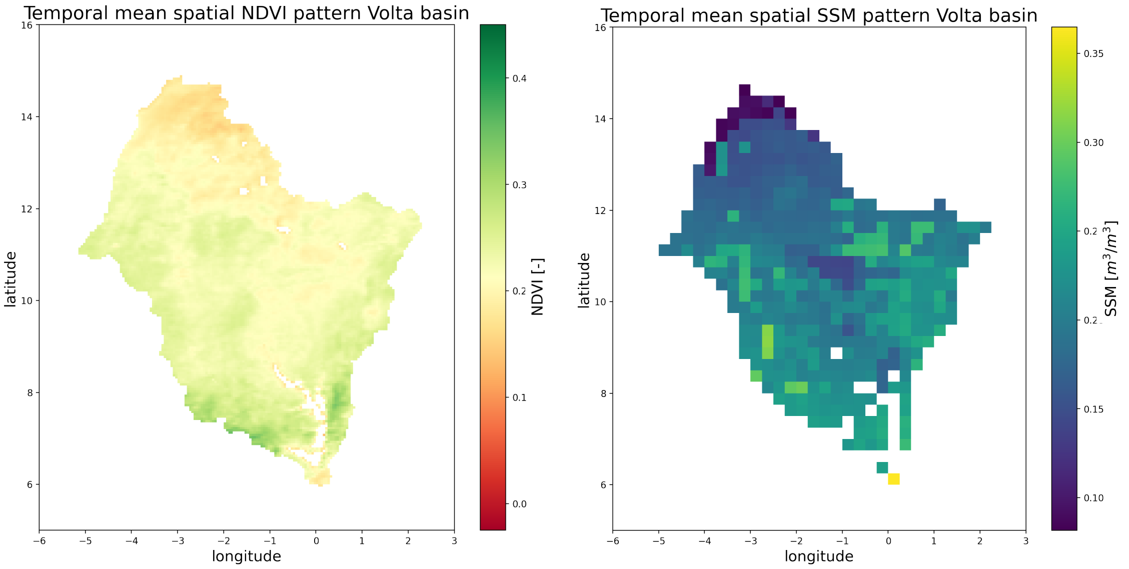


Figure 26: Mean spatial patterns of NDVI (left) and SM (right) in the Volta basin in the model period

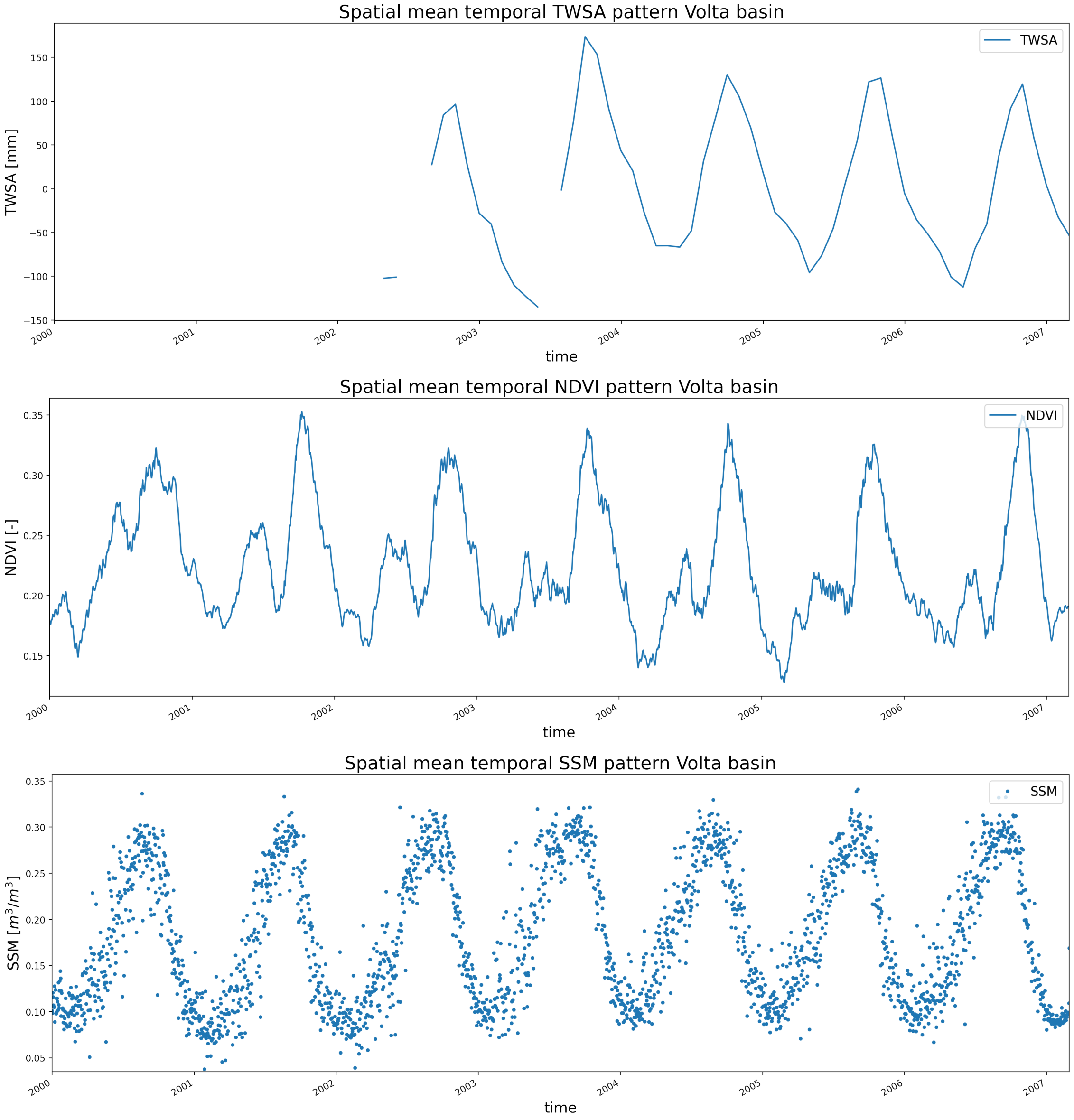


Figure 27: Mean temporal patterns of TWSA (top), NDVI (middle) and SSM (bottom) in the Volta basin in the model period

D Water balance method used for the determination of T and $S_{u,max}$

In this appendix, the method used to determine the characteristic time length T and the depth of the unsaturated zone $S_{u,max}$, which is directly used as the depth of the soil moisture reservoir (FC) in the wflow hbv model, is explained. The method was developed by and taken from Bouaziz et al. (2020).

Determining $S_{u,max}$

First, the mean annual precipitation $\overline{P}$ per subcatchment was determined for the 10-year period from 1997 up to and including 2006 in mm/year. Then, the timeseries of precipitation and PET from each cell were taken, and the interception evaporation for each cell was determined assuming interception capacities of 0.5, 1.0, 2.0 and 3.0 mm. Then also the mean annual $\overline{PET}$ and $\overline{E_i}$ were determined from this analysis per subcatchment. The streamflow observations were used to determine to mean annual runoff $\overline{Q_{river}}$ in mm/year too, which allowed for an estimation of the annual mean actual evaporation $\overline{E_a}$ in mm/year from the water balance given in Equation 16, because the mean annual groundwater loss $\overline{Q_{gw,loss}}$ was estimated to be negligible in the annual water balance (Bouaziz et al., 2020). This annual mean actual evaporation was then scaled to timeseries of daily values per subcatchment using Equation 17. Then, an initial storage deficit of zero is assumed at the end of the wet period, which was defined from April up to and including October, leaving the months November till March for the dry period. The annual maximum storage deficit per subcatchment was then determined from the daily storage deficit timeseries using Equation 18, with T_0 being the start of the dry season and T_1 being the start of the next dry season. In this way, 9 annual maximum storage deficits were identified for every subcatchment, because only 9 complete dry seasons were present in the 10-year period. An example of such a storage deficit timeseries is given in Figure 28.

$$\frac{dS}{dt} \approx 0 \approx \overline{P} - \overline{E_i} - \overline{E_a} - \overline{Q_{river}} - \overline{Q_{gw,loss}} \tag{16}$$

$$E_a(t) = \Big(PET(t) - E_i(t)\Big) \cdot \frac{\overline{E_a}}{(\overline{PET} - \overline{E_i})} \tag{17}$$

$$S_{u,max}(t) = min \int_{T_0}^{T_1} P(t) - E_i(t) - E_a(t)\, dt \tag{18}$$

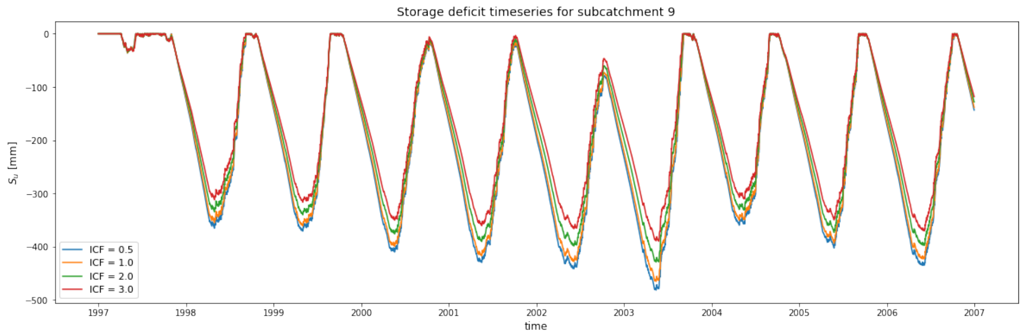


Figure 28: Storage deficit timeseries for the model period for subcatchment 9, assuming 4 different interception capacities (in mm).

For each subcatchment, the identified 9 annual maximum storage deficits were used in a Gumbel extrapolation to identify the storage deficit with a return period of 20 years. An example of this extrapolation can be seen in Figure 29, in which an ICF of 2 mm was used. The method assumes that this specific storage deficit is equal to the depth of the root-zone, and hence can be used directly as the depth of the soil moisture reservoir in the hbv model (parameter FC). However, a dependence on the interception capacity of the interception reservoir is observed. See Figure 28 and Figure 30. This dependence is almost perfectly linear and therefore the capacity of the soil moisture reservoir FC is directly related to the capacity of the interception reservoir in the model implementation. The linear relation between interception reservoir capacity and the soil moisture reservoir capacity is used to determine the capacity of the soil moisture reservoir. Therefore the interception reservoir capacity can be kept as calibration parameter, from which the soil moisture reservoir capacity is determined in every model run.

An explanation of this inverse relation between ICF and $S_{u,max}$ is that the total (potential) evaporation is fixed, but that it is not known how much of that evaporation actually can be attributed to interception evaporation, soil evaporation, transpiration and possibly open water evaporation. This means that the higher the ICF, the more interception evaporation occurs, and thus less transpiration and soil evaporation (in the model taken together in one term as the actual evaporation E_a), is realised by a smaller unsaturated zone.

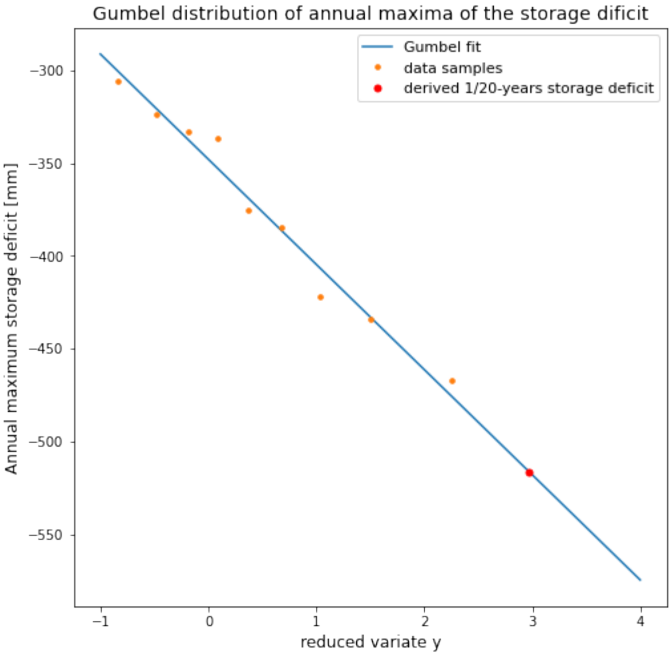


Figure 29: Gumbel extrapolation of annual storage deficit maxima for subcatchment 1 using a ICF of 2 mm.

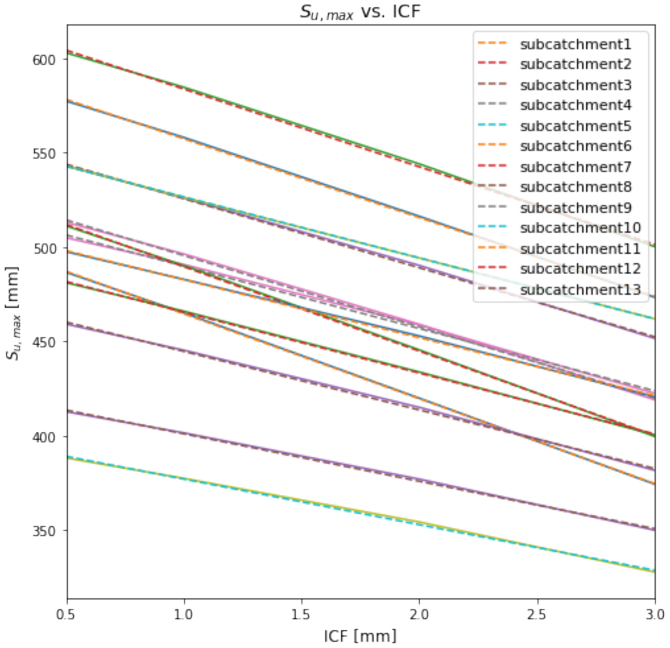


Figure 30: The almost perfectly linear relation between $S_{u,max}$ and ICF is shown in this figure for every subcatchment.

Determining the characteristic time length T

Now the model can be run with a certain combination of parameter values for ICF (based on land use), from which the FC parameter value per subcatchment and land use class is determined. This results in a timeseries of the amount of water in the soil moisture reservoir per cell in the model. These timeseries are upscaled to a subcatchment/land use class resolution, as are the SSM observations from the ESA CCI SM dataset. The T-value is determined by comparison of the modelled SM timeseries and 100 timeseries of 'observed' RS observations, each representing a different T-value per subcatchment/land use class. The T-value is a measure for how fast water sinks from the topsoil to the deeper parts of the unsaturated zone. The T-value is a non-physical parameter that summarizes the effect of different variables influences water movement in the soil, like soil depth, diffusive soil properties and vegetation and evaporation (Bouaziz et al., 2020). The T-value is the only parameter needed in this method to translate topsoil SM observations to SM observations representative for the complete unsaturated zone. For the formulas used for this transformation, see the supplementary material of Bouaziz et al. (2020) pages 3 and 4.

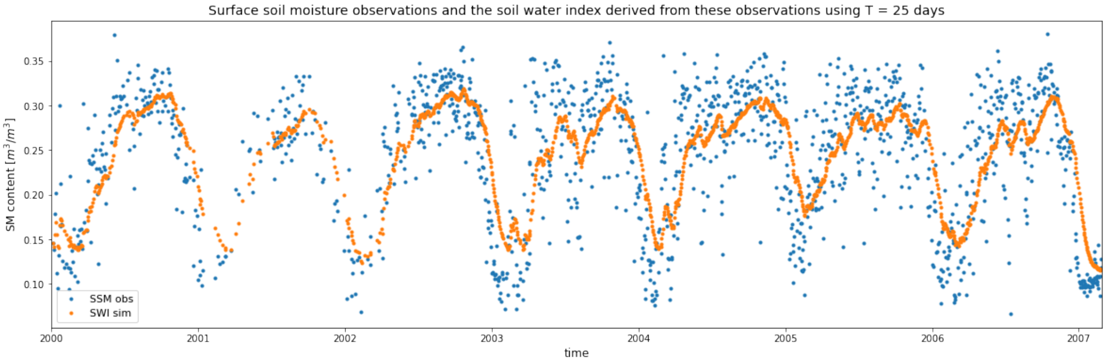


Figure 31: Surface soil moisture observations, translated to soil water index values representative for the complete unsaturated zone using a T-value of 25 days.

The T-value is now determined per subcatchment/land use class by calculating the spearman rank order correlation between the modelled SM timeseries per subcatchment/land use class, and the 100 SWI timeseries derived from RS SSM observations using T-values from 1 till 100. The T-value that results in the highest correlation is chosen as the T-value for that specific subcatchment/land use class. An example of the result of such a comparison is shown in Figure 32, in which an optimal T-value for each subcatchtment/land use class is given. Using this map of T-values, the corresponding SWI timeseries per cell are extracted from a predefined database of SWI timeseries derived from RS SSM observations for all 100 T-values. These SWI timeseries are then finally used for the model assessment using E_{SP} and E_{TMP}. Since observations from the ESA CCI SM dataset are available at 0.25° resolution, the simulated SM timeseries are also upscaled to this resolution for model assessment.

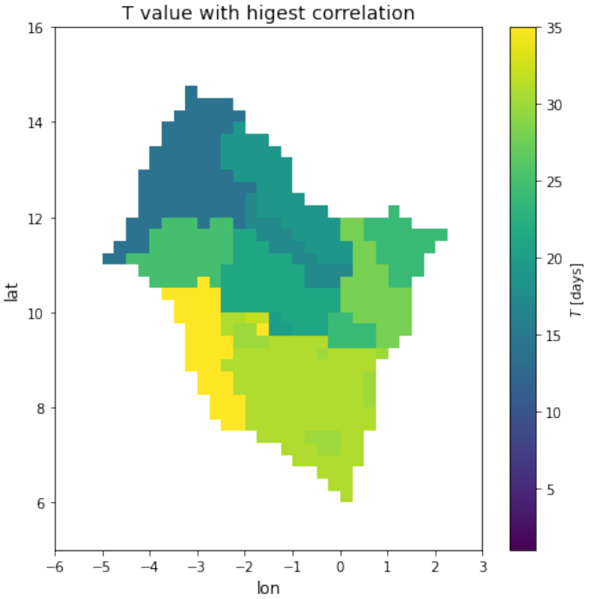


Figure 32: Optimal T-values for each subcatchment/land use class for a specific run, based on spearman rank order correlation

It can be expected that a deeper unsaturated zone results in higher T-values, because the distance that a water particle has to travel inside the unsaturated zone is then longer. Hence, there should be a strong positive relation between the T-value and the FC parameter. This expected relation was also found and an example of the results of a model run are shown in Figure 33, in which each dot represents the FC-parameter versus the T-value of a specific subcatchment/land use class. For this specific model run, an R^2 of 0.82 was found, but R^2 values generally vary between 0.77 and 0.83.

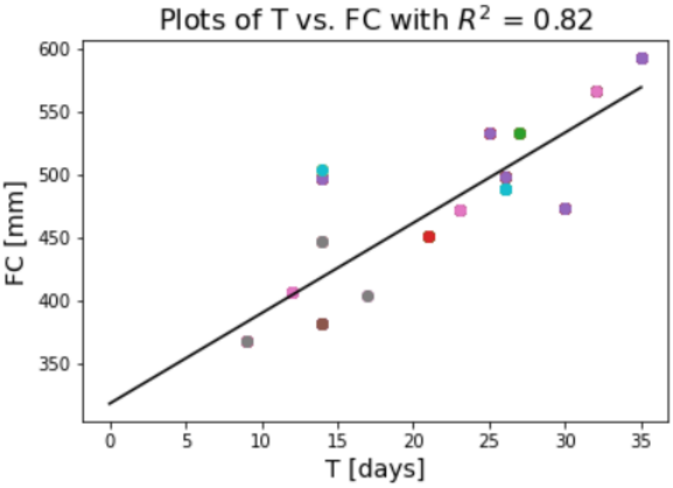


Figure 33: Example of the relation between the T-value and depth of the unsaturated root-zone per subcatchment/land use class.

E Baseflow recession curve analysis

In this appendix, a base flow recession curve analysis is performed. From this analysis it is possible to derive the linear reservoir coefficients ($K4$) of the lower zone reservoir (V_{LZ}) for each subcatchment. Recession curves were selected based on a minimum number of consecutive days with decreasing streamflow. Recession curve selections based on a minimum of 5, 10, 15, 20, 25, 30 and 35 days were used. The selected streamflow timeseries were transformed using a log-transformation. In Figure 34, an example is shown for the streamflow observations at Sabari for the full model period. It can be seen that the recession curves almost perfectly plot along a straight line and that the slope of the recession curve plots is very similar in each year.

If the lower zone reservoir is assumed the respond like a linear reservoir, the baseflow recession constant can be determined using Equation 19, which describes the outflow of a linear reservoir with reservoir constant k. Equation 20 describes the log-transformation shown in Figure 34, from which k can be solved according to Equation 21 (Ramirez, 2012).

$$Q(t) = Q(t_0)e^{-k(t-t_0)} \tag{19}$$

$$ln(Q(t)) = -k(t - t_0) + ln(Q(t_0)) \tag{20}$$

$$k = \frac{ln(Q(t_0)) - ln(Q(t))}{t - t_0} \tag{21}$$

The baseflow recession constant was determined for every gauging station and for every minimum number of days of recession. These constants are plotted in Figure 35 for every gauging station and minimum number of recession days separately. The results often show very similar baseflow constant values for different minimum days of recession. This is the case for Lawra, Chache, Bui Amont, Nawuni, Daboya, Porga, Saboba and Sabari. However, the stations Boromo and especially Yaragu and Pwalagu show a larger spread in baseflow constant values and Bamboi is somewhere in between. The stations with the largest spread in their recession constants are also the stations that were not used for calibration or evaluation because of inconsistent data, which is of course related to each other. The baseflow recession constant for each subcatchment was taken equal to the baseflow recession constant derived from the longest period of recession available. So for most of the subcatchments, the pink data points in Figure 35 were used. These selected absolute baseflow recession constant values are given in Table 12.

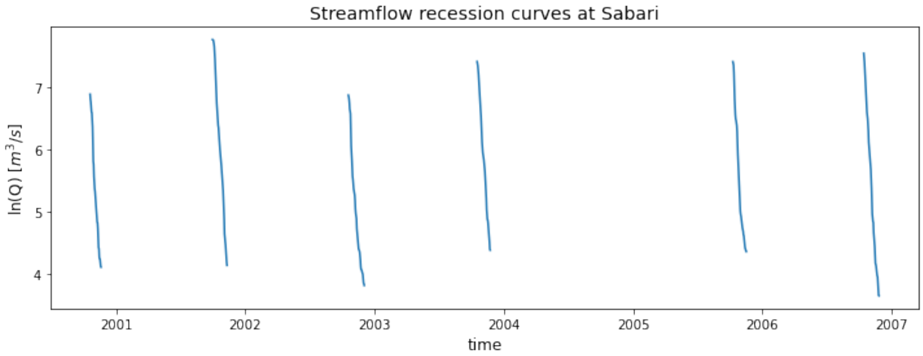


Figure 34: Baseflow recession curve of the log-transformed streamflow observations at Sabari. The recession curve plots approximate a straight line, of which the slope is equal to the linear reservoir coefficient of the lower zone in the runoff response routine ($K4$). The minimum length of the recession curve used in this plot is 30 days.

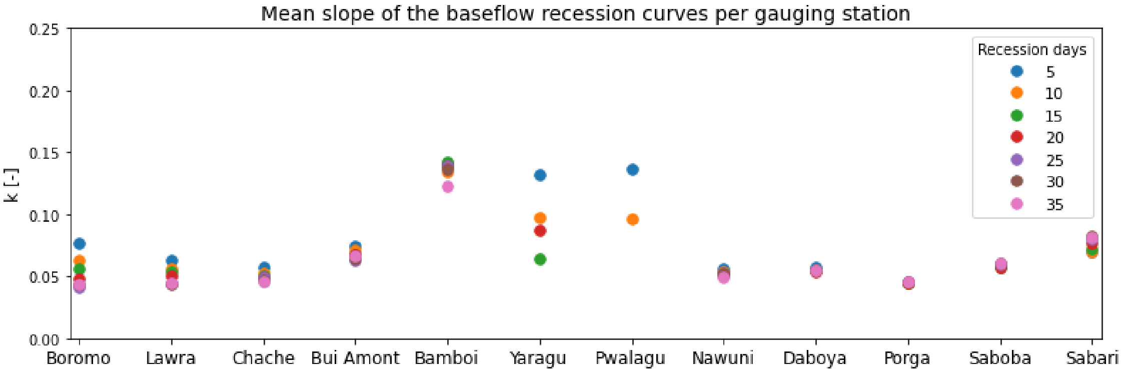


Figure 35: Mean slope of the baseflow recession curve per gauging stations plotted using a log-transformation. The minimum number of consecutive days of recession used for the recession curve selection is indicated in the legend. The reservoir coefficients of each subcatchment are directly derived from this plot.

It is observed that baseflow recession constants are generally increasing in the downstream direction of the main branches of the Volta. This holds especially for the Black Volta (Boromo, Lawra, Chache, Bui Amont and Bamboi) and the Oti (Porga, Saboba and Sabari). Baseflow recession constants in the White Volta are either not that reliable (Yaragu and Pwalagu) or are very close to each other (Nawuni and Daboya). These results could also have been expected because higher flows go down relatively faster. The k-values of the stations Nawuni and Daboya are very similar because these stations are located very close to each other.

Table 12: Overview of the determined K4 parameter values from baseflow recession curve analysis.

Subcatchment	**K4** $[d^{-1}]$
Boromo	0.044
Lawra	0.045
Chache	0.045
Bui Amont	0.066
Bamboi	0.123
Yaragu	0.087
Pwalagu	0.096
Nawuni	0.050
Daboya	0.055
Porga	0.046
Saboba	0.060
Sabari	0.082

F Amount of model runs per calibration

As was explained in subsubsection 3.3.4, it may be the case that good performing model solutions can be found by DDS in the 11-dimensional calibration problem in a relatively limited amount of model runs, because the correlation between parameters reduces the effective parameter space of the calibration problem. In this Appendix, the results of a small rest are discussed, on which the total number of model runs needed per calibration run is based. The test consists of doing multiple calibration runs for the model for the simplest scenario (Q-only), while keeping track of the result (being the model performance) but also of the effort (being the amount of model runs, which is assumed to be linearly related to computation time).

Important to interpret the results of this test is to realise that although DDS is a relatively smart algorithm, good at finding global optima of high-dimensional problems, it is still a partly stochastic algorithm and especially for a lower amount of runs with a low amount of starting samples (5 in our case), there is a large chance DDS will get stuck in local optima of the parameter space. Therefore, it can be expected that DDS will give a different solution with a possibly different 'best' parameter set for every calibration run. The goal of is test is therefore to base the final number of model runs within a calibration run on the trade-off between effort and result.

The results of the test described above are given in Figure 36, with the result (model performance in mean KGE value) on the y-axis and the effort (run time of DDS) on the x-axis. The number of model runs per calibration run is indicated with the color of the data points. It can be seen that the best results are obtained with 1.000 model runs per calibration run. However, this also takes much more time than using less runs per calibration. It is also observed that for 100 or a lower amount of runs, the model performance can be good, but can also still be improved by doing more runs, for instance by doing 200 runs. The results of the calibration run with 500 model runs is somewhat lower than expected, this could be because DDS was stuck in a local optimum here. Based on this Figure it was decided that 200 model runs should be sufficient to approach a good solution, especially if this calibration run is performed multiple times. For 200 runs and more, also the 'best' performing parameter sets seem to converge to a global optimum (not shown here), while this is much less the case for a lower amount of model runs.

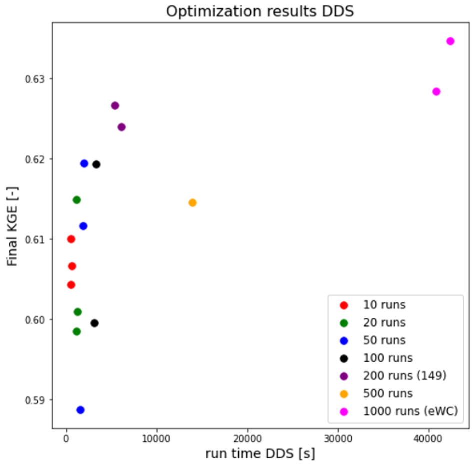


Figure 36: Results of several calibration runs with the result (KGE performance score) on he y-axis and the effort (run-time) on the x-axis. The colors of the data points indicate the number of model runs per calibration run. Calibration runs with 1.000 runs were performed on the HPC computer (indicated with eWC) and the shorter runs were performed on the Cartesius computer.

G Results Scenario 1: Q-only Calibration

Table 13: Streamflow performance scores per gauging station for simulation 1 calibrated on Q-only.

Simulation 1	**Calibration period**						**Evaluation period**					
Stat. / func.	WA_{KGE}	KGE_{OH}	KGE_{BC}	KGE_{FDC}	KGE_{ACF}	KGE_{R_c}	WA_{KGE}	KGE_{OH}	KGE_{BC}	KGE_{FDC}	KGE_{ACF}	KGE_{R_c}
Boromo	-0.28	-0.5	-0.2	0.01	0.15	-0.67	0.05	0.21	-0.04	0.19	0.86	-1.07
Lawra	0.61	0.62	0.64	0.76	0.75	0.3	0.65	0.66	0.63	0.8	0.91	0.22
Chache	0.77	0.75	0.78	0.85	0.72	0.75	0.69	0.73	0.75	0.82	0.54	0.59
Bui Amont	0.69	0.67	0.67	0.76	0.84	0.51	0.68	0.83	0.46	0.48	0.7	0.8
Bamboi	0.34	0.7	-0.61	-0.11	0.53	0.9	-0.49	0.63	-2.48	-2.48	0.21	0.69
Yaragu	-2.37	0.4	-0.2	-15.4	0.47	0.43	-1.15	-0.1	-3.12	-3.09	0.11	-0.49
Pwalagu	-0.55	0.29	-1.08	-3.35	0.77	-0.11	-0.36	0.29	-1.66	-1.65	0.69	-0.06
Nawuni	0.23	0.49	-0.36	-0.38	0.69	0.47	0.17	0.56	-0.56	-0.57	0.7	0.38
Daboya	0.49	0.7	0.14	0.15	0.61	0.68	0.51	0.88	-0.0	0.01	0.71	0.63
Porga	0.92	0.87	0.88	0.94	0.98	0.98	0.82	0.91	0.63	0.68	0.91	0.91
Saboba	0.47	0.57	0.16	0.17	0.82	0.55	0.37	0.56	0.03	0.03	0.54	0.54
Sabari	0.37	0.52	0.13	-0.08	0.69	0.49	0.44	0.46	0.3	0.3	0.68	0.45
mean cal catch	**0.64**	0.66	0.55	0.57	0.8	0.6	**0.61**	0.69	0.47	0.52	0.71	0.58
mean eval catch	**0.36**	0.6	-0.11	-0.11	0.65	0.57	**0.34**	0.72	-0.28	-0.28	0.71	0.51

Stat.: Gauging station. func.: Objective function. cal catch: calibration catchment. eval catch: evaluation catchments.
The performance scores of station Bamboi were not used for the calculation of the means.

Table 14: Streamflow performance scores per gauging station for simulation 2, calibrated on Q-only.

Simulation 2	**Calibration period**						**Evaluation period**					
Stat. / func.	WA_{KGE}	KGE_{OH}	KGE_{BC}	KGE_{FDC}	KGE_{ACF}	KGE_{R_c}	WA_{KGE}	KGE_{OH}	KGE_{BC}	KGE_{FDC}	KGE_{ACF}	KGE_{R_c}
Boromo	-0.25	-0.49	-0.12	0.06	0.21	-0.69	0.09	0.24	0.03	0.23	0.85	-1.05
Lawra	0.64	0.61	0.69	0.79	0.79	0.32	0.66	0.66	0.68	0.83	0.91	0.22
Chache	0.75	0.73	0.77	0.81	0.7	0.77	0.67	0.71	0.74	0.79	0.51	0.6
Bui Amont	0.7	0.71	0.67	0.73	0.86	0.55	0.67	0.83	0.43	0.45	0.67	0.85
Bamboi	0.32	0.68	-0.64	-0.15	0.52	0.89	-0.53	0.62	-2.59	-2.59	0.18	0.69
Yaragu	-1.12	0.43	-0.13	-8.22	0.49	0.48	-1.09	-0.07	-2.99	-2.97	0.09	-0.41
Pwalagu	0.17	0.45	-0.01	-0.17	0.57	-0.25	0.07	0.49	-0.41	-0.39	0.53	-0.27
Nawuni	0.24	0.5	-0.31	-0.33	0.66	0.47	0.19	0.56	-0.51	-0.53	0.65	0.42
Daboya	0.51	0.71	0.18	0.19	0.58	0.68	0.53	0.89	0.04	0.04	0.66	0.7
Porga	0.92	0.9	0.88	0.92	0.99	0.94	0.82	0.93	0.62	0.66	0.88	0.91
Saboba	0.47	0.54	0.2	0.21	0.79	0.52	0.37	0.53	0.06	0.07	0.51	0.52
Sabari	0.37	0.49	0.18	-0.03	0.66	0.46	0.43	0.43	0.32	0.32	0.65	0.43
mean cal catch	**0.64**	0.66	0.56	0.57	0.8	0.59	**0.6**	0.68	0.48	0.52	0.69	0.59
mean eval catch	**0.37**	0.61	-0.06	-0.07	0.62	0.57	**0.36**	0.73	-0.24	-0.24	0.66	0.56

Stat.: Gauging station. func.: Objective function. cal catch: calibration catchment. eval catch: evaluation catchments.
The performance scores of station Bamboi were not used for the calculation of the means.

Table 15: Spatial and temporal mean RS performance scores for simulation 1, calibrated on Q-only.

Simulation 1	**TWSA**	**NDVI vs. AET**	**SM**
$E_{SP_{cal,cal}}$	-0.84	-0.69	-0.15
$E_{SP_{cal,eval}}$	-0.91	-0.87	-0.15
$E_{SP_{eval,cal}}$	-2.06	-0.71	-0.52
$E_{SP_{eval,eval}}$	-1.61	-0.97	-0.47
$E_{TMP_{cal,cal}}$	0.39	-1.36	-0.82
$E_{TMP_{cal,eval}}$	0.33	-1.43	-0.9
$E_{TMP_{eval,cal}}$	0.39	-1.01	-0.81
$E_{TMP_{eval,eval}}$	0.32	-1.17	-0.96

First cal/eval in subscript indicates catchment, second cal/eval in subscript indicates period.

Table 16: Spatial and temporal mean RS performance scores for simulation 2, calibrated on Q-only.

Simulation 2	**TWSA**	**NDVI vs. AET**	**SM**
$E_{SP_{cal,cal}}$	-0.76	-0.8	-0.2
$E_{SP_{cal,eval}}$	-0.76	-1.01	-0.23
$E_{SP_{eval,cal}}$	-1.87	-0.83	-0.56
$E_{SP_{eval,eval}}$	-1.41	-1.13	-0.54
$E_{TMP_{cal,cal}}$	0.35	-1.56	-0.93
$E_{TMP_{cal,eval}}$	0.29	-1.62	-0.99
$E_{TMP_{eval,cal}}$	0.33	-1.21	-0.9
$E_{TMP_{eval,eval}}$	0.25	-1.36	-1.03

First cal/eval in subscript indicates catchment, second cal/eval in subscript indicates period.

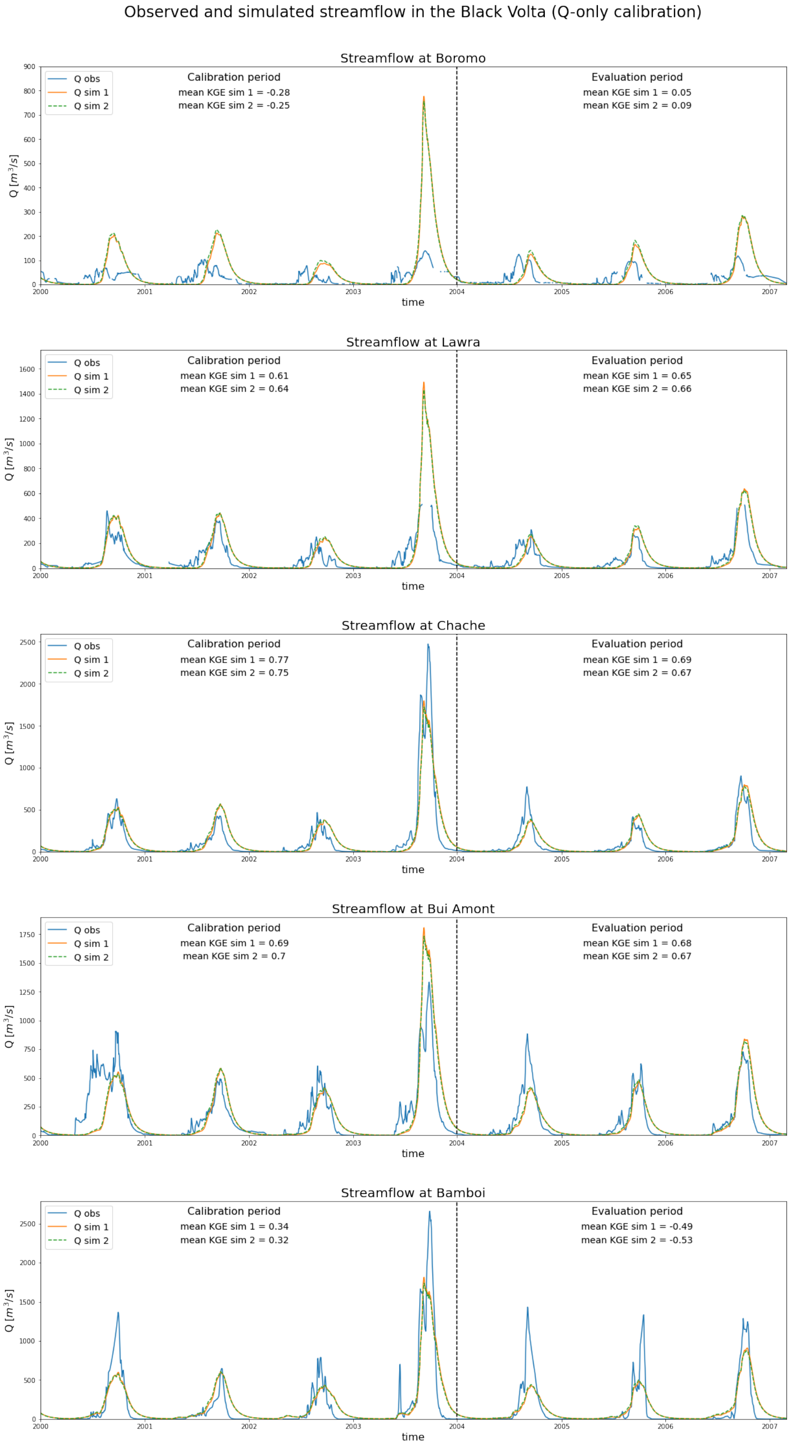


Figure 37: Hydrographs of the stations in the Black Volta based on calibration on Q only. These stations are calibration stations. The stations Boromo and Bamboi were excluded from calibration because of low data quality at Boromo, and far too high flow observations at Bamboi. The streamflow observations at Bamboi shown in this plot are scaled by catchment area to match the flow at the upstream station Bui Amont. The mean KGE values given in the plots are mean values of the 5 objective functions used for calibration as was explained in subsubsection 3.3.2.

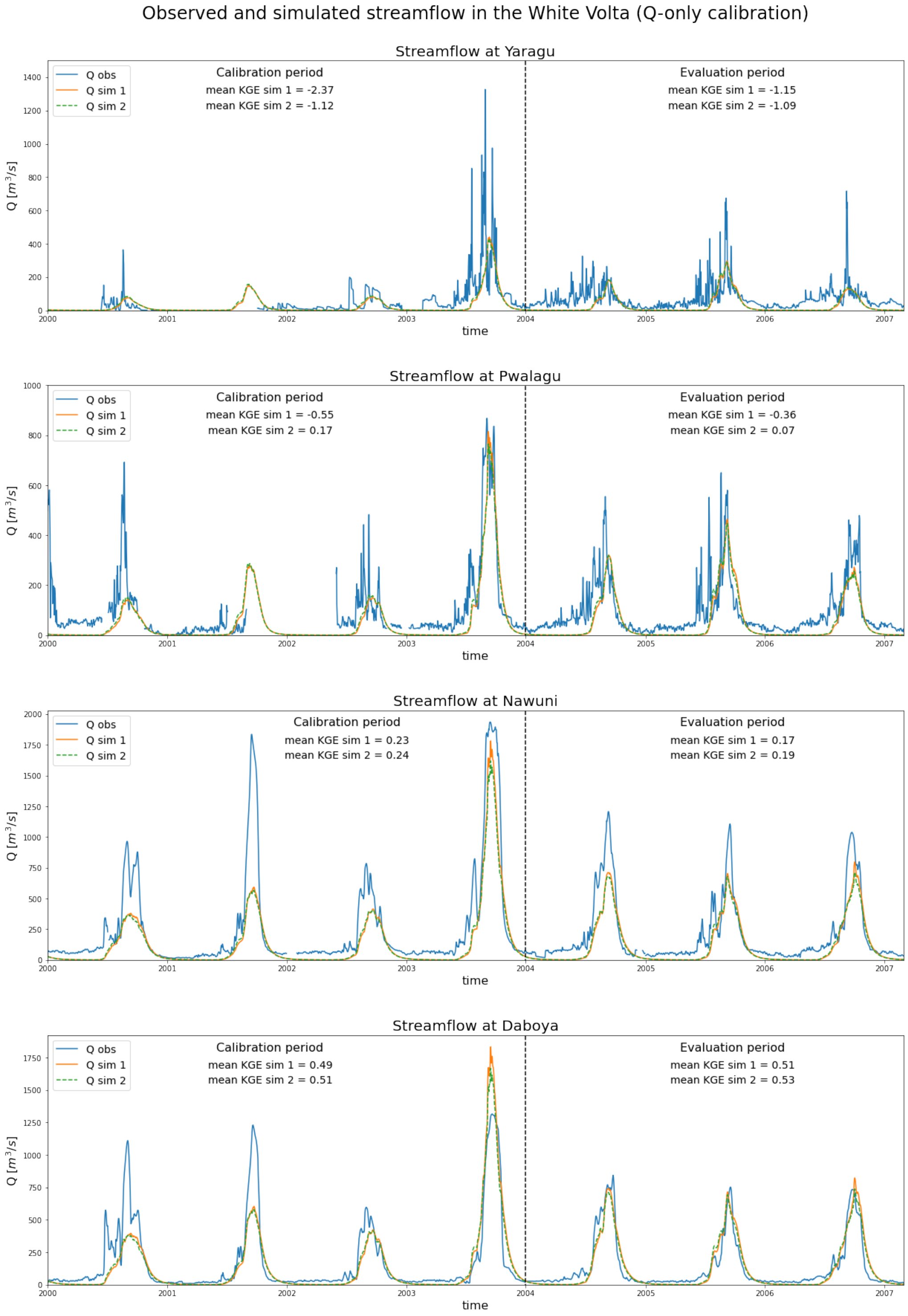


Figure 38: Hydrographs of the stations in the White Volta based on calibration on Q only. These stations are evaluation stations. The stations Yaragu and Pwalagu were excluded from evaluation because of the high influence of reservoir management on the river flow in this area. The mean KGE values given in the plots are mean values of the 5 objective functions used for calibration as was explained in subsubsection 3.3.2.

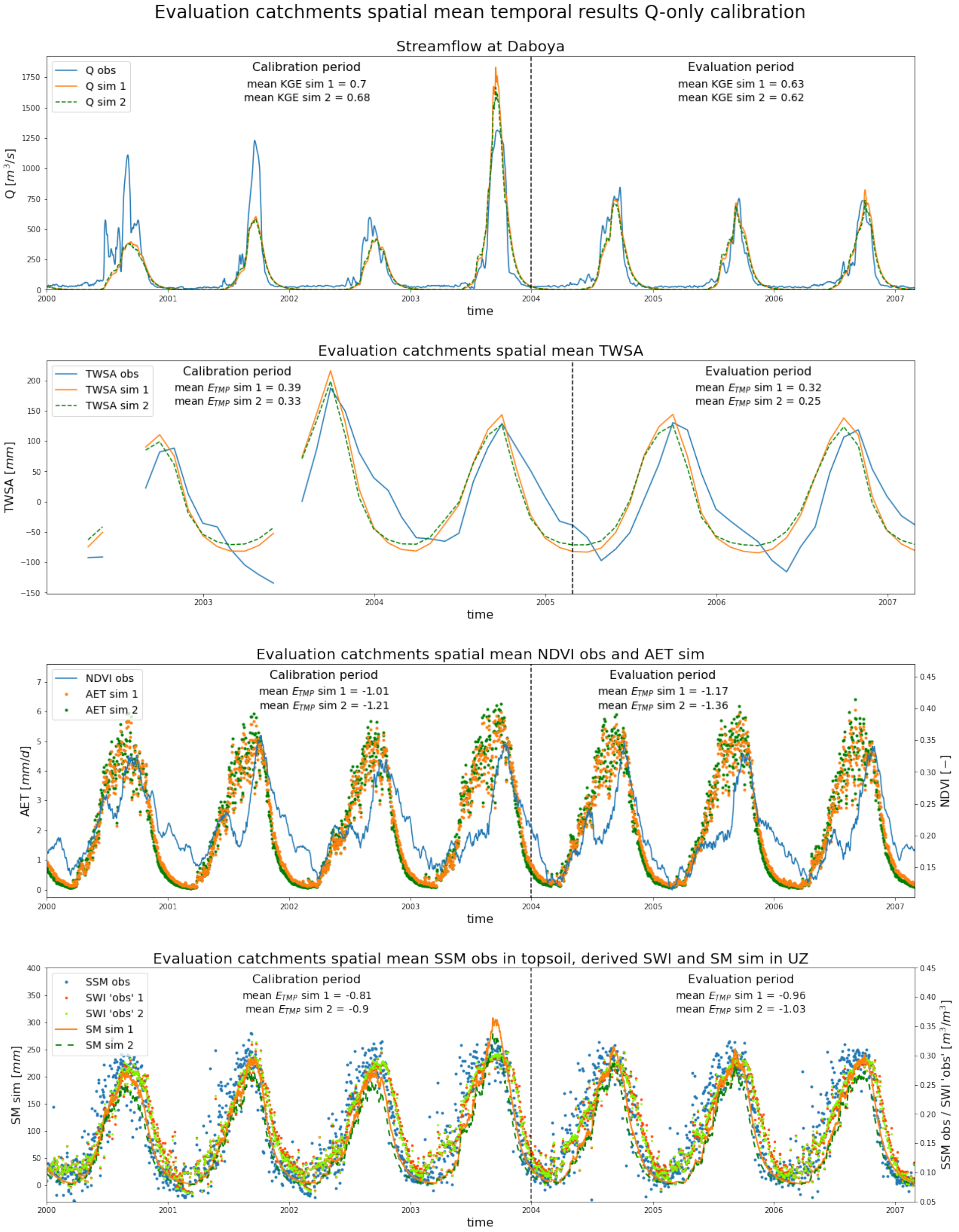


Figure 39: Timeseries of streamflow observations and simulations at Daboya (top). Timeseries of the mean TWSA observations and simulations within the evaluation catchments (second from above). Timeseries of the mean NDVI observations and mean AET simulations within the evaluation catchments (third from above). Timeseries of the mean surface soil moisture observations and the mean simulated amount of water in the soil moisture reservoir (SM) within the evaluation catchments (bottom).

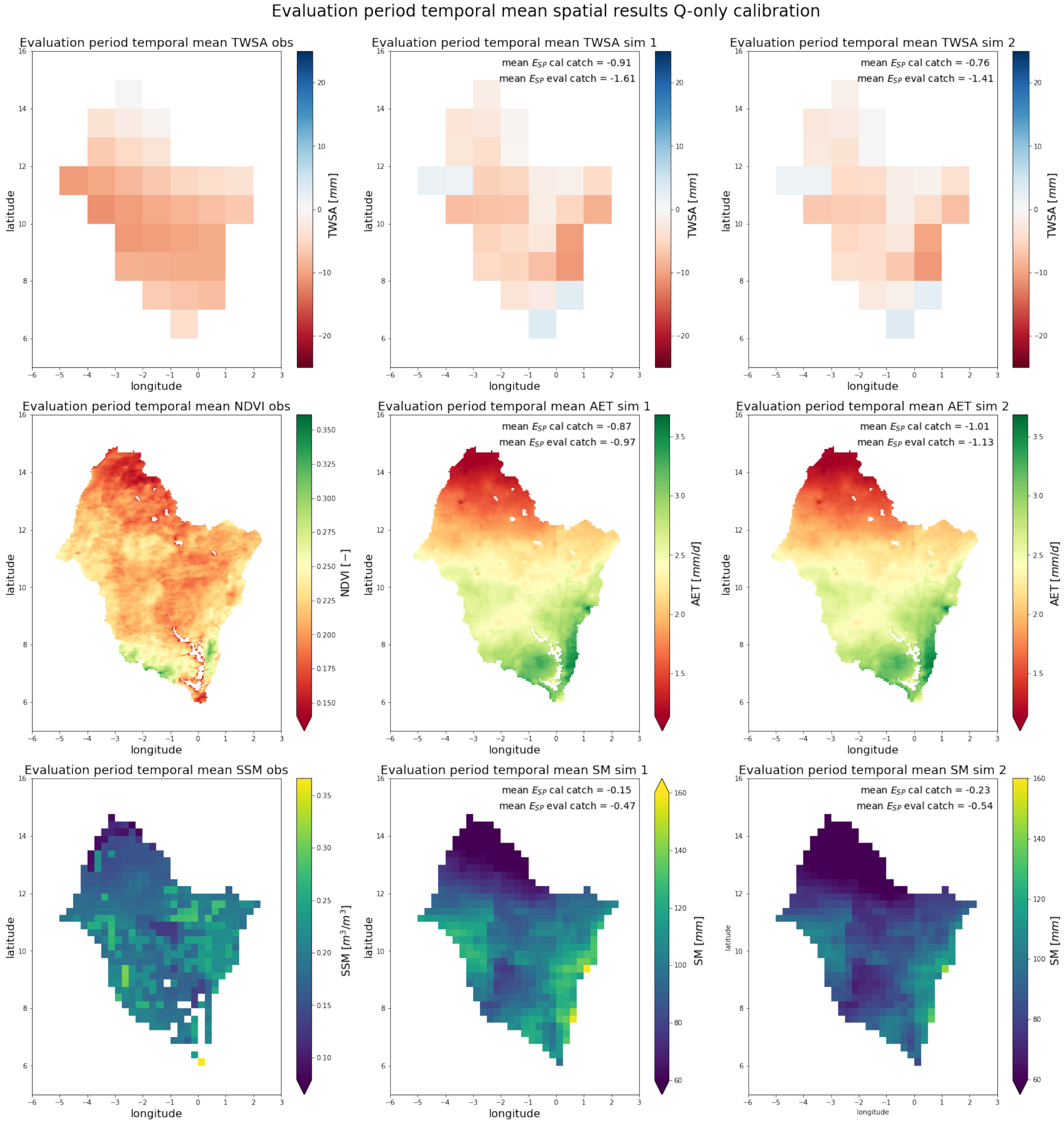

Evaluation period temporal mean spatial results Q-only calibration
Evaluation period temporal mean TWSA obs
Evaluation period temporal mean TWSA sim 1
mean E_{SP} cal catch = -0.91
mean E_{SP} eval catch = -1.61
Evaluation period temporal mean TWSA sim 2
mean E_{SP} cal catch = -0.76
mean E_{SP} eval catch = -1.41
TWSA $[mm]$
20
10
0
−10
−20
Evaluation period temporal mean NDVI obs
NDVI $[-]$
0.350
0.325
0.300
0.275
0.250
0.225
0.200
0.175
0.150
Evaluation period temporal mean AET sim 1
mean E_{SP} cal catch = -0.87
mean E_{SP} eval catch = -0.97
Evaluation period temporal mean AET sim 2
mean E_{SP} cal catch = -1.01
mean E_{SP} eval catch = -1.13
AET $[mm/d]$
3.5
3.0
2.5
2.0
1.5
Evaluation period temporal mean SSM obs
SSM $[m^3/m^3]$
0.35
0.30
0.25
0.20
0.15
0.10
Evaluation period temporal mean SM sim 1
mean E_{SP} cal catch = -0.15
mean E_{SP} eval catch = -0.47
Evaluation period temporal mean SM sim 2
mean E_{SP} cal catch = -0.23
mean E_{SP} eval catch = -0.54
SM $[mm]$
160
140
120
100
80
60
latitude
16
14
12
10
8
6
longitude
−6
−5
−4
−3
−2
−1
0
1
2
3

Figure 40: Spatial plots of the mean TWSA observations and simulations in the evaluation period in the Volta basin (top). Spatial plots of the mean NDVI observations and AET simulations in the evaluation period in the Volta basin (middle). Spatial plots of the mean surface soil moisture observations and the mean simulated amount of water in the soil moisture reservoir in the evaluation period in the Volta basin (bottom).

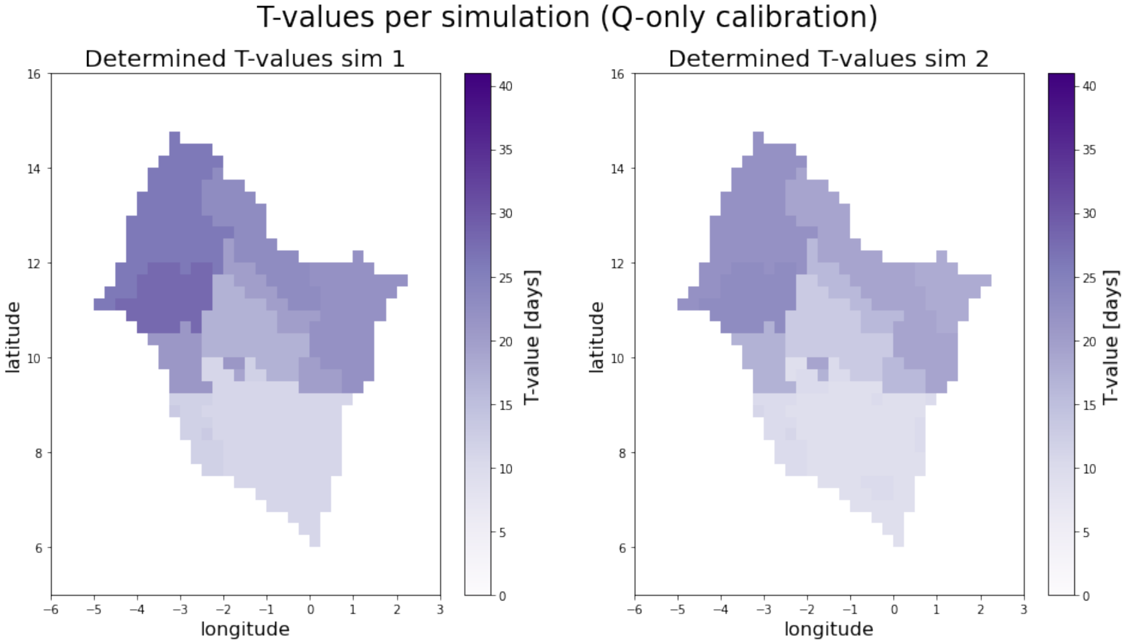

T-values per simulation (Q-only calibration)
Determined T-values sim 1
Determined T-values sim 2
latitude
longitude
T-value [days]
16
14
12
10
8
6
−6
−5
−4
−3
−2
−1
0
1
2
3
40
35
30
25
20
15
10
5
0

Figure 41: The determined T-values per subcatchment / land use class for simulation 1 and 2.

Table 17: Overview of the parameter values found in the 2 calibration runs based on Q-only

Parameter	ICF_{nf}	ICF_f	$CEVPF_{nf}$	$CEVPF_f$	β	LP	$PERC$	Q_{cf}	SUZ	$K0$	$K_{QuickFlow}$
Unit	$[mm]$	$[mm]$	$[-]$	$[-]$	$[-]$	$[-]$	$[mm/d]$	$[mm/d]$	$[mm]$	$[d^{-1}]$	$[d^{-1}]$
Sim 1	0.31	1.09	1.01	1.1	3.21	0.40	4.43	2.50	25.0	0.3	0.45
Sim 2	0.52	1.1	1.08	1.1	2.83	0.37	5.6	1.83	24.36	0.3	0.85
mean	0.41	1.09	1.04	1.1	3.02	0.39	5.02	2.16	24.68	0.3	0.65

H Results Scenario 2: Q+SM Calibration

Table 18: Streamflow performance scores per gauging station for simulation 1 calibrated on Q+SM

Simulation 1	**Calibration period**						**Evaluation period**					
Stat. / func.	WA_{KGE}	KGE_{OH}	KGE_{BC}	KGE_{FDC}	KGE_{ACF}	KGE_{R_c}	WA_{KGE}	KGE_{OH}	KGE_{BC}	KGE_{FDC}	KGE_{ACF}	KGE_{R_c}
Lawra	0.28	-0.06	0.63	0.81	0.77	-0.48	0.28	-0.02	0.62	0.77	0.91	-0.63
Chache	0.42	0.26	0.6	0.63	0.7	0.06	0.24	0.01	0.56	0.59	0.53	-0.26
Bui Amont	0.5	0.4	0.55	0.6	0.86	0.17	0.32	0.21	0.24	0.25	0.69	0.31
Bamboi	-0.06	0.22	-1.1	-0.48	0.52	0.29	-0.96	0.26	-3.51	-3.5	0.2	0.68
Nawuni	0.44	0.81	-0.16	-0.17	0.68	0.74	0.31	0.79	-0.39	-0.4	0.69	0.42
Daboya	0.52	0.66	0.28	0.29	0.6	0.65	0.31	0.41	0.1	0.11	0.7	0.15
Porga	0.63	0.37	0.74	0.76	0.99	0.49	0.52	0.4	0.47	0.5	0.92	0.44
Saboba	0.66	0.85	0.34	0.35	0.8	0.81	0.56	0.84	0.2	0.21	0.53	0.77
Sabari	0.56	0.77	0.32	0.15	0.66	0.73	0.62	0.7	0.48	0.48	0.67	0.69
mean cal catch	**0.51**	0.43	0.53	0.55	0.8	0.3	**0.42**	0.36	0.43	0.47	0.71	0.22
mean eval catch	**0.48**	0.73	0.06	0.06	0.64	0.69	**0.31**	0.6	-0.14	-0.14	0.7	0.28

Stat.: Gauging station. func.: Objective function. cal catch: calibration catchment. eval catch: evaluation catchments.
The performance scores of station Bamboi were not used for the calculation of the means.

Table 19: Streamflow performance scores per gauging station for simulation 2 calibrated on Q+SM

Simulation 2	**Calibration period**						**Evaluation period**					
Stat. / func.	WA_{KGE}	KGE_{OH}	KGE_{BC}	KGE_{FDC}	KGE_{ACF}	KGE_{R_c}	WA_{KGE}	KGE_{OH}	KGE_{BC}	KGE_{FDC}	KGE_{ACF}	KGE_{R_c}
Lawra	0.27	-0.09	0.64	0.82	0.77	-0.47	0.27	-0.05	0.63	0.77	0.91	-0.63
Chache	0.41	0.23	0.59	0.62	0.71	0.07	0.23	-0.03	0.55	0.58	0.53	-0.27
Bui Amont	0.49	0.38	0.54	0.59	0.86	0.19	0.3	0.17	0.22	0.24	0.69	0.3
Bamboi	-0.08	0.19	-1.13	-0.5	0.52	0.28	-0.99	0.23	-3.55	-3.55	0.19	0.65
Nawuni	0.46	0.82	-0.14	-0.16	0.69	0.77	0.32	0.79	-0.38	-0.39	0.69	0.45
Daboya	0.52	0.64	0.29	0.3	0.61	0.67	0.31	0.39	0.11	0.12	0.7	0.16
Porga	0.62	0.35	0.74	0.75	0.99	0.52	0.52	0.38	0.46	0.49	0.92	0.46
Saboba	0.67	0.86	0.35	0.36	0.8	0.82	0.57	0.85	0.21	0.22	0.53	0.8
Sabari	0.57	0.78	0.34	0.16	0.66	0.73	0.63	0.71	0.49	0.49	0.67	0.71
mean cal catch	**0.51**	0.42	0.53	0.55	0.8	0.31	**0.42**	0.34	0.43	0.46	0.71	0.23
mean eval catch	**0.49**	0.73	0.07	0.07	0.65	0.72	**0.31**	0.59	-0.13	-0.13	0.7	0.3

Stat.: Gauging station. func.: Objective function. cal catch: calibration catchment. eval catch: evaluation catchments.
The performance scores of station Bamboi were not used for the calculation of the means.

Table 20: Spatial and temporal mean performance RS scores for simulation 1, calibrated on Q+SM.

Simulation 1	**TWSA**	**SM**
$E_{SP_{cal,cal}}$	-0.95	-0.04
$E_{SP_{cal,eval}}$	-1.09	0.0
$E_{SP_{eval,cal}}$	-2.34	-0.4
$E_{SP_{eval,eval}}$	-1.83	-0.32
$E_{TMP_{cal,cal}}$	0.43	-0.43
$E_{TMP_{cal,eval}}$	0.37	-0.57
$E_{TMP_{eval,cal}}$	0.45	-0.46
$E_{TMP_{eval,eval}}$	0.39	-0.65

First cal/eval in subscript indicates catchment, second cal/eval in subscript indicates period.

Table 21: Spatial and temporal mean RS performance scores for simulation 2, calibrated on Q+SM.

Simulation 2	**TWSA**	**SM**
$E_{SP_{cal,cal}}$	-0.94	-0.03
$E_{SP_{cal,eval}}$	-1.07	0.0
$E_{SP_{eval,cal}}$	-2.32	-0.4
$E_{SP_{eval,eval}}$	-1.82	-0.32
$E_{TMP_{cal,cal}}$	0.43	-0.43
$E_{TMP_{cal,eval}}$	0.37	-0.57
$E_{TMP_{eval,cal}}$	0.46	-0.46
$E_{TMP_{eval,eval}}$	0.39	-0.65

First cal/eval in subscript indicates catchment, second cal/eval in subscript indicates period.

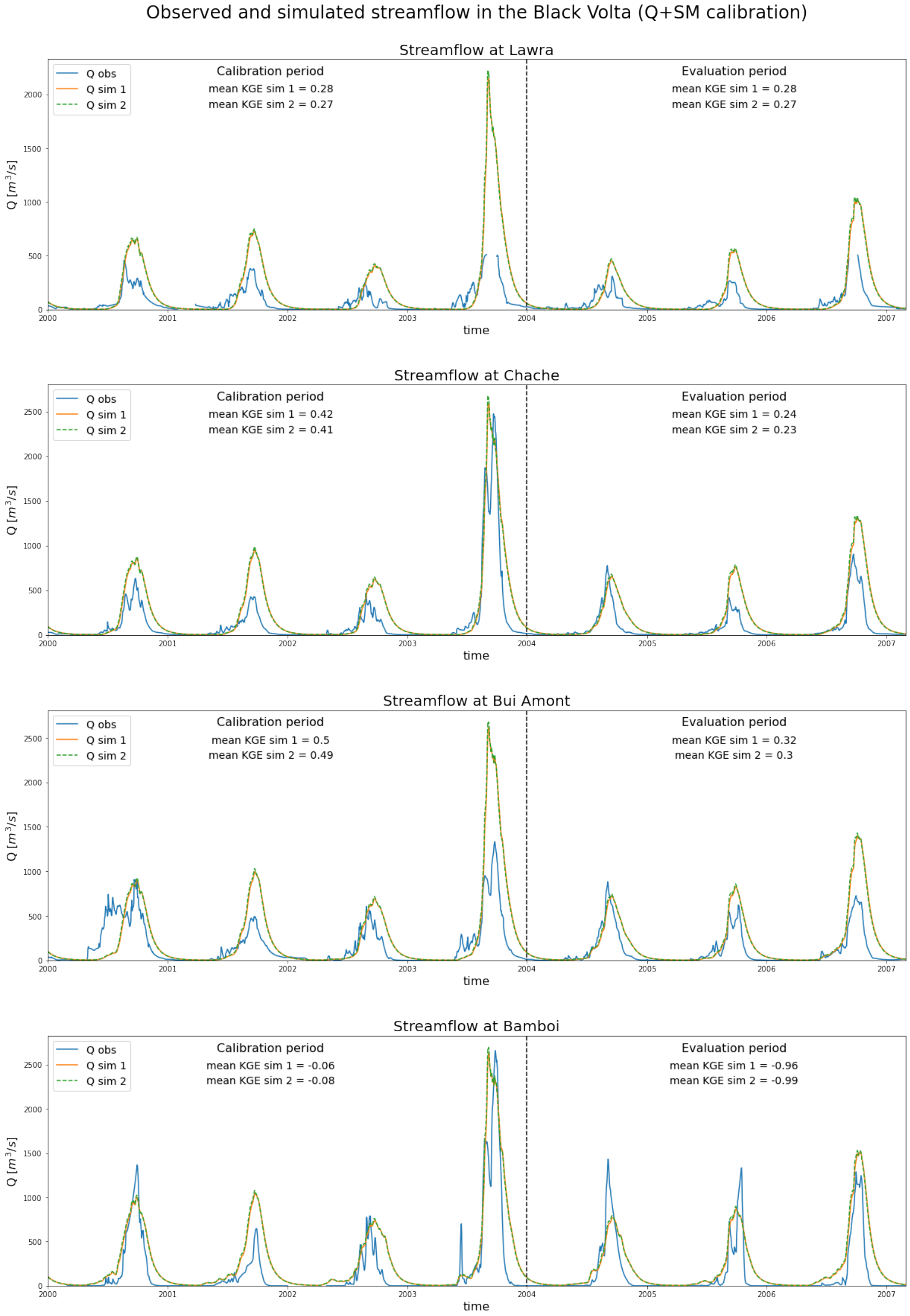


Figure 42: Hydrographs of the stations in the Black Volta based on calibration on Q+SM. These stations are calibration stations. The stations Boromo and Bamboi were excluded from calibration because of low data quality at Boromo, and far too high flow observations at Bamboi. The streamflow observations at Bamboi shown in this plot are scaled by catchment area to match the flow at the upstream station Bui Amont. The mean KGE values given in the plots are mean values of the 5 objective functions used for calibration as was explained in subsubsection 3.3.2.

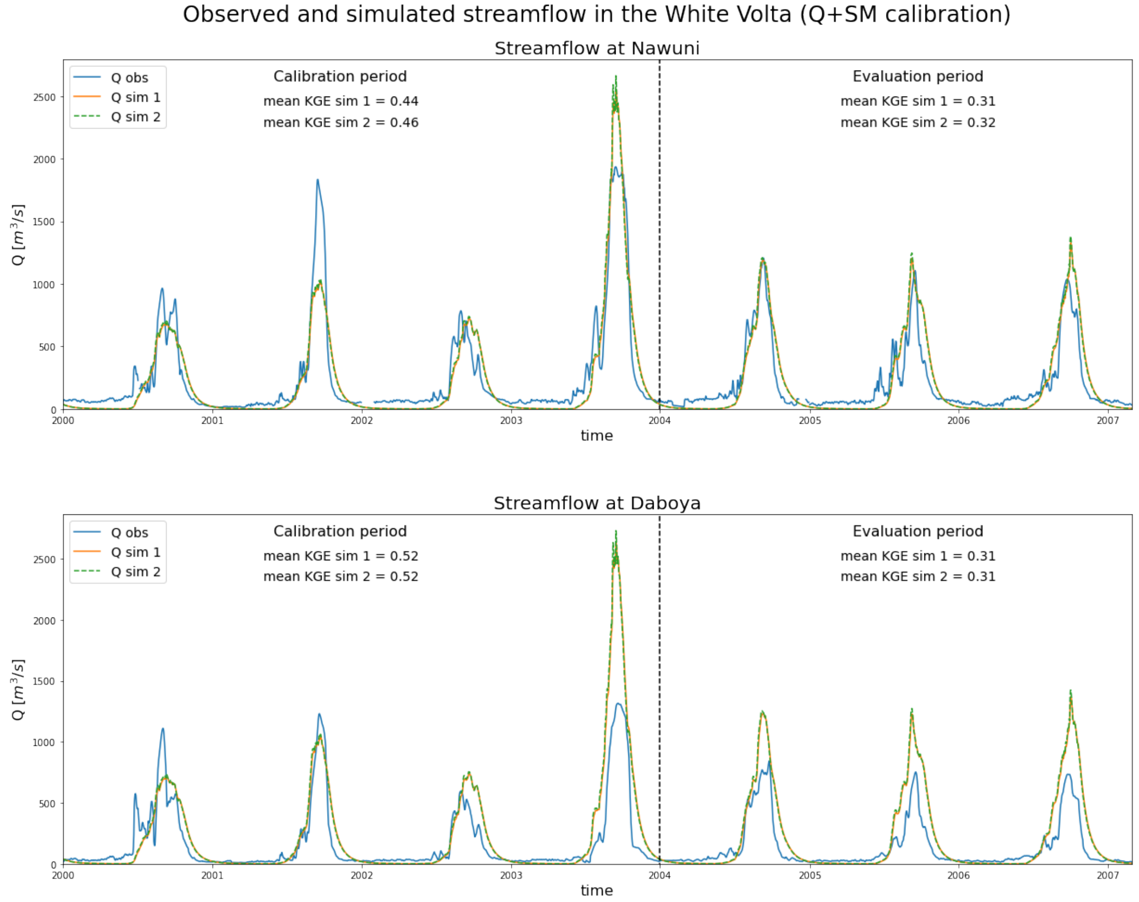


Figure 43: Hydrographs of the stations in the White Volta based on calibration on Q+SM. These stations are evaluation stations. The stations Yaragu and Pwalagu were excluded from evaluation because of the high influence of reservoir management on the river flow in this area. The mean KGE values given in the plots are mean values of the 5 objective functions used for calibration as was explained in subsubsection 3.3.2.

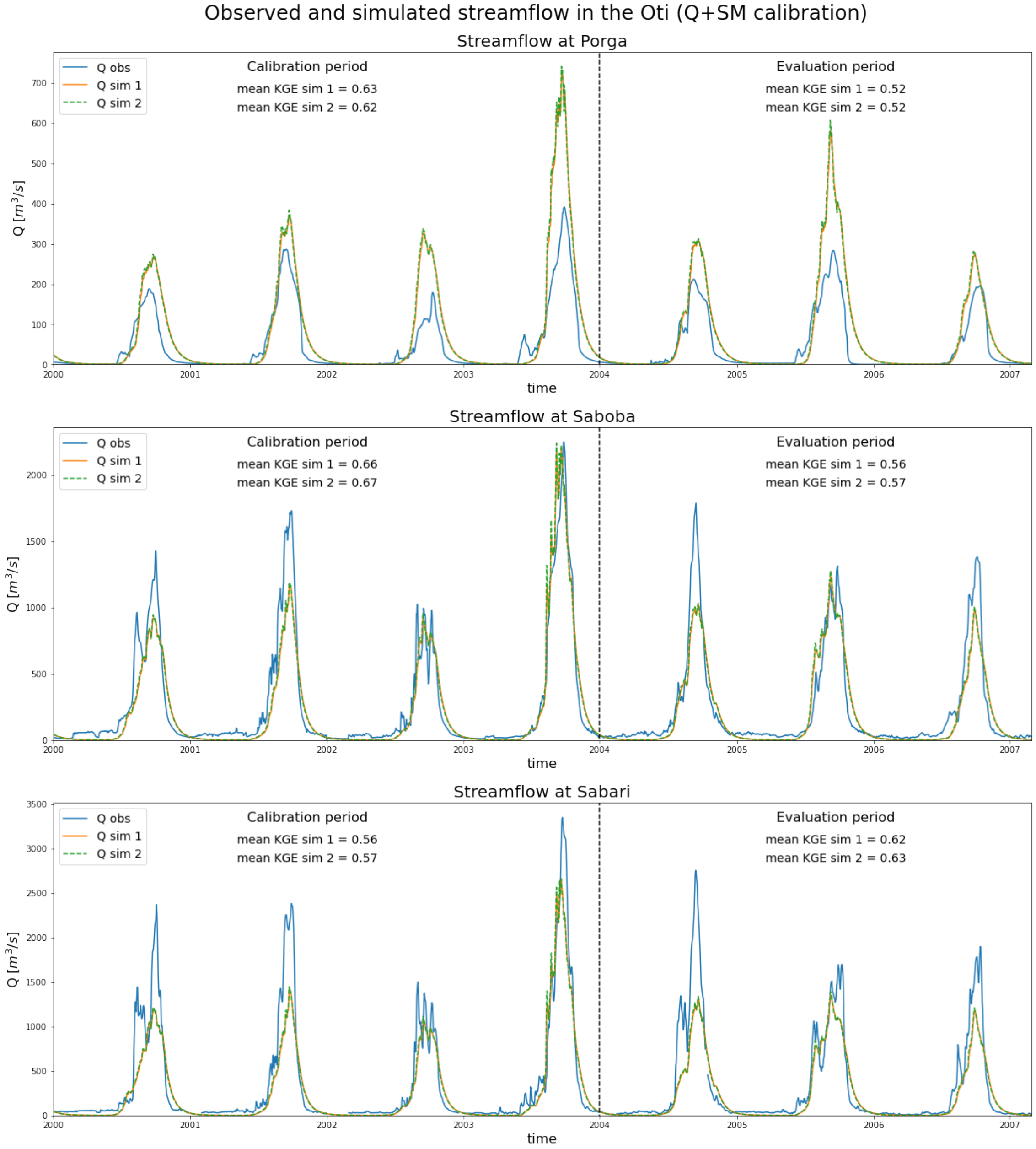


Figure 44: Hydrographs of the stations in the Oti based on calibration on Q+SM. These stations are calibration stations. The mean KGE values given in the plots are mean values of the 5 objective functions used for calibration as was explained in subsubsection 3.3.2.

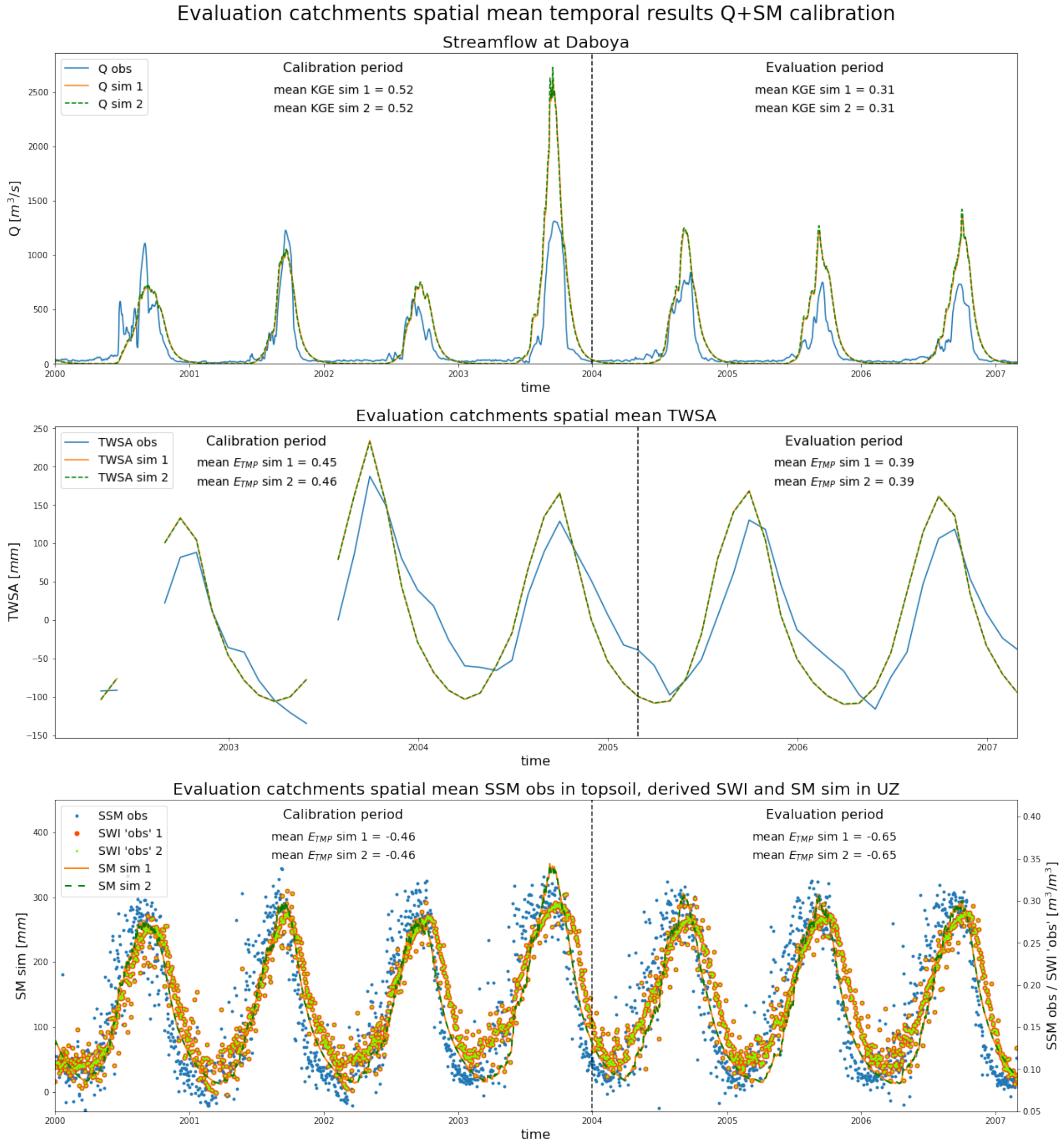


Figure 45: Timeseries of streamflow observations and simulations at Daboya (top). Timeseries of the mean TWSA observations and simulations within the evaluation catchments (middle). Timeseries of the mean surface soil moisture observations and the mean simulated amount of water in the soil moisture reservoir (SM) within the evaluation catchments (bottom).

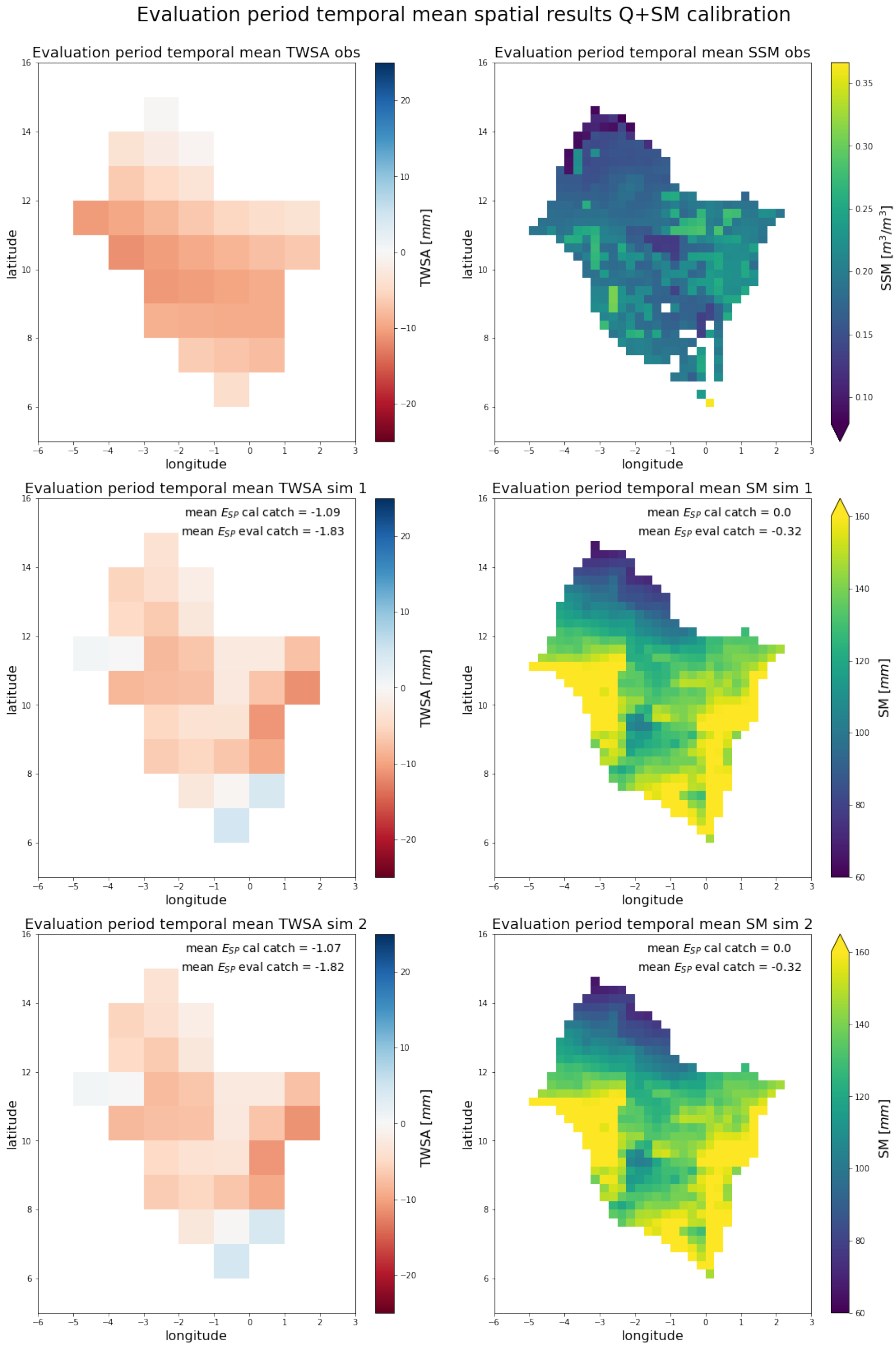

Evaluation period temporal mean spatial results Q+SM calibration
Evaluation period temporal mean TWSA obs
Evaluation period temporal mean SSM obs
Evaluation period temporal mean TWSA sim 1
mean E_{SP} cal catch = -1.09
mean E_{SP} eval catch = -1.83
Evaluation period temporal mean SM sim 1
mean E_{SP} cal catch = 0.0
mean E_{SP} eval catch = -0.32
Evaluation period temporal mean TWSA sim 2
mean E_{SP} cal catch = -1.07
mean E_{SP} eval catch = -1.82
Evaluation period temporal mean SM sim 2
mean E_{SP} cal catch = 0.0
mean E_{SP} eval catch = -0.32
latitude
longitude
TWSA $[mm]$
SSM $[m^3/m^3]$
SM $[mm]$
16
14
12
10
8
6
−6
−5
−4
−3
−2
−1
0
1
2
3
20
10
0
−10
−20
0.35
0.30
0.25
0.20
0.15
0.10
160
140
120
100
80
60

Figure 46: Spatial plots of the mean TWSA observations and simulations in the evaluation period in the Volta basin (left). Spatial plots of the mean surface soil moisture observations and the mean simulated amount of water in the soil moisture reservoir in the evaluation period in the Volta basin (right).

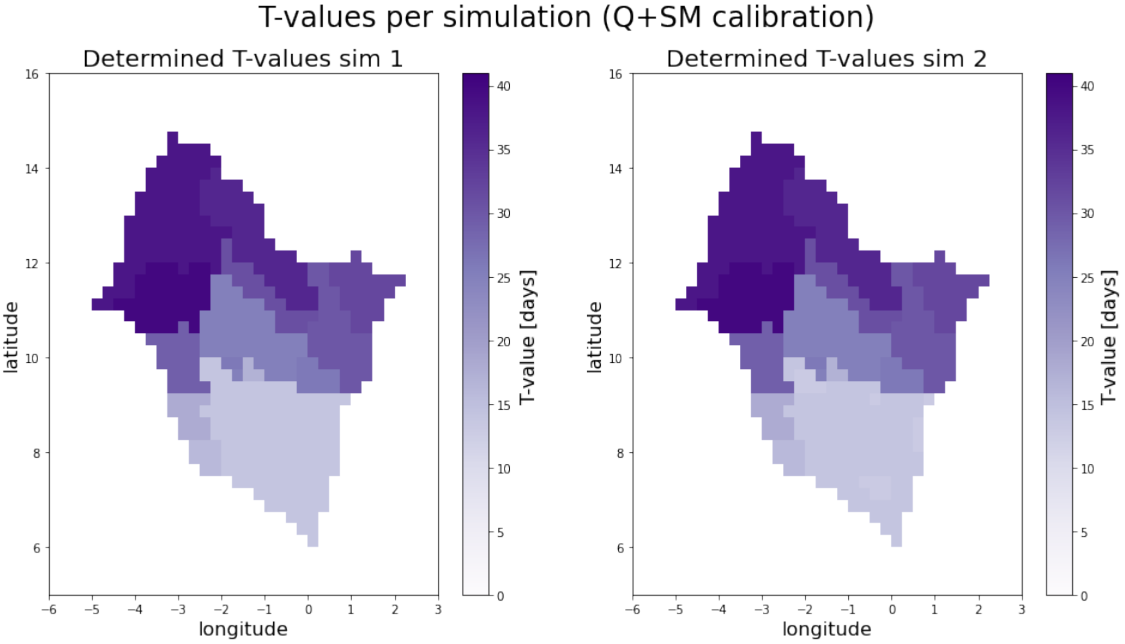

T-values per simulation (Q+SM calibration)
Determined T-values sim 1
Determined T-values sim 2
latitude
longitude
T-value [days]
16
14
12
10
8
6
−6
−5
−4
−3
−2
−1
0
1
2
3
0
5
10
15
20
25
30
35
40

Figure 47: The determined T-values per subcatchment / land use class for simulation 1 and 2.

Table 22: Overview of the parameter values found in the 2 calibration runs based on Q+SM

Parameter	ICF_{nf}	ICF_f	$CEVPF_{nf}$	$CEVPF_f$	β	LP	$PERC$	Q_{cf}	SUZ	$K0$	$K_{QuickFlow}$
Unit	$[mm]$	$[mm]$	$[-]$	$[-]$	$[-]$	$[-]$	$[mm/d]$	$[mm/d]$	$[mm]$	$[d^{-1}]$	$[d^{-1}]$
Sim 1	0.3	0.83	0.9	1.1	3.5	0.55	4.33	0.6	21.01	0.16	0.45
Sim 2	0.3	0.8	0.9	1.1	3.43	0.55	5.37	2.5	14.95	0.29	0.52
mean	0.3	0.81	0.9	1.1	3.46	0.55	4.85	1.55	17.98	0.22	0.48

I Results Scenario 3: Q+TWSA Calibration

Table 23: Streamflow performance scores per gauging station for simulation 1 calibrated on Q+TWSA

Simulation 1	**Calibration period**						**Evaluation period**					
Stat. / func.	WA_{KGE}	KGE_{OH}	KGE_{BC}	KGE_{FDC}	KGE_{ACF}	KGE_{R_c}	WA_{KGE}	KGE_{OH}	KGE_{BC}	KGE_{FDC}	KGE_{ACF}	KGE_{R_c}
Lawra	0.6	0.52	0.7	0.81	0.79	0.27	0.62	0.57	0.69	0.84	0.9	0.13
Chache	0.72	0.68	0.75	0.78	0.71	0.71	0.62	0.63	0.72	0.75	0.51	0.51
Bui Amont	0.71	0.72	0.66	0.72	0.85	0.58	0.65	0.79	0.41	0.42	0.67	0.86
Bamboi	0.29	0.66	-0.7	-0.18	0.53	0.85	-0.57	0.62	-2.7	-2.7	0.18	0.73
Nawuni	0.27	0.53	-0.28	-0.3	0.66	0.51	0.21	0.6	-0.5	-0.52	0.64	0.48
Daboya	0.53	0.76	0.2	0.2	0.58	0.72	0.54	0.92	0.04	0.04	0.65	0.74
Porga	0.91	0.87	0.87	0.9	0.99	0.93	0.81	0.91	0.61	0.64	0.88	0.94
Saboba	0.48	0.56	0.22	0.22	0.79	0.53	0.38	0.55	0.07	0.08	0.51	0.53
Sabari	0.39	0.5	0.2	-0.01	0.66	0.47	0.44	0.45	0.34	0.34	0.64	0.45
mean cal catch	**0.63**	0.64	0.57	0.57	0.8	0.58	**0.59**	0.65	0.47	0.51	0.69	0.57
mean eval catch	**0.4**	0.65	-0.04	-0.05	0.62	0.62	**0.38**	0.76	-0.23	-0.24	0.64	0.61

Stat.: Gauging station. func.: Objective function. cal catch: calibration catchment. eval catch: evaluation catchments.
The performance scores of station Bamboi were not used for the calculation of the means.

Table 24: Streamflow performance scores per gauging station for simulation 2 calibrated on Q+TWSA

Simulation 2	**Calibration period**						**Evaluation period**					
Stat. / func.	WA_{KGE}	KGE_{OH}	KGE_{BC}	KGE_{FDC}	KGE_{ACF}	KGE_{R_c}	WA_{KGE}	KGE_{OH}	KGE_{BC}	KGE_{FDC}	KGE_{ACF}	KGE_{R_c}
Lawra	0.59	0.42	0.75	0.88	0.86	0.21	0.59	0.49	0.75	0.86	0.83	0.09
Chache	0.66	0.62	0.69	0.71	0.67	0.66	0.57	0.56	0.66	0.68	0.47	0.48
Bui Amont	0.71	0.75	0.62	0.66	0.9	0.61	0.61	0.73	0.35	0.36	0.62	0.88
Bamboi	0.24	0.61	-0.78	-0.26	0.5	0.8	-0.65	0.59	-2.89	-2.88	0.13	0.73
Nawuni	0.31	0.55	-0.16	-0.18	0.61	0.52	0.26	0.63	-0.37	-0.39	0.59	0.53
Daboya	0.57	0.78	0.29	0.29	0.53	0.74	0.58	0.93	0.13	0.14	0.6	0.78
Porga	0.87	0.85	0.82	0.83	0.95	0.93	0.77	0.86	0.55	0.57	0.83	0.95
Saboba	0.51	0.57	0.32	0.32	0.76	0.53	0.42	0.56	0.18	0.19	0.48	0.55
Sabari	0.42	0.5	0.3	0.12	0.63	0.47	0.47	0.46	0.43	0.43	0.6	0.46
mean cal catch	**0.63**	0.62	0.58	0.59	0.8	0.57	**0.57**	0.61	0.49	0.52	0.64	0.57
mean eval catch	**0.44**	0.67	0.06	0.06	0.57	0.63	**0.42**	0.78	-0.12	-0.13	0.6	0.65

Stat.: Gauging station. func.: Objective function. cal catch: calibration catchment. eval catch: evaluation catchments.
The performance scores of station Bamboi were not used for the calculation of the means.

Table 25: Spatial and temporal mean RS performance scores for simulation 1, calibrated on Q+TWSA.

Simulation 1	**TWSA**	**SM**
$E_{SP_{cal,cal}}$	-0.69	-0.22
$E_{SP_{cal,eval}}$	-0.7	-0.24
$E_{SP_{eval,cal}}$	-1.79	-0.58
$E_{SP_{eval,eval}}$	-1.32	-0.56
$E_{TMP_{cal,cal}}$	0.34	-0.95
$E_{TMP_{cal,eval}}$	0.29	-1.01
$E_{TMP_{eval,cal}}$	0.32	-0.92
$E_{TMP_{eval,eval}}$	0.24	-1.07

First cal/eval in subscript indicates catchment, second cal/eval in subscript indicates period.

Table 26: Spatial and temporal mean RS performance scores for simulation 2, calibrated on Q+TWSA.

Simulation 2	**TWSA**	**SM**
$E_{SP_{cal,cal}}$	-0.62	-0.43
$E_{SP_{cal,eval}}$	-0.56	-0.59
$E_{SP_{eval,cal}}$	-1.49	-0.79
$E_{SP_{eval,eval}}$	-1.14	-0.84
$E_{TMP_{cal,cal}}$	0.22	-1.2
$E_{TMP_{cal,eval}}$	0.15	-1.23
$E_{TMP_{eval,cal}}$	0.18	-1.14
$E_{TMP_{eval,eval}}$	0.08	-1.28

First cal/eval in subscript indicates catchment, second cal/eval in subscript indicates period.

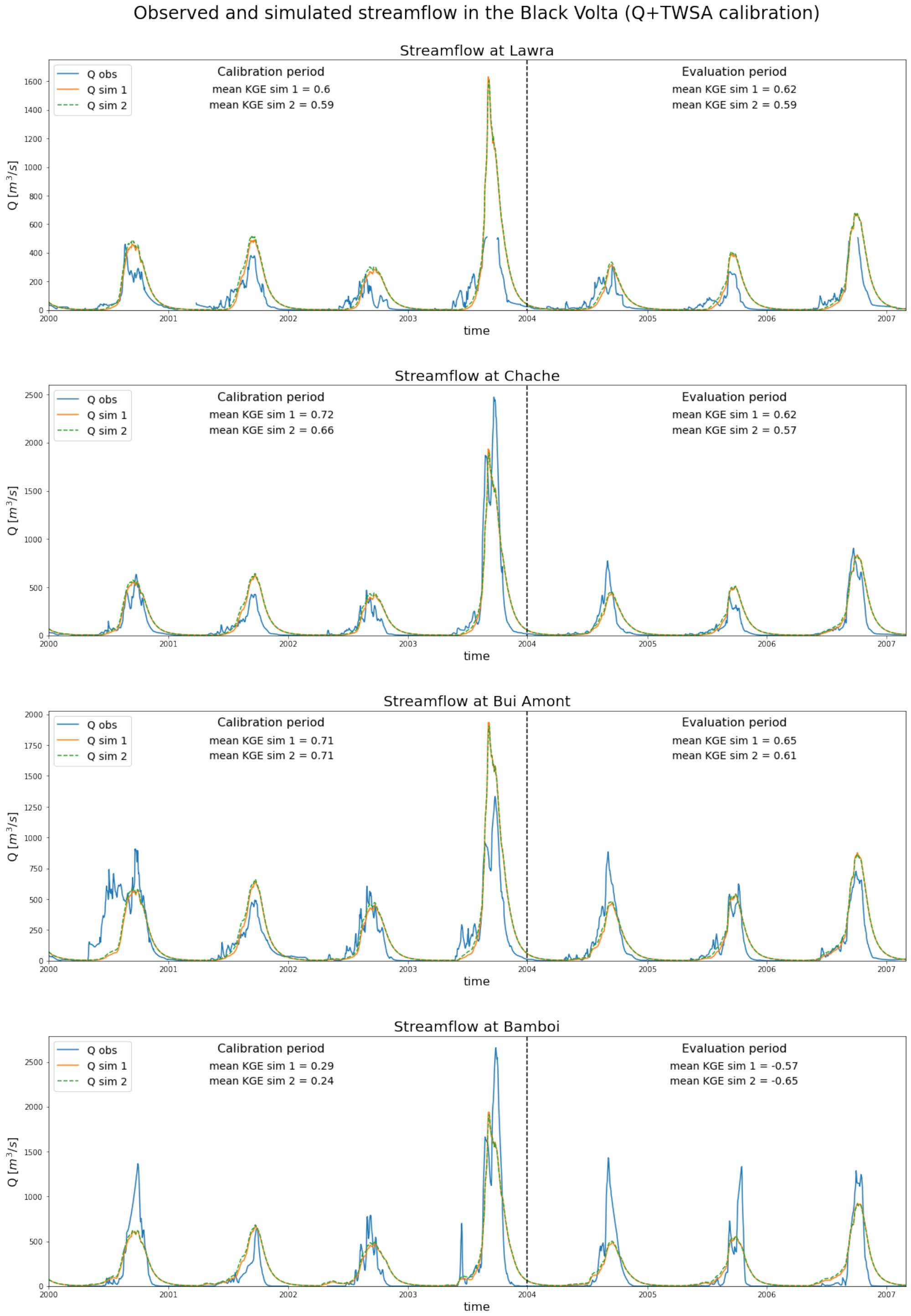


Figure 48: Hydrographs of the stations in the Black Volta based on calibration on Q+TWSA. These stations are calibration stations. The stations Boromo and Bamboi were excluded from calibration because of low data quality at Boromo, and far too high flow observations at Bamboi. The streamflow observations at Bamboi shown in this plot are scaled by catchment area to match the flow at the upstream station Bui Amont. The mean KGE values given in the plots are mean values of the 5 objective functions used for calibration as was explained in subsubsection 3.3.2.

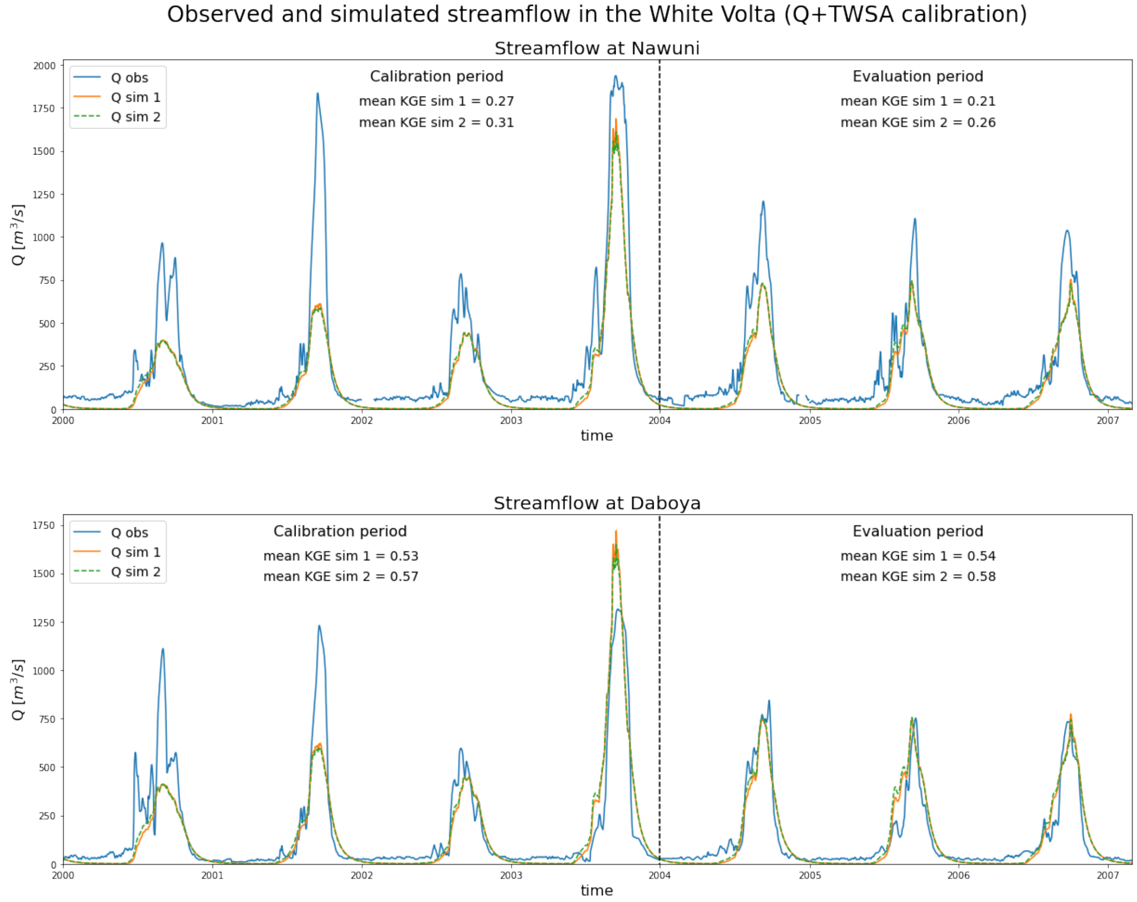


Figure 49: Hydrographs of the stations in the White Volta based on calibration on Q+TWSA. These stations are evaluation stations. The stations Yaragu and Pwalagu were excluded from evaluation because of the high influence of reservoir management on the river flow in this area. The mean KGE values given in the plots are mean values of the 5 objective functions used for calibration as was explained in subsubsection 3.3.2.

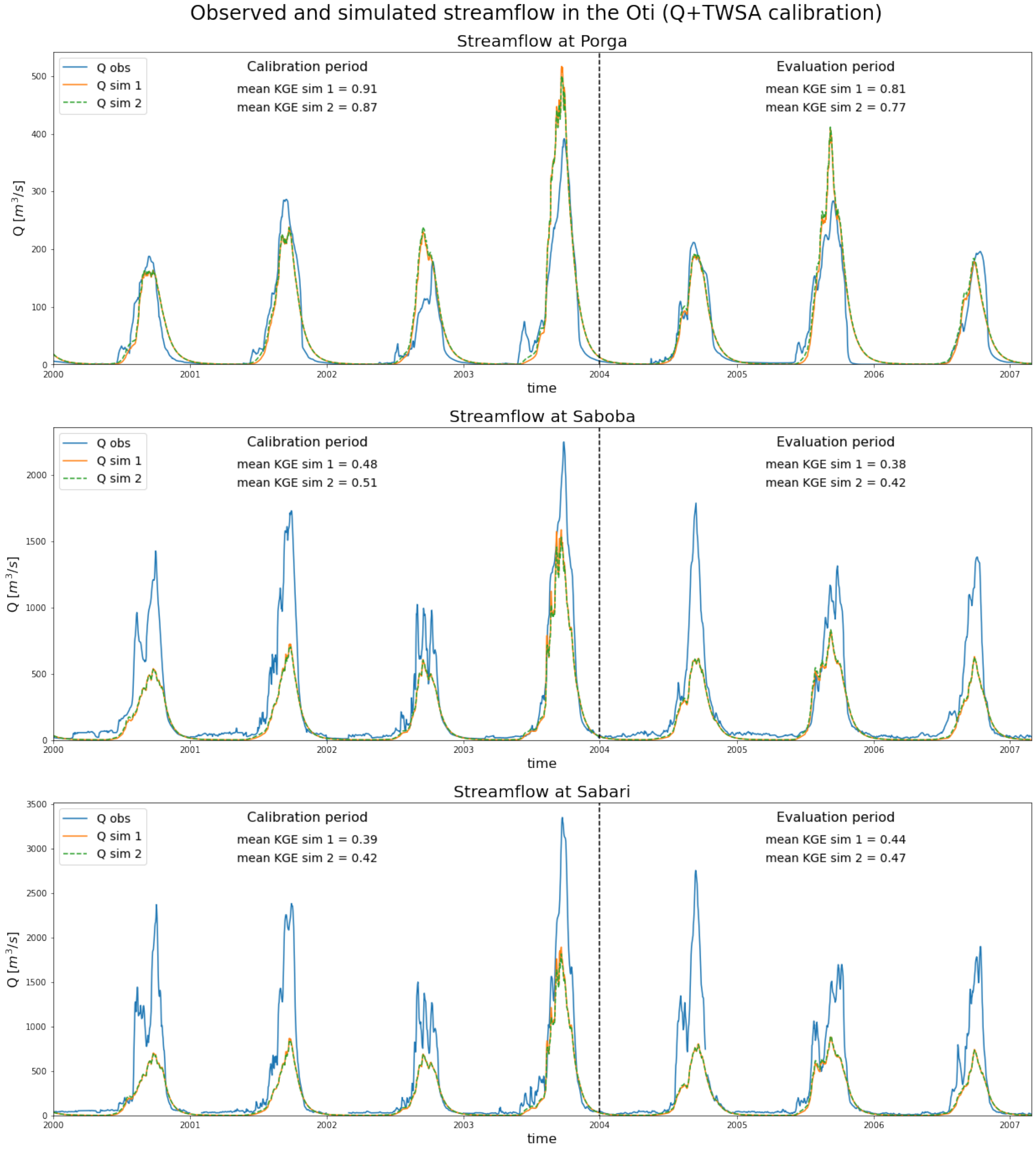


Figure 50: Hydrographs of the stations in the Oti based on calibration on Q+TWSA. These stations are calibration stations. The mean KGE values given in the plots are mean values of the 5 objective functions used for calibration as was explained in subsubsection 3.3.2.

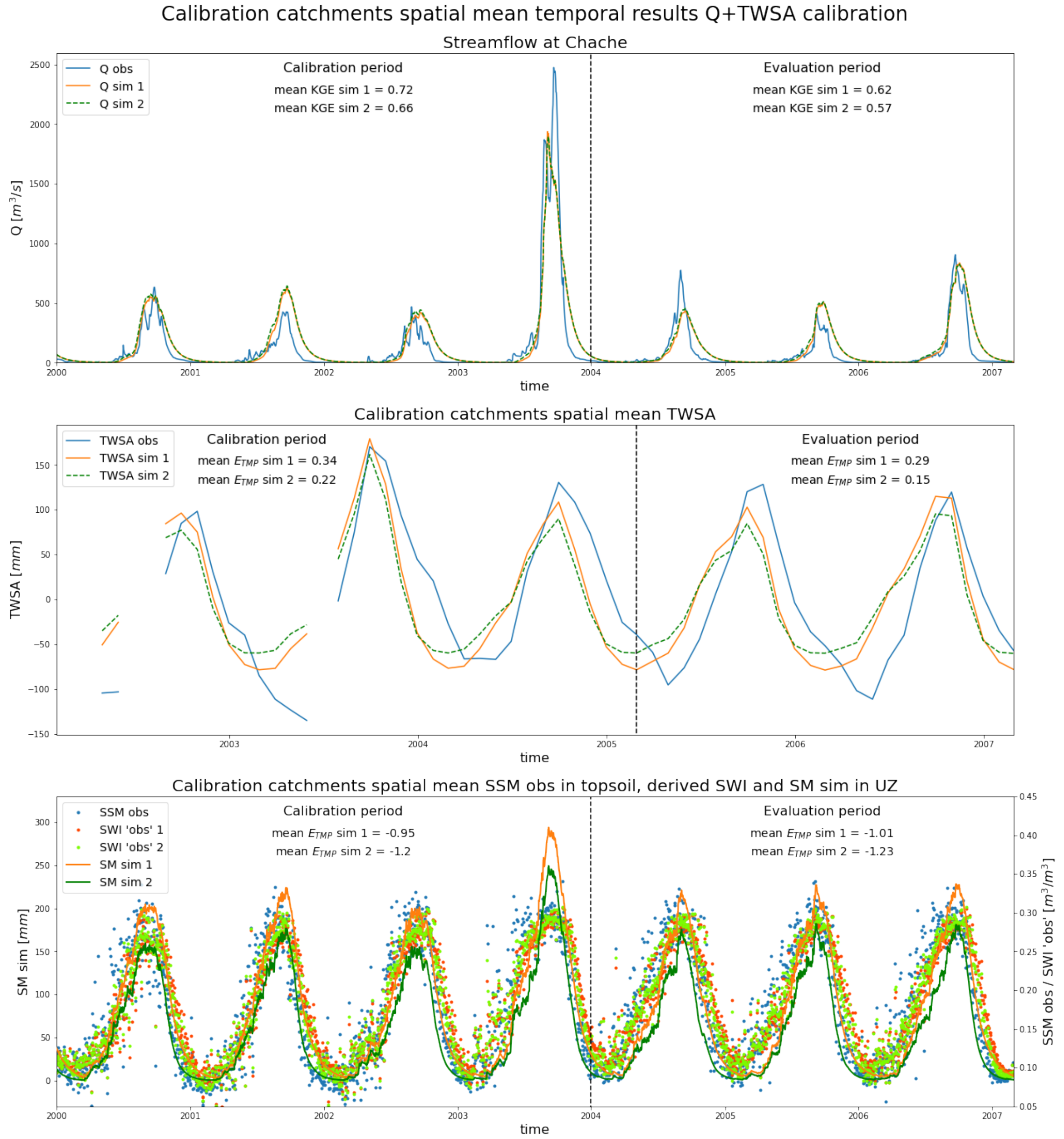


Figure 51: Timeseries of streamflow observations and simulations at Chache (top). Timeseries of the mean TWSA observations and simulations within the calibration catchments (middle). Timeseries of the mean surface soil moisture observations and the mean simulated amount of water in the soil moisture reservoir (SM) within the calibration catchments (bottom).

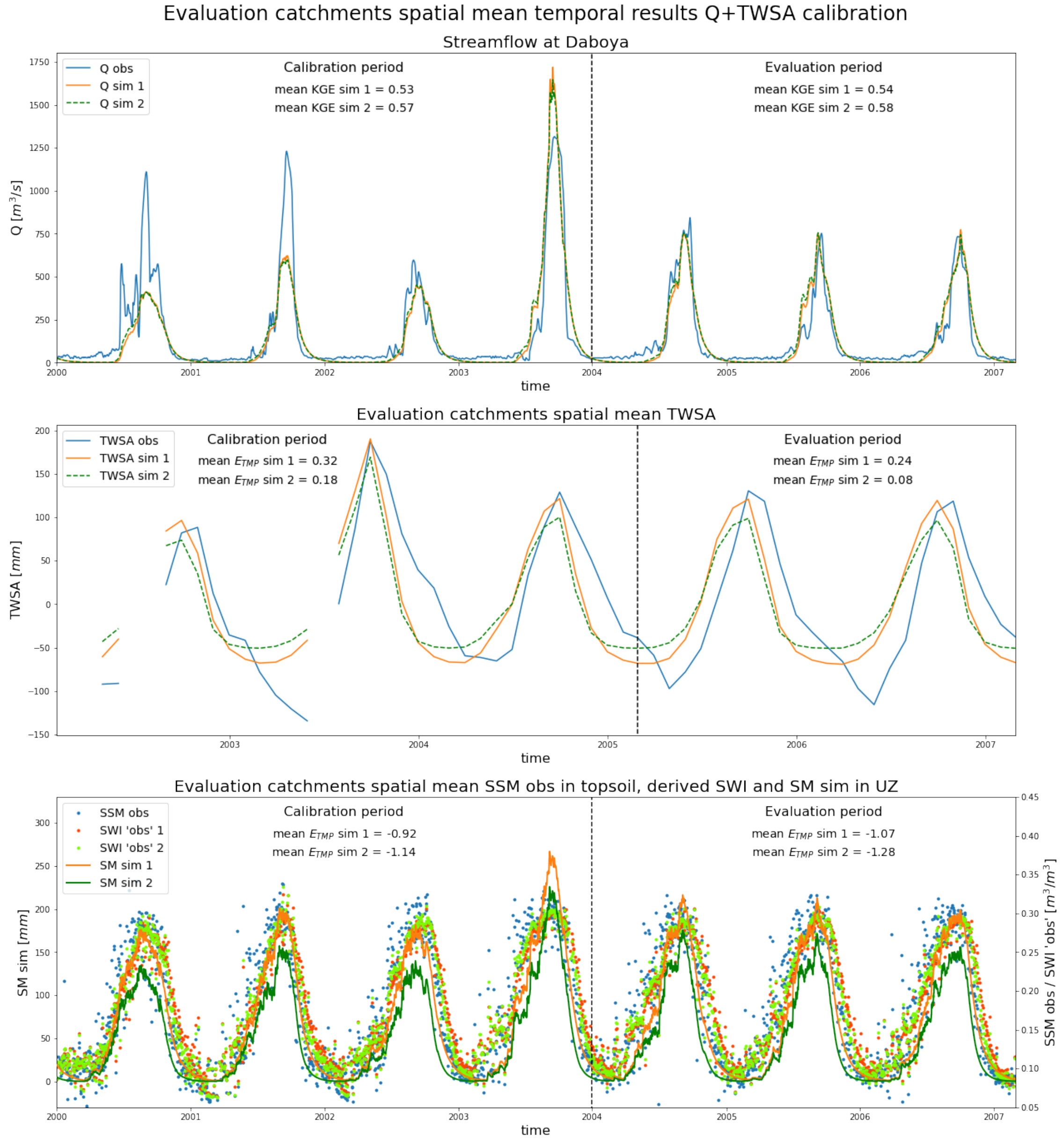


Figure 52: Timeseries of streamflow observations and simulations at Daboya (top). Timeseries of the mean TWSA observations and simulations within the evaluation catchments (middle). Timeseries of the mean surface soil moisture observations and the mean simulated amount of water in the soil moisture reservoir (SM) within the evaluation catchments (bottom).

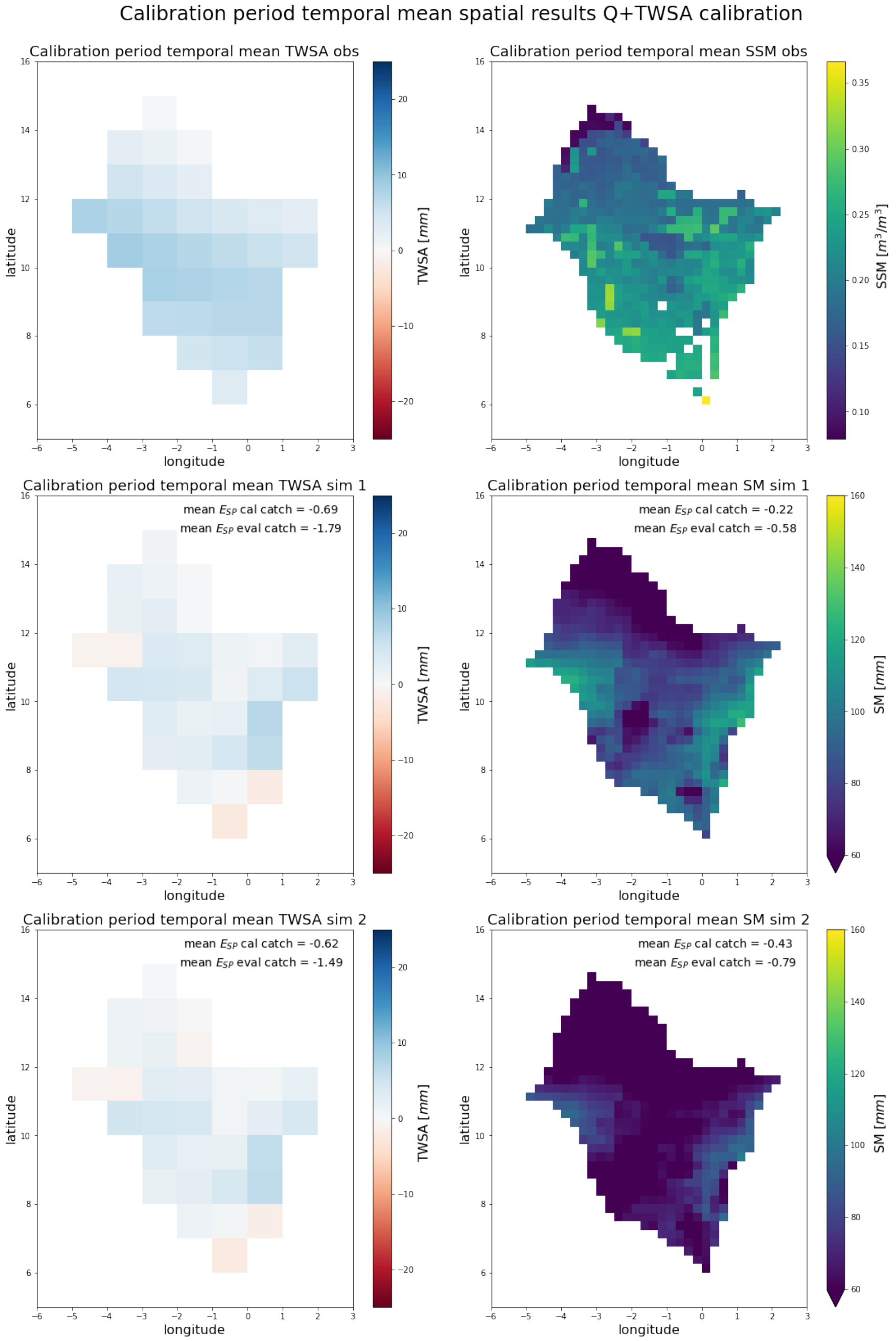

Calibration period temporal mean spatial results Q+TWSA calibration
Calibration period temporal mean TWSA obs
Calibration period temporal mean SSM obs
Calibration period temporal mean TWSA sim 1
mean E_{SP} cal catch = -0.69
mean E_{SP} eval catch = -1.79
Calibration period temporal mean SM sim 1
mean E_{SP} cal catch = -0.22
mean E_{SP} eval catch = -0.58
Calibration period temporal mean TWSA sim 2
mean E_{SP} cal catch = -0.62
mean E_{SP} eval catch = -1.49
Calibration period temporal mean SM sim 2
mean E_{SP} cal catch = -0.43
mean E_{SP} eval catch = -0.79
latitude
longitude
TWSA [mm]
SSM [m^3/m^3]
SM [mm]

Figure 53: Spatial plots of the mean TWSA observations and simulations in the calibration period in the Volta basin (left). Spatial plots of the mean surface soil moisture observations and the mean simulated amount of water in the soil moisture reservoir in the calibration period in the Volta basin (right).

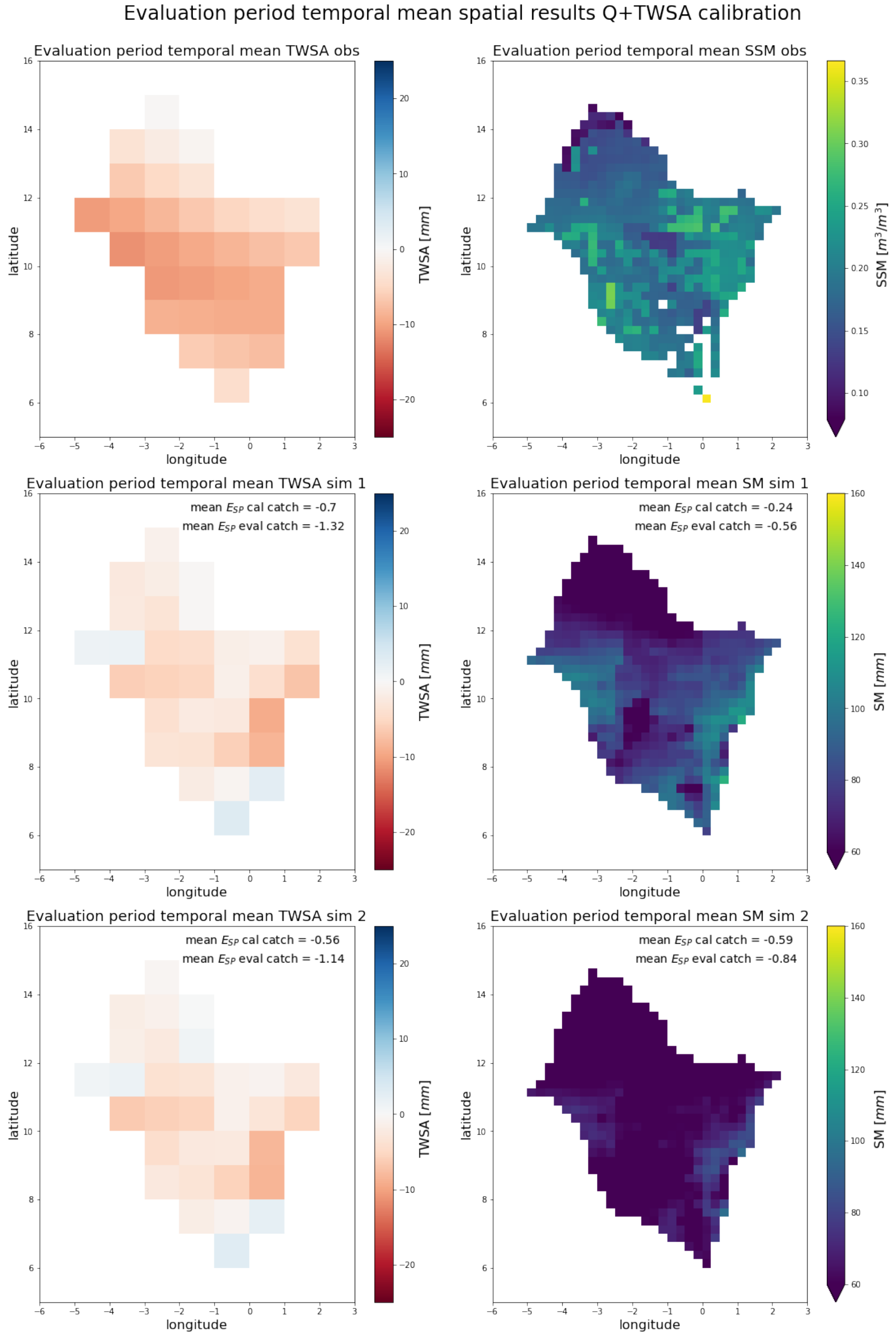

Evaluation period temporal mean spatial results Q+TWSA calibration
Evaluation period temporal mean TWSA obs
Evaluation period temporal mean SSM obs
Evaluation period temporal mean TWSA sim 1
mean E_{SP} cal catch = -0.7
mean E_{SP} eval catch = -1.32
Evaluation period temporal mean SM sim 1
mean E_{SP} cal catch = -0.24
mean E_{SP} eval catch = -0.56
Evaluation period temporal mean TWSA sim 2
mean E_{SP} cal catch = -0.56
mean E_{SP} eval catch = -1.14
Evaluation period temporal mean SM sim 2
mean E_{SP} cal catch = -0.59
mean E_{SP} eval catch = -0.84
latitude
longitude
TWSA [mm]
SSM [m^3/m^3]
SM [mm]

Figure 54: Spatial plots of the mean TWSA observations and simulations in the evaluation period in the Volta basin (left). Spatial plots of the mean surface soil moisture observations and the mean simulated amount of water in the soil moisture reservoir in the evaluation period in the Volta basin (right).

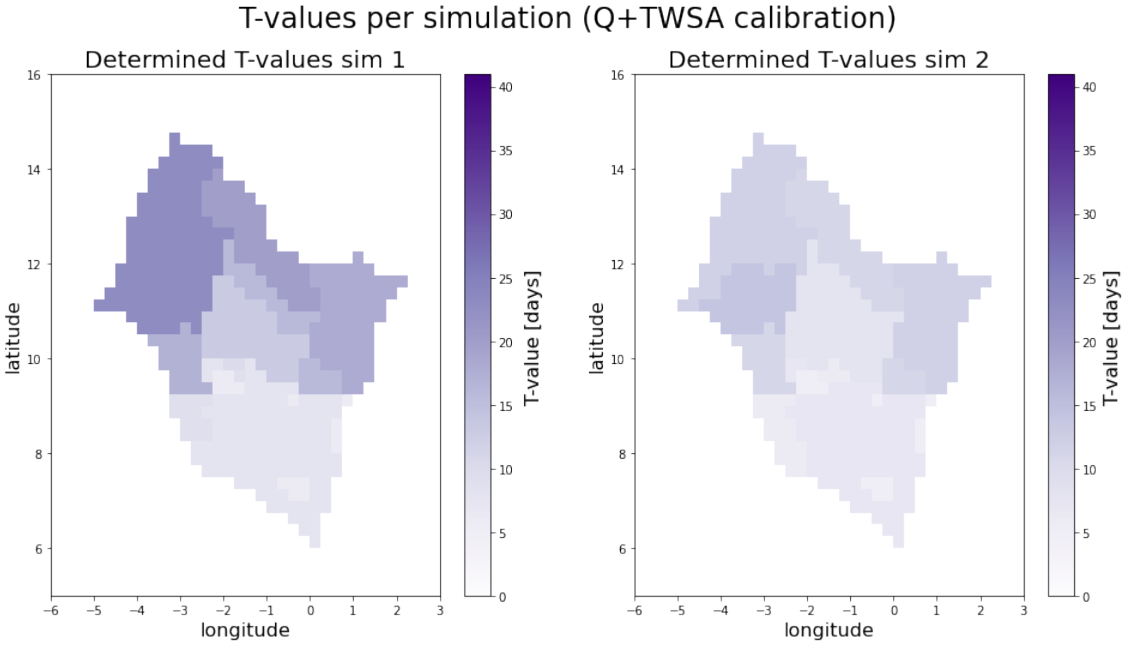

T-values per simulation (Q+TWSA calibration)
Determined T-values sim 1
Determined T-values sim 2
latitude
longitude
T-value [days]
16
14
12
10
8
6
−6
−5
−4
−3
−2
−1
0
1
2
3
0
5
10
15
20
25
30
35
40

Figure 55: The determined T-values per subcatchment / land use class for simulation 1 and 2.

Table 27: Overview of the parameter values found in the 2 calibration runs based on Q+TWSA

Parameter	ICF_{nf}	ICF_f	$CEVPF_{nf}$	$CEVPF_f$	β	LP	$PERC$	Q_{cf}	SUZ	$K0$	$K_{QuickFlow}$
Unit	$[mm]$	$[mm]$	$[-]$	$[-]$	$[-]$	$[-]$	$[mm/d]$	$[mm/d]$	$[mm]$	$[d^{-1}]$	$[d^{-1}]$
Sim 1	0.8	0.95	1.1	1.75	2.7	0.39	3.51	0.23	11.87	0.3	0.75
Sim 2	0.8	1.0	1.1	1.75	2.11	0.26	3.5	0.48	22.07	0.28	0.51
mean	0.8	0.98	1.1	1.75	2.41	0.32	3.51	0.36	16.97	0.29	0.63

J Results Scenario 4: Q+SM+TWSA Calibration

Table 28: Streamflow performance scores per gauging station for simulation 1 calibrated on Q+SM+TWSA

Simulation 1	**Calibration period**						**Evaluation period**					
Stat. / func.	WA_{KGE}	KGE_{OH}	KGE_{BC}	KGE_{FDC}	KGE_{ACF}	KGE_{R_c}	WA_{KGE}	KGE_{OH}	KGE_{BC}	KGE_{FDC}	KGE_{ACF}	KGE_{R_c}
Lawra	0.29	-0.05	0.64	0.81	0.75	-0.39	0.3	0.02	0.63	0.77	0.91	-0.55
Chache	0.44	0.27	0.61	0.64	0.73	0.12	0.26	0.02	0.57	0.6	0.54	-0.21
Bui Amont	0.51	0.41	0.56	0.61	0.84	0.23	0.34	0.22	0.25	0.26	0.7	0.35
Bamboi	-0.04	0.23	-1.09	-0.46	0.54	0.32	-0.94	0.28	-3.47	-3.47	0.21	0.69
Nawuni	0.45	0.8	-0.17	-0.19	0.71	0.77	0.31	0.78	-0.4	-0.41	0.71	0.45
Daboya	0.53	0.65	0.27	0.28	0.63	0.69	0.32	0.41	0.09	0.1	0.72	0.19
Porga	0.64	0.37	0.75	0.77	0.96	0.56	0.54	0.41	0.48	0.51	0.94	0.5
Saboba	0.65	0.84	0.32	0.33	0.83	0.8	0.56	0.84	0.19	0.19	0.54	0.79
Sabari	0.56	0.77	0.31	0.12	0.69	0.72	0.62	0.7	0.46	0.47	0.68	0.69
mean cal catch	**0.52**	0.43	0.53	0.55	0.8	0.34	**0.44**	0.37	0.43	0.47	0.72	0.26
mean eval catch	**0.49**	0.72	0.05	0.05	0.67	0.73	**0.31**	0.6	-0.16	-0.16	0.72	0.32

Stat.: Gauging station. func.: Objective function. cal catch: calibration catchment. eval catch: evaluation catchments.
The performance scores of station Bamboi were not used for the calculation of the means.

Table 29: Streamflow performance scores per gauging station for simulation 2 calibrated on Q+SM+TWSA

Simulation 2	**Calibration period**						**Evaluation period**					
Stat. / func.	WA_{KGE}	KGE_{OH}	KGE_{BC}	KGE_{FDC}	KGE_{ACF}	KGE_{R_c}	WA_{KGE}	KGE_{OH}	KGE_{BC}	KGE_{FDC}	KGE_{ACF}	KGE_{R_c}
Lawra	0.28	-0.07	0.64	0.81	0.75	-0.42	0.29	-0.01	0.64	0.77	0.91	-0.58
Chache	0.43	0.25	0.61	0.63	0.74	0.1	0.25	-0.0	0.57	0.59	0.54	-0.23
Bui Amont	0.5	0.39	0.56	0.6	0.83	0.22	0.32	0.2	0.24	0.25	0.7	0.33
Bamboi	-0.06	0.21	-1.1	-0.48	0.54	0.3	-0.96	0.25	-3.5	-3.5	0.2	0.67
Nawuni	0.45	0.81	-0.16	-0.18	0.71	0.78	0.31	0.78	-0.4	-0.4	0.71	0.46
Daboya	0.52	0.64	0.28	0.29	0.63	0.68	0.31	0.39	0.1	0.1	0.72	0.17
Porga	0.63	0.37	0.75	0.77	0.96	0.56	0.54	0.4	0.48	0.5	0.94	0.49
Saboba	0.66	0.84	0.33	0.33	0.83	0.8	0.56	0.84	0.19	0.2	0.54	0.79
Sabari	0.56	0.77	0.31	0.13	0.69	0.72	0.62	0.7	0.47	0.47	0.68	0.7
mean cal catch	**0.51**	0.42	0.53	0.55	0.8	0.33	**0.43**	0.35	0.43	0.47	0.72	0.25
mean eval catch	**0.49**	0.72	0.06	0.06	0.67	0.73	**0.31**	0.58	-0.15	-0.15	0.71	0.31

Stat.: Gauging station. func.: Objective function. cal catch: calibration catchment. eval catch: evaluation catchments.
The performance scores of station Bamboi were not used for the calculation of the means.

Table 30: Spatial and temporal mean RS performance scores for simulation 1, calibrated on Q+SM+TWSA.

Simulation 1	**TWSA**	**SM**
$E_{SP_{cal,cal}}$	-0.9	-0.04
$E_{SP_{cal,eval}}$	-1.03	-0.0
$E_{SP_{eval,cal}}$	-2.26	-0.41
$E_{SP_{eval,eval}}$	-1.78	-0.33
$E_{TMP_{cal,cal}}$	0.44	-0.46
$E_{TMP_{cal,eval}}$	0.39	-0.6
$E_{TMP_{eval,cal}}$	0.47	-0.49
$E_{TMP_{eval,eval}}$	0.4	-0.68

First cal/eval in subscript indicates catchment, second cal/eval in subscript indicates period.

Table 31: Spatial and temporal mean RS performance scores for simulation 2, calibrated on Q+SM+TWSA.

Simulation 2	**TWSA**	**SM**
$E_{SP_{cal,cal}}$	-0.86	-0.04
$E_{SP_{cal,eval}}$	-0.98	-0.01
$E_{SP_{eval,cal}}$	-2.2	-0.42
$E_{SP_{eval,eval}}$	-1.74	-0.35
$E_{TMP_{cal,cal}}$	0.45	-0.51
$E_{TMP_{cal,eval}}$	0.41	-0.65
$E_{TMP_{eval,cal}}$	0.47	-0.54
$E_{TMP_{eval,eval}}$	0.41	-0.74

First cal/eval in subscript indicates catchment, second cal/eval in subscript indicates period.

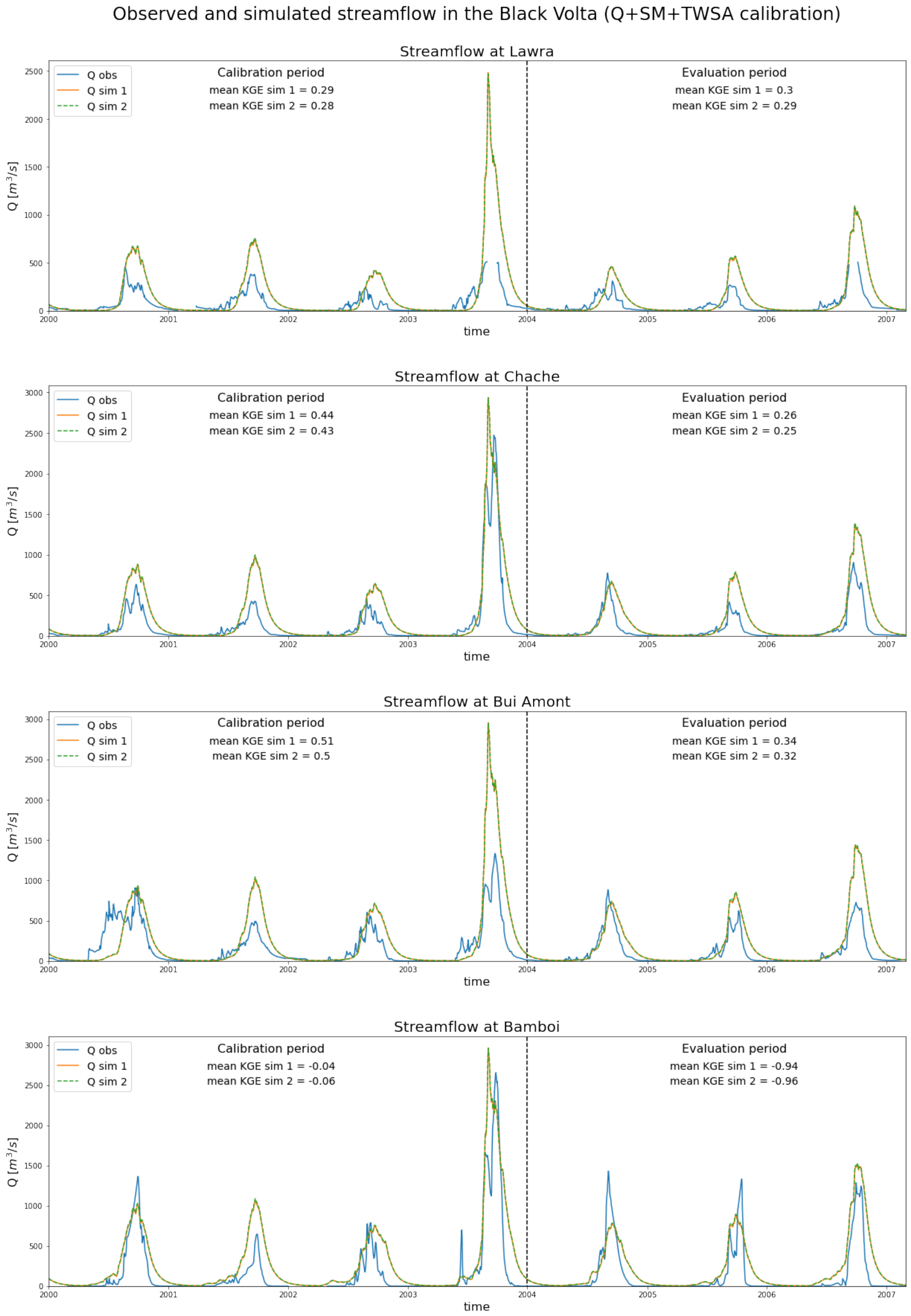


Figure 56: Hydrographs of the stations in the Black Volta based on calibration on Q+SM+TWSA. These stations are calibration stations. The stations Boromo and Bamboi were excluded from calibration because of low data quality at Boromo, and far too high flow observations at Bamboi. The streamflow observations at Bamboi shown in this plot are scaled by catchment area to match the flow at the upstream station Bui Amont. The mean KGE values given in the plots are mean values of the 5 objective functions used for calibration as was explained in subsubsection 3.3.2.

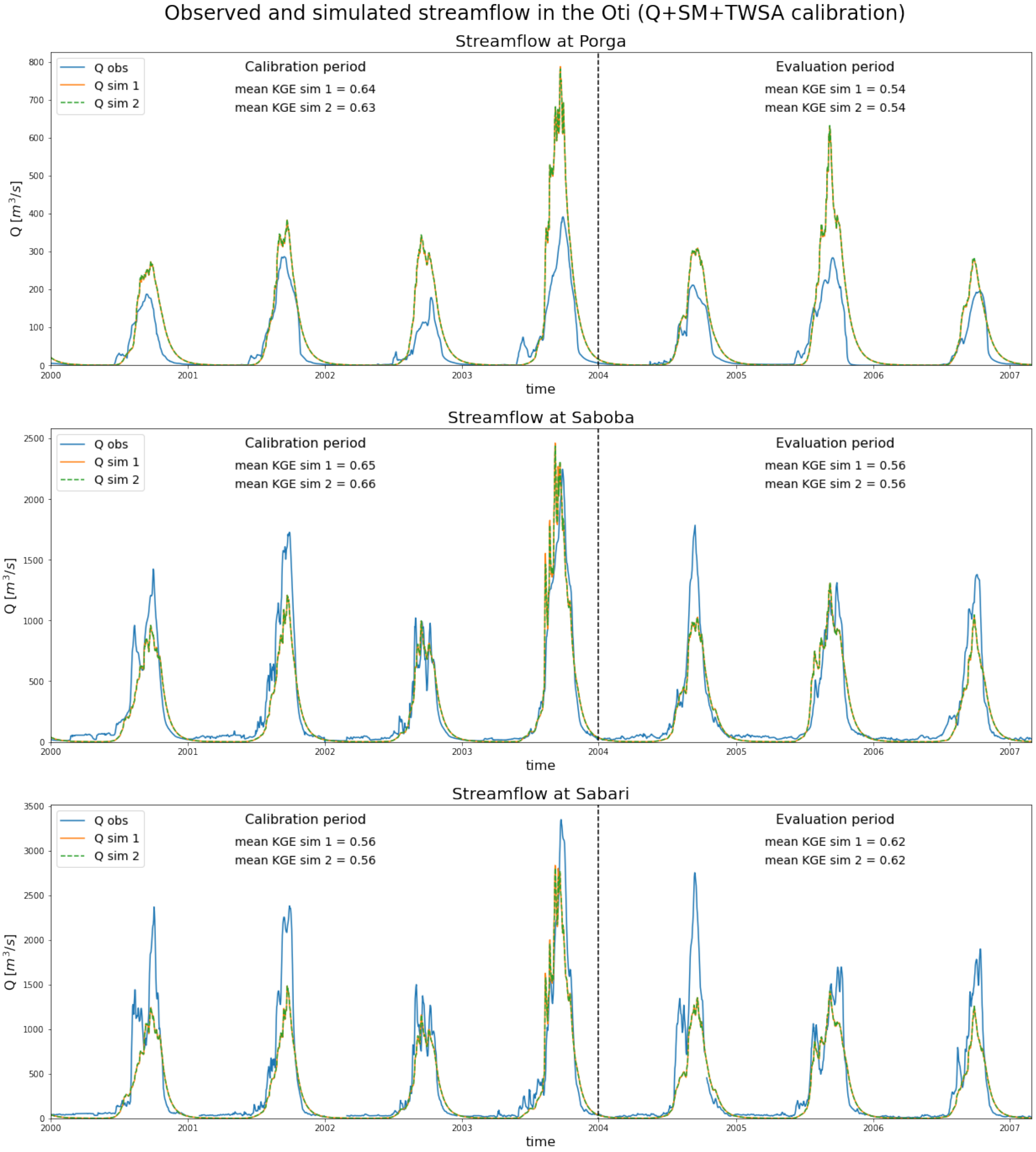


Figure 57: Hydrographs of the stations in the Oti based on calibration on Q+SM+TWSA. These stations are calibration stations. The mean KGE values given in the plots are mean values of the 5 objective functions used for calibration as was explained in subsubsection 3.3.2.

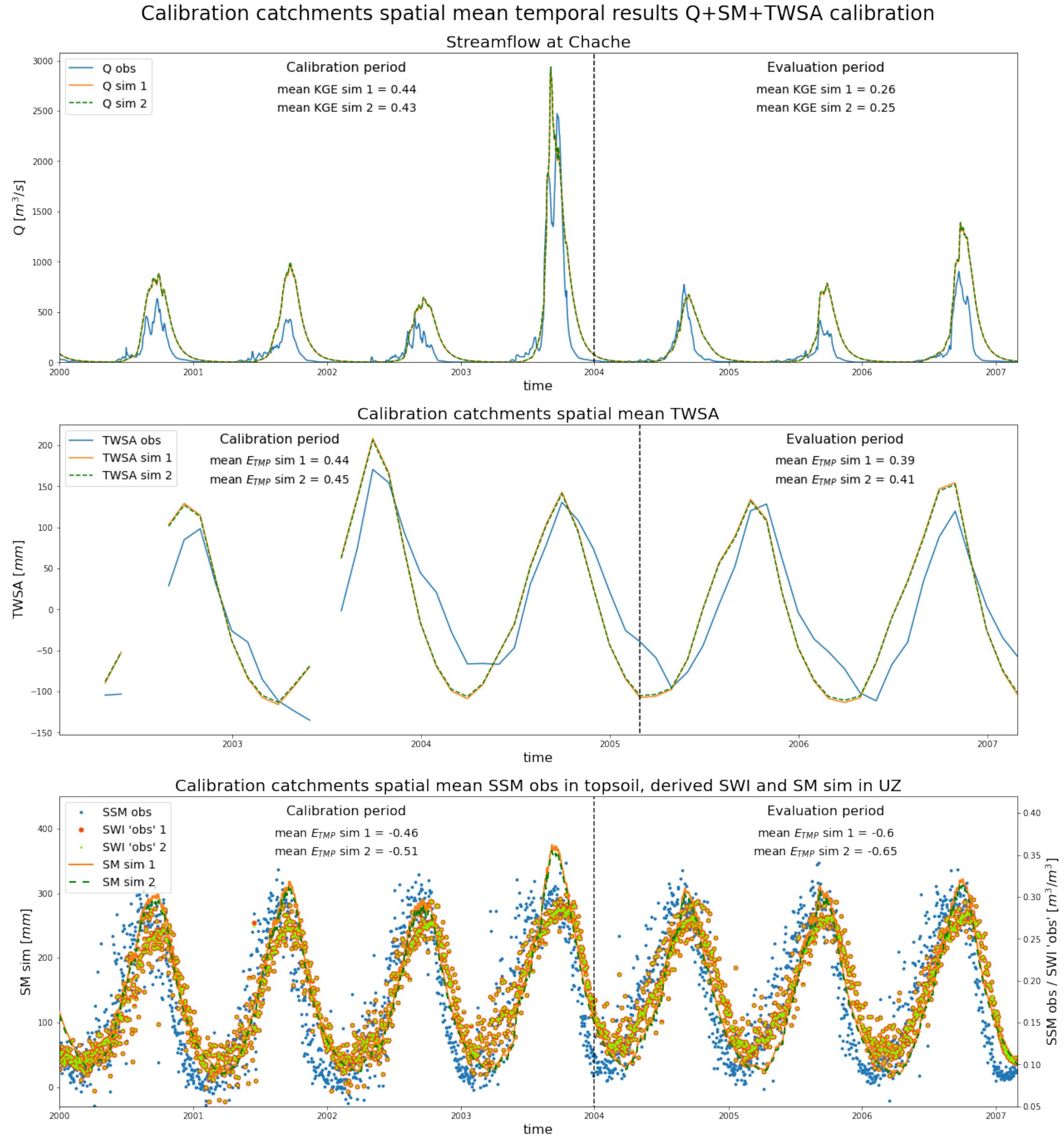


Figure 58: Timeseries of streamflow observations and simulations at Chache (top). Timeseries of the mean TWSA observations and simulations within the calibration catchments (middle). Timeseries of the mean surface soil moisture observations and the mean simulated amount of water in the soil moisture reservoir (SM) within the calibration catchments (bottom).

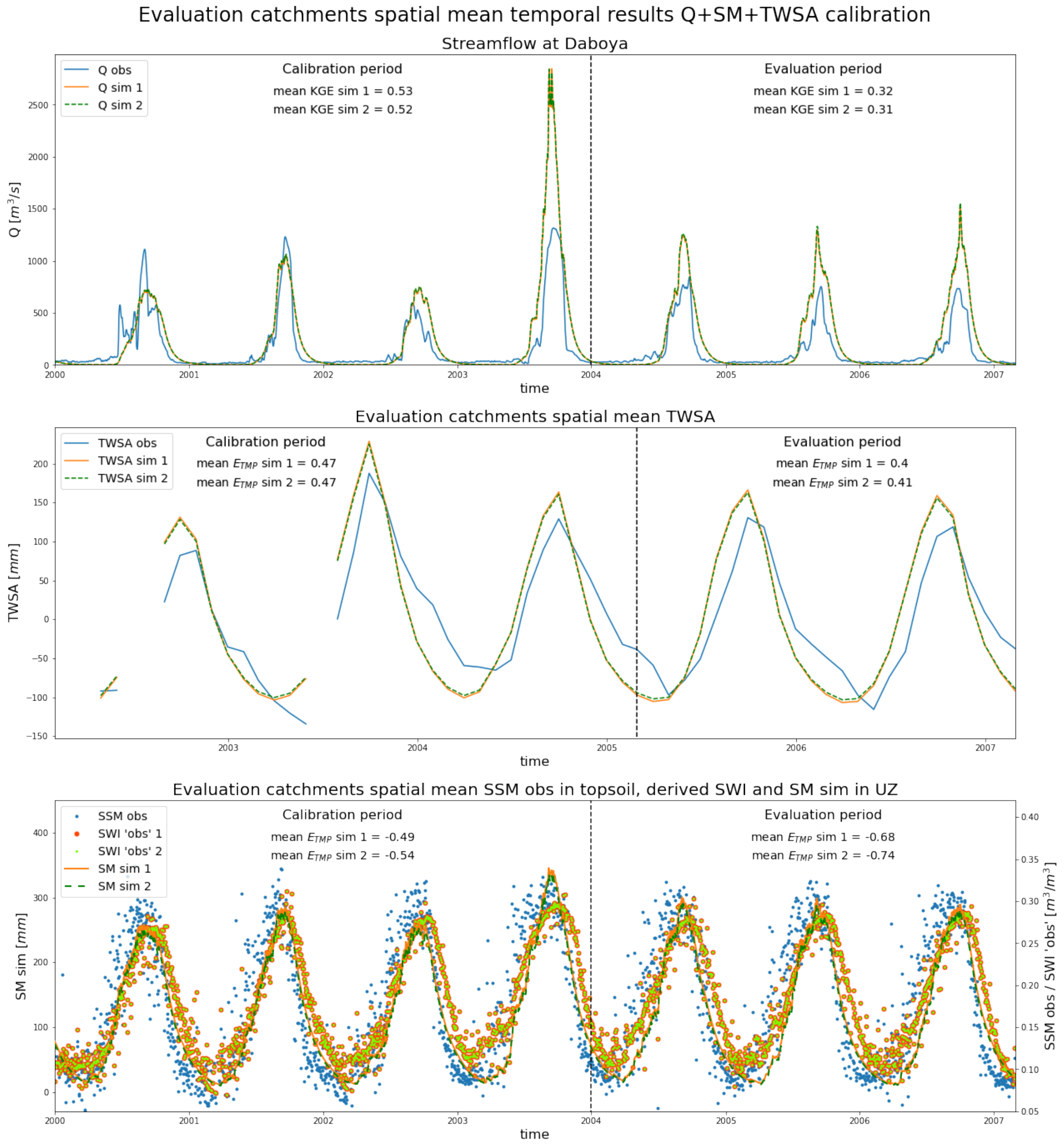


Figure 59: Timeseries of streamflow observations and simulations at Daboya (top). Timeseries of the mean TWSA observations and simulations within the evaluation catchments (middle). Timeseries of the mean surface soil moisture observations and the mean simulated amount of water in the soil moisture reservoir (SM) within the evaluation catchments (bottom).

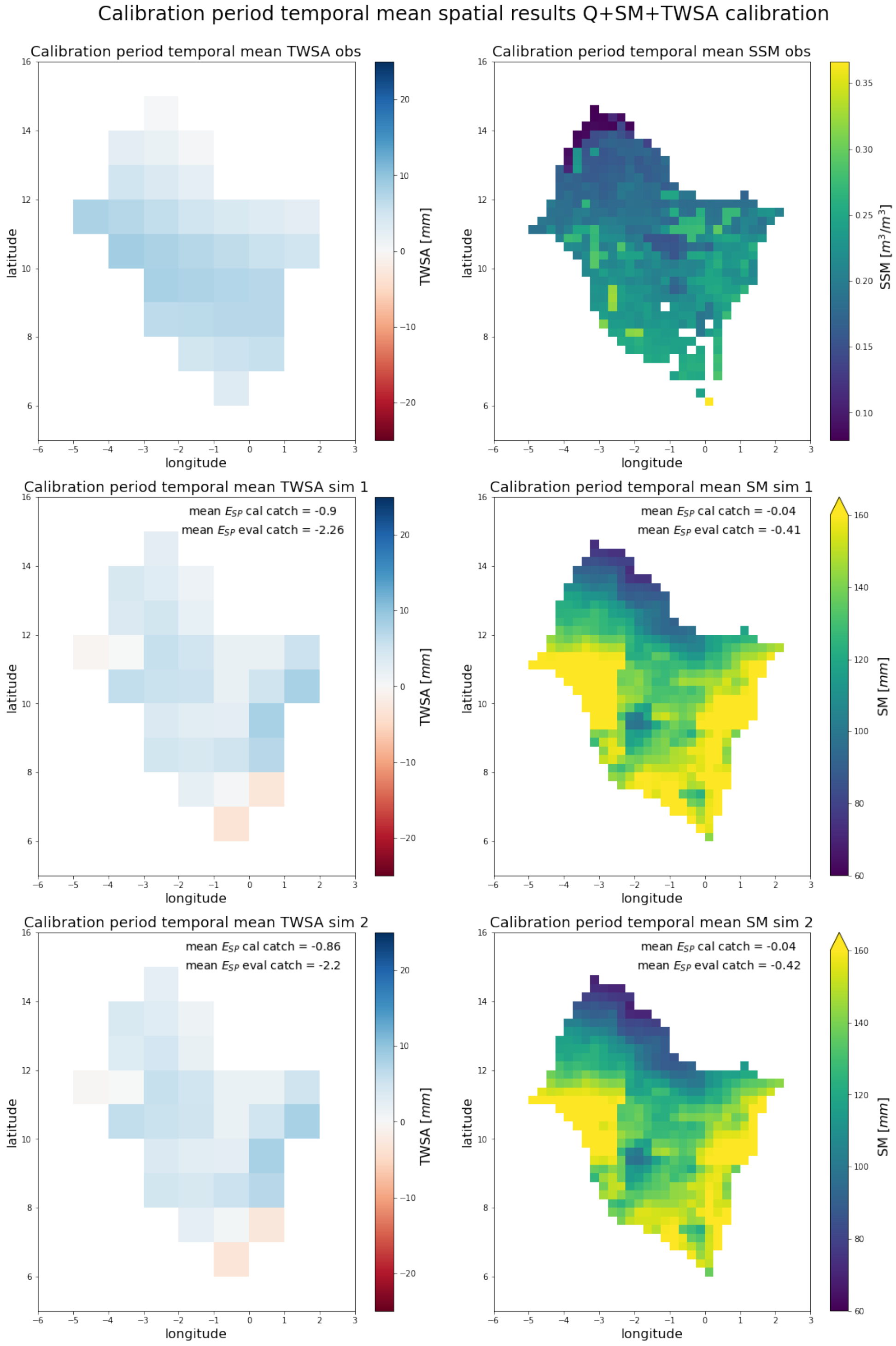
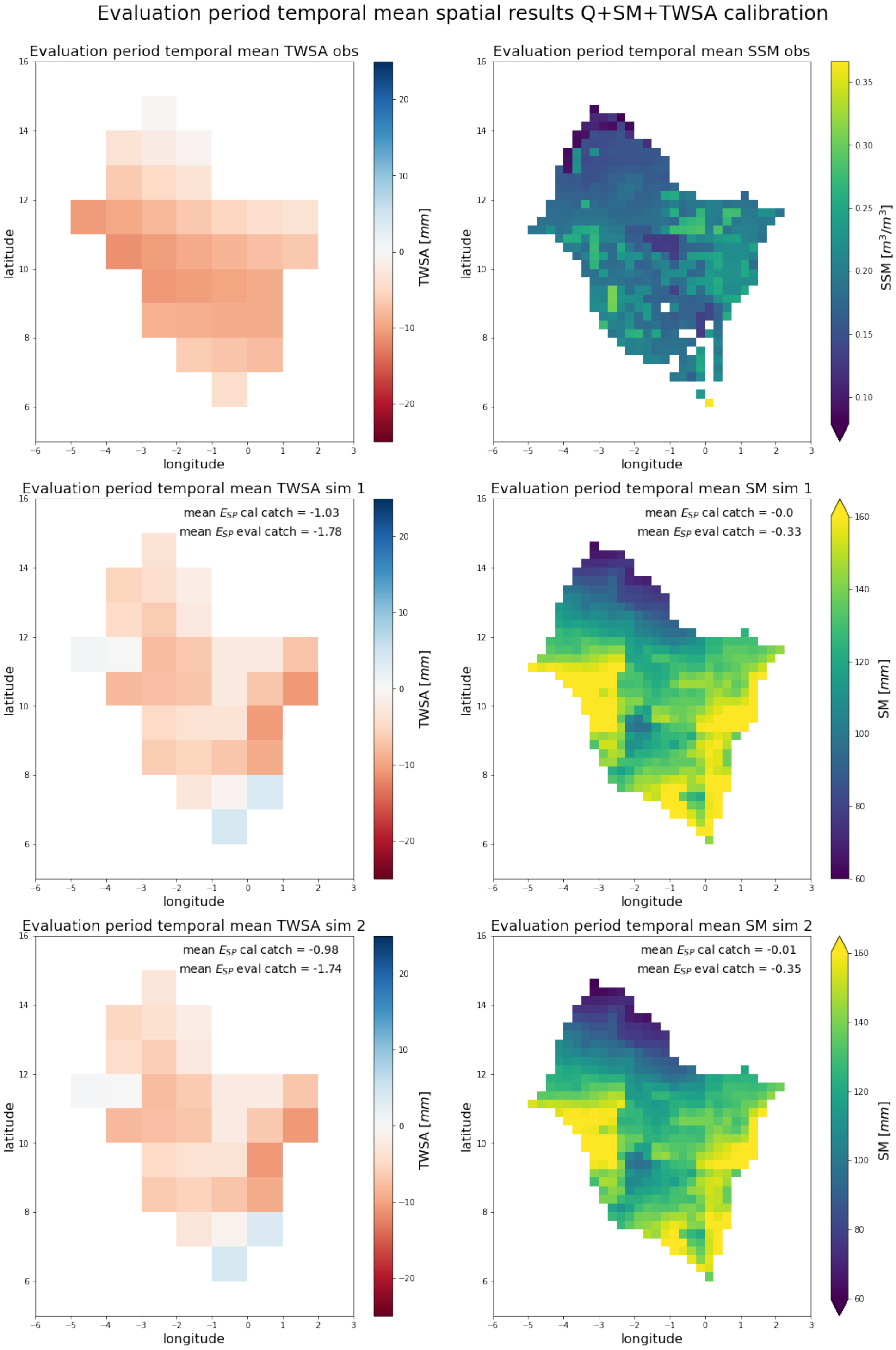

Calibration period temporal mean spatial results Q+SM+TWSA calibration
Calibration period temporal mean TWSA obs
Calibration period temporal mean SSM obs
Calibration period temporal mean TWSA sim 1
mean E_{SP} cal catch = -0.9
mean E_{SP} eval catch = -2.26
Calibration period temporal mean SM sim 1
mean E_{SP} cal catch = -0.04
mean E_{SP} eval catch = -0.41
Calibration period temporal mean TWSA sim 2
mean E_{SP} cal catch = -0.86
mean E_{SP} eval catch = -2.2
Calibration period temporal mean SM sim 2
mean E_{SP} cal catch = -0.04
mean E_{SP} eval catch = -0.42
latitude
longitude
TWSA $[mm]$
SSM $[m^3/m^3]$
SM $[mm]$

Figure 60: Spatial plots of the mean TWSA observations and simulations in the calibration period in the Volta basin (left). Spatial plots of the mean surface soil moisture observations and the mean simulated amount of water in the soil moisture reservoir in the calibration period in the Volta basin (right).



Evaluation period temporal mean spatial results Q+SM+TWSA calibration
Evaluation period temporal mean TWSA obs
Evaluation period temporal mean SSM obs
Evaluation period temporal mean TWSA sim 1
mean E_{SP} cal catch = -1.03
mean E_{SP} eval catch = -1.78
Evaluation period temporal mean SM sim 1
mean E_{SP} cal catch = -0.0
mean E_{SP} eval catch = -0.33
Evaluation period temporal mean TWSA sim 2
mean E_{SP} cal catch = -0.98
mean E_{SP} eval catch = -1.74
Evaluation period temporal mean SM sim 2
mean E_{SP} cal catch = -0.01
mean E_{SP} eval catch = -0.35
latitude
longitude
TWSA $[mm]$
SSM $[m^3/m^3]$
SM $[mm]$

Figure 61: Spatial plots of the mean TWSA observations and simulations in the evaluation period in the Volta basin (left). Spatial plots of the mean surface soil moisture observations and the mean simulated amount of water in the soil moisture reservoir in the evaluation period in the Volta basin (right).

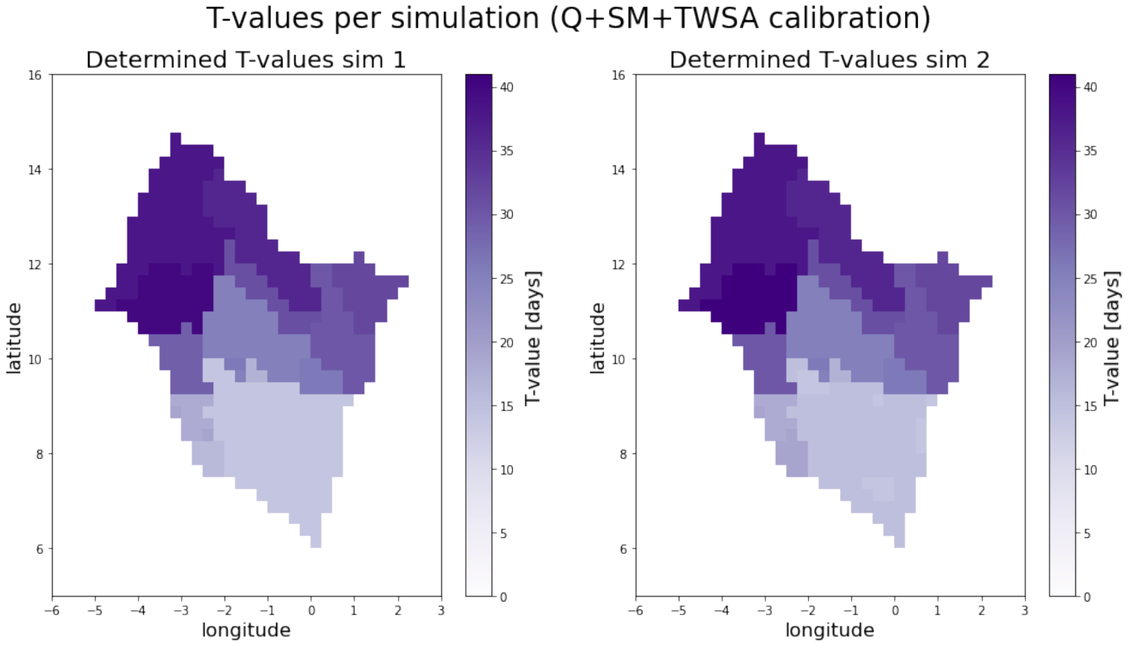

T-values per simulation (Q+SM+TWSA calibration)
Determined T-values sim 1
Determined T-values sim 2
latitude
longitude
T-value [days]
16
14
12
10
8
6
−6
−5
−4
−3
−2
−1
0
1
2
3
40
35
30
25
20
15
10
5
0

Figure 62: The determined T-values per subcatchment / land use class for simulation 1 and 2.

Table 32: Overview of the parameter values found in the 2 calibration runs based on Q+SM+TWSA

Parameter	ICF_{nf}	ICF_f	$CEVPF_{nf}$	$CEVPF_f$	β	LP	$PERC$	Q_{cf}	SUZ	$K0$	$K_{QuickFlow}$
Unit	$[mm]$	$[mm]$	$[-]$	$[-]$	$[-]$	$[-]$	$[mm/d]$	$[mm/d]$	$[mm]$	$[d^{-1}]$	$[d^{-1}]$
Sim 1	0.48	1.18	0.9	1.11	3.42	0.55	3.51	1.17	16.5	0.3	0.9
Sim 2	0.74	1.19	0.9	1.1	3.29	0.55	3.5	2.1	17.8	0.3	0.45
mean	0.61	1.18	0.9	1.1	3.36	0.55	3.5	1.64	17.15	0.3	0.68

K Results Scenario 5: SM+TWSA Calibration

Table 33: Streamflow performance scores per gauging station for simulation 1 calibrated on SM+TWSA

Simulation 1	**Calibration period**						**Evaluation period**					
Stat. / func.	WA_{KGE}	KGE_{OH}	KGE_{BC}	KGE_{FDC}	KGE_{ACF}	KGE_{R_c}	WA_{KGE}	KGE_{OH}	KGE_{BC}	KGE_{FDC}	KGE_{ACF}	KGE_{R_c}
Lawra	-0.75	-2.17	0.4	0.55	0.92	-2.22	-1.06	-2.45	0.38	0.4	0.71	-3.09
Chache	-0.41	-1.13	0.15	0.16	0.61	-1.21	-0.96	-2.09	0.09	0.09	0.41	-2.29
Bui Amont	-0.22	-0.88	0.15	0.16	0.96	-0.93	-0.79	-1.74	-0.31	-0.3	0.55	-1.31
Bamboi	-1.04	-1.23	-1.99	-1.24	0.45	-1.0	-2.41	-1.46	-5.42	-5.42	0.05	-0.64
Nawuni	0.4	0.44	0.24	0.22	0.54	0.55	0.14	0.15	-0.02	-0.02	0.53	0.04
Daboya	0.16	-0.27	0.52	0.53	0.46	-0.04	-0.21	-0.84	0.33	0.33	0.54	-0.84
Porga	0.06	-0.64	0.38	0.38	0.88	-0.08	-0.11	-0.69	0.08	0.09	0.79	-0.28
Saboba	0.71	0.7	0.67	0.68	0.68	0.85	0.58	0.63	0.54	0.55	0.42	0.74
Sabari	0.71	0.78	0.69	0.57	0.55	0.89	0.77	0.79	0.8	0.81	0.53	0.91
mean cal catch	**0.02**	-0.56	0.41	0.42	0.77	-0.45	**-0.26**	-0.92	0.26	0.27	0.57	-0.89
mean eval catch	**0.28**	0.08	0.38	0.38	0.5	0.26	**-0.04**	-0.35	0.15	0.15	0.53	-0.4

Stat.: Gauging station. func.: Objective function. cal catch: calibration catchment. eval catch: evaluation catchments.
The performance scores of station Bamboi were not used for the calculation of the means.

Table 34: Streamflow performance scores per gauging station for simulation 2 calibrated on SM+TWSA

Simulation 2	**Calibration period**						**Evaluation period**					
Stat. / func.	WA_{KGE}	KGE_{OH}	KGE_{BC}	KGE_{FDC}	KGE_{ACF}	KGE_{R_c}	WA_{KGE}	KGE_{OH}	KGE_{BC}	KGE_{FDC}	KGE_{ACF}	KGE_{R_c}
Lawra	-0.77	-2.19	0.4	0.54	0.92	-2.25	-1.07	-2.48	0.38	0.4	0.71	-3.12
Chache	-0.42	-1.15	0.15	0.15	0.61	-1.22	-0.97	-2.11	0.08	0.08	0.41	-2.31
Bui Amont	-0.23	-0.89	0.15	0.16	0.96	-0.94	-0.8	-1.75	-0.31	-0.31	0.55	-1.32
Bamboi	-1.04	-1.24	-2.0	-1.24	0.45	-1.01	-2.42	-1.48	-5.43	-5.43	0.05	-0.65
Nawuni	0.4	0.44	0.24	0.23	0.54	0.54	0.13	0.14	-0.01	-0.02	0.53	0.03
Daboya	0.16	-0.28	0.52	0.53	0.46	-0.04	-0.21	-0.84	0.33	0.33	0.54	-0.85
Porga	0.05	-0.65	0.37	0.38	0.88	-0.09	-0.12	-0.71	0.08	0.08	0.79	-0.3
Saboba	0.71	0.69	0.67	0.68	0.68	0.85	0.58	0.62	0.54	0.55	0.42	0.73
Sabari	0.71	0.78	0.7	0.57	0.55	0.89	0.77	0.79	0.8	0.81	0.53	0.91
mean cal catch	**0.01**	-0.57	0.41	0.41	0.76	-0.46	**-0.27**	-0.94	0.26	0.27	0.57	-0.9
mean eval catch	**0.28**	0.08	0.38	0.38	0.5	0.25	**-0.04**	-0.35	0.16	0.16	0.53	-0.41

Stat.: Gauging station. func.: Objective function. cal catch: calibration catchment. eval catch: evaluation catchments.
The performance scores of station Bamboi were not used for the calculation of the means.

Table 35: Spatial and temporal mean RS performance scores for simulation 1, calibrated on SM+TWSA.

Simulation 1	**TWSA**	**SM**
$E_{SP_{cal,cal}}$	-0.79	-0.06
$E_{SP_{cal,eval}}$	-0.81	-0.02
$E_{SP_{eval,cal}}$	-2.04	-0.42
$E_{SP_{eval,eval}}$	-1.48	-0.35
$E_{TMP_{cal,cal}}$	0.45	-0.45
$E_{TMP_{cal,eval}}$	0.41	-0.59
$E_{TMP_{eval,cal}}$	0.45	-0.46
$E_{TMP_{eval,eval}}$	0.4	-0.66

First cal/eval in subscript indicates catchment, second cal/eval in subscript indicates period.

Table 36: Spatial and temporal mean RS performance scores for simulation 2, calibrated on SM+TWSA.

Simulation 2	**TWSA**	**SM**
$E_{SP_{cal,cal}}$	-0.79	-0.06
$E_{SP_{cal,eval}}$	-0.81	-0.02
$E_{SP_{eval,cal}}$	-2.06	-0.42
$E_{SP_{eval,eval}}$	-1.5	-0.35
$E_{TMP_{cal,cal}}$	0.45	-0.43
$E_{TMP_{cal,eval}}$	0.41	-0.57
$E_{TMP_{eval,cal}}$	0.45	-0.45
$E_{TMP_{eval,eval}}$	0.4	-0.64

First cal/eval in subscript indicates catchment, second cal/eval in subscript indicates period.

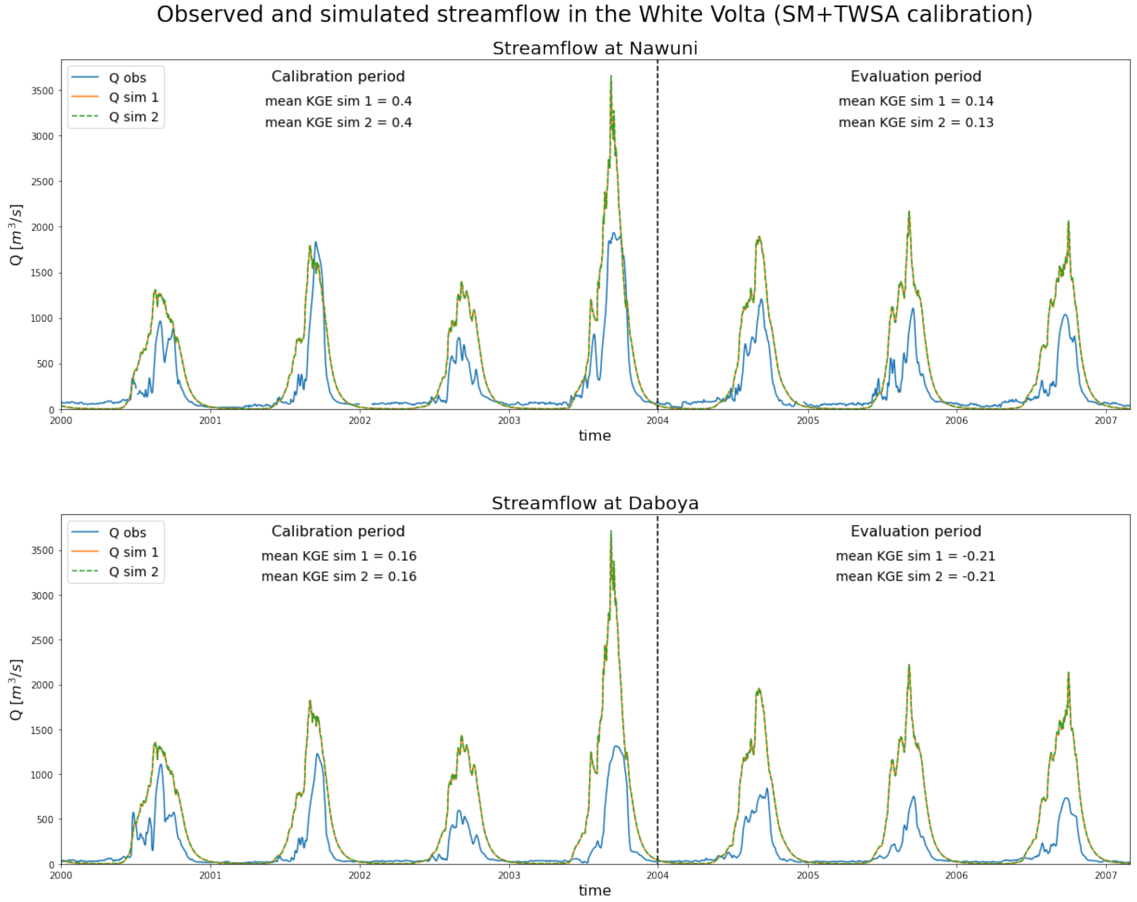


Figure 63: Hydrographs of the stations in the White Volta based on calibration on SM+TWSA. These stations are evaluation stations. The stations Yaragu and Pwalagu were excluded from evaluation because of the high influence of reservoir management on the river flow in this area. The mean KGE values given in the plots are mean values of the 5 objective functions used for calibration as was explained in subsubsection 3.3.2.

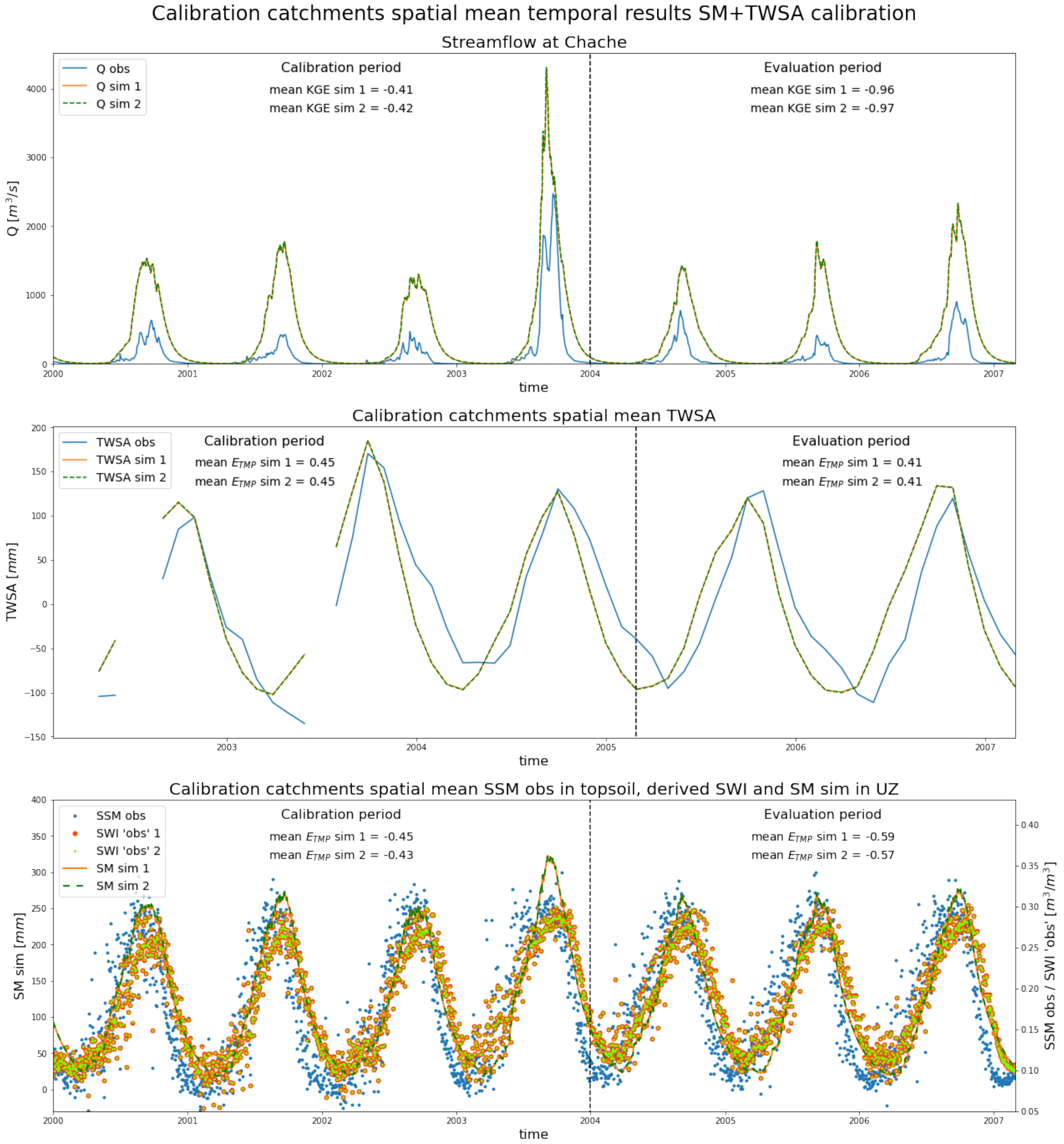


Figure 64: Timeseries of streamflow observations and simulations at Chache (top). Timeseries of the mean TWSA observations and simulations within the calibration catchments (middle). Timeseries of the mean surface soil moisture observations and the mean simulated amount of water in the soil moisture reservoir (SM) within the calibration catchments (bottom).

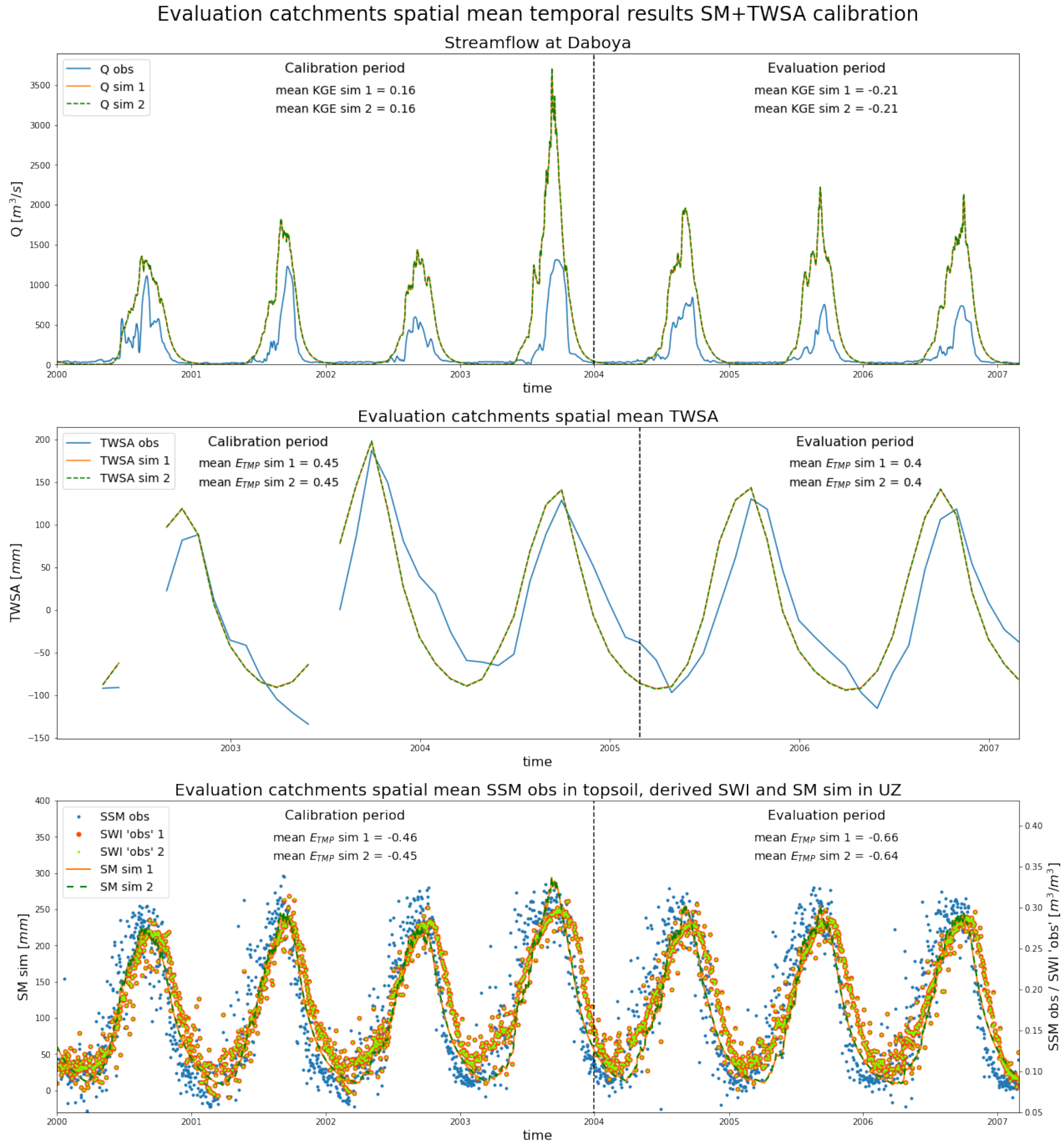


Figure 65: Timeseries of streamflow observations and simulations at Daboya (top). Timeseries of the mean TWSA observations and simulations within the evaluation catchments (middle). Timeseries of the mean surface soil moisture observations and the mean simulated amount of water in the soil moisture reservoir (SM) within the evaluation catchments (bottom).

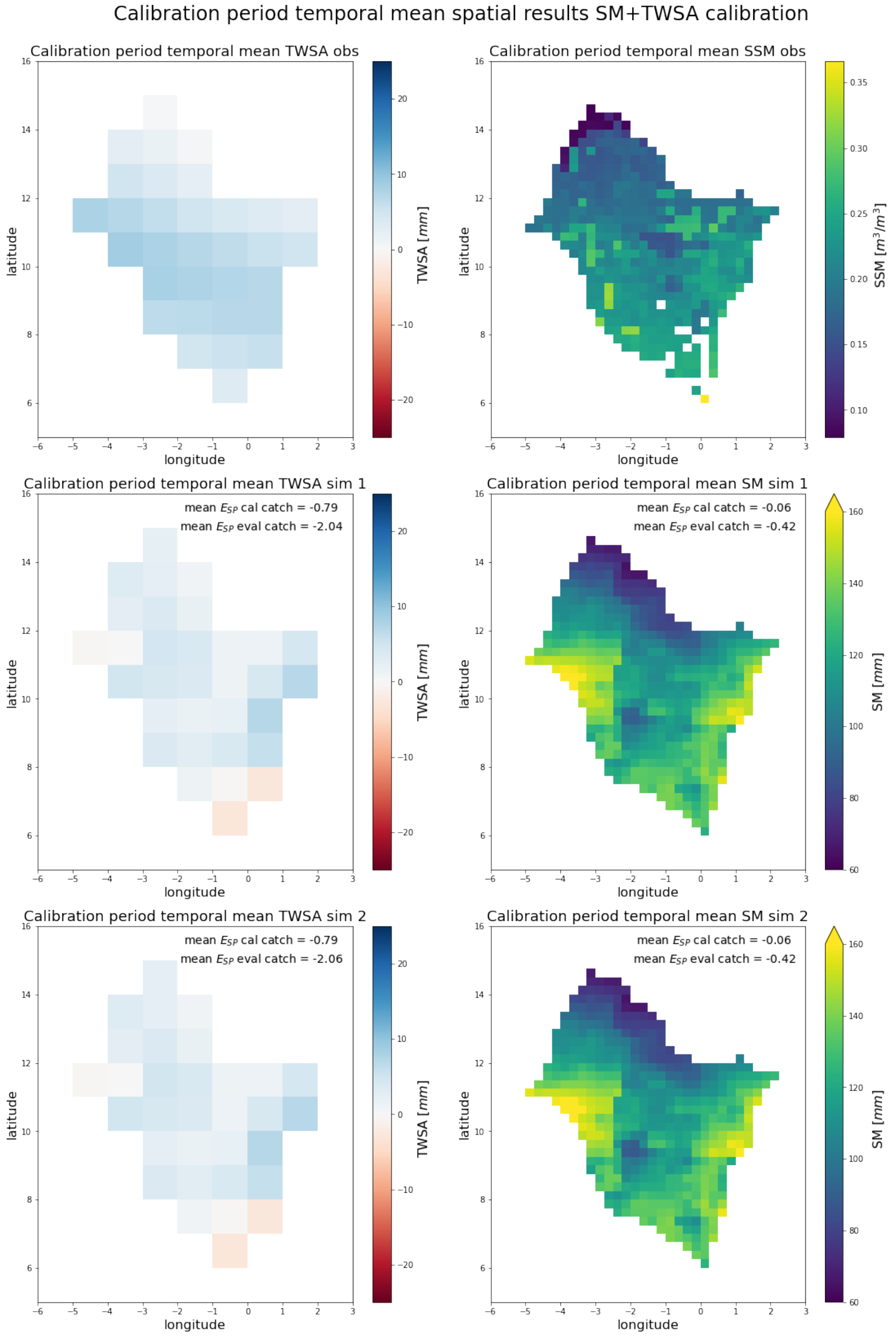

Calibration period temporal mean spatial results SM+TWSA calibration
Calibration period temporal mean TWSA obs
Calibration period temporal mean SSM obs
Calibration period temporal mean TWSA sim 1
mean E_{SP} cal catch = -0.79
mean E_{SP} eval catch = -2.04
Calibration period temporal mean SM sim 1
mean E_{SP} cal catch = -0.06
mean E_{SP} eval catch = -0.42
Calibration period temporal mean TWSA sim 2
mean E_{SP} cal catch = -0.79
mean E_{SP} eval catch = -2.06
Calibration period temporal mean SM sim 2
mean E_{SP} cal catch = -0.06
mean E_{SP} eval catch = -0.42
latitude
longitude
TWSA $[mm]$
SSM $[m^3/m^3]$
SM $[mm]$
16
14
12
10
8
6
−6
−5
−4
−3
−2
−1
0
1
2
3
20
10
0
−10
−20
0.35
0.30
0.25
0.20
0.15
0.10
160
140
120
100
80
60

Figure 66: Spatial plots of the mean TWSA observations and simulations in the calibration period in the Volta basin (left). Spatial plots of the mean surface soil moisture observations and the mean simulated amount of water in the soil moisture reservoir in the calibration period in the Volta basin (right).

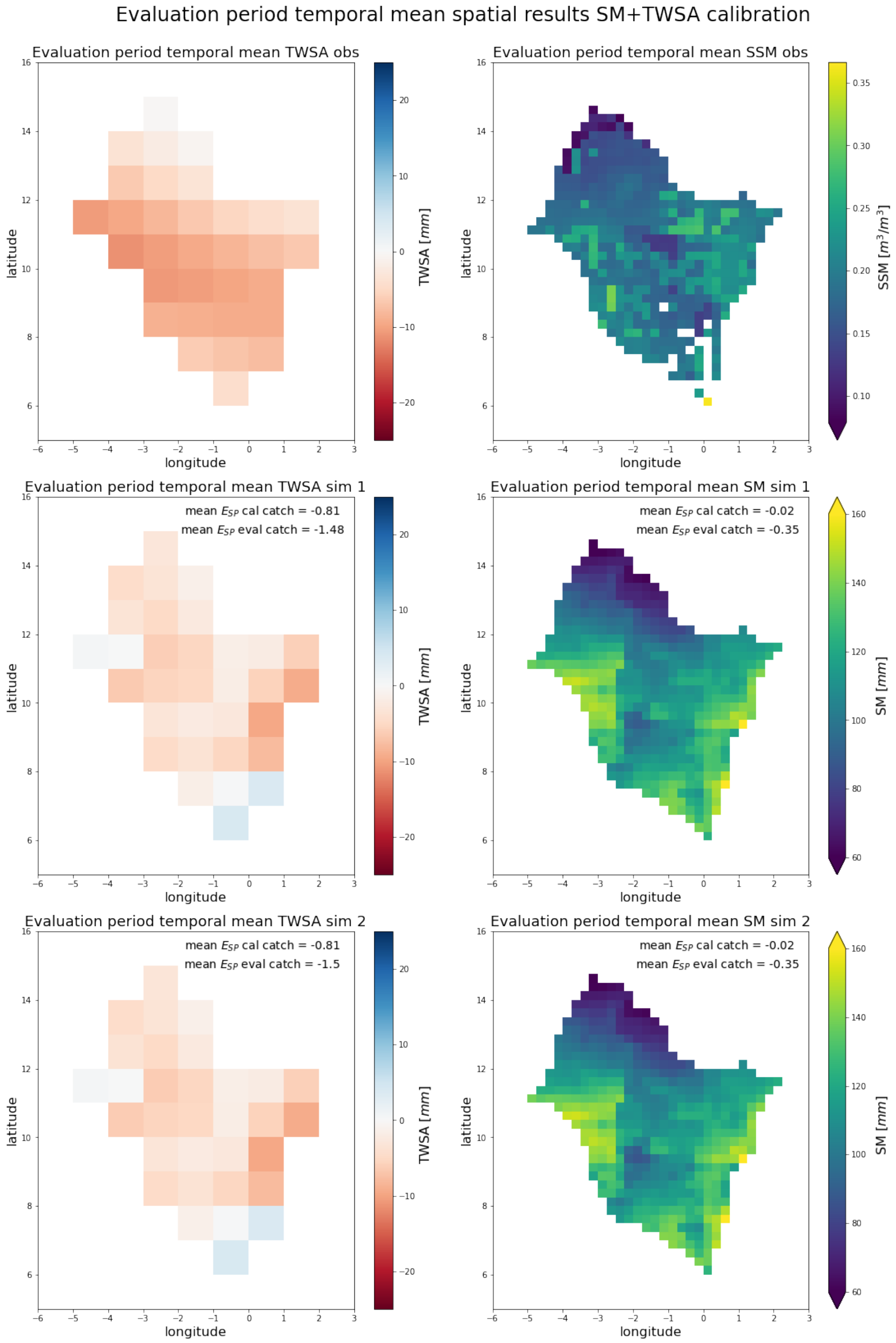

Evaluation period temporal mean spatial results SM+TWSA calibration
Evaluation period temporal mean TWSA obs
Evaluation period temporal mean SSM obs
Evaluation period temporal mean TWSA sim 1
mean E_{SP} cal catch = -0.81
mean E_{SP} eval catch = -1.48
Evaluation period temporal mean SM sim 1
mean E_{SP} cal catch = -0.02
mean E_{SP} eval catch = -0.35
Evaluation period temporal mean TWSA sim 2
mean E_{SP} cal catch = -0.81
mean E_{SP} eval catch = -1.5
Evaluation period temporal mean SM sim 2
mean E_{SP} cal catch = -0.02
mean E_{SP} eval catch = -0.35
latitude
longitude
TWSA $[mm]$
SSM $[m^3/m^3]$
SM $[mm]$
16
14
12
10
8
6
−6
−5
−4
−3
−2
−1
0
1
2
3
20
10
0
−10
−20
0.35
0.30
0.25
0.20
0.15
0.10
160
140
120
100
80
60

Figure 67: Spatial plots of the mean TWSA observations and simulations in the evaluation period in the Volta basin (left). Spatial plots of the mean surface soil moisture observations and the mean simulated amount of water in the soil moisture reservoir in the evaluation period in the Volta basin (right).

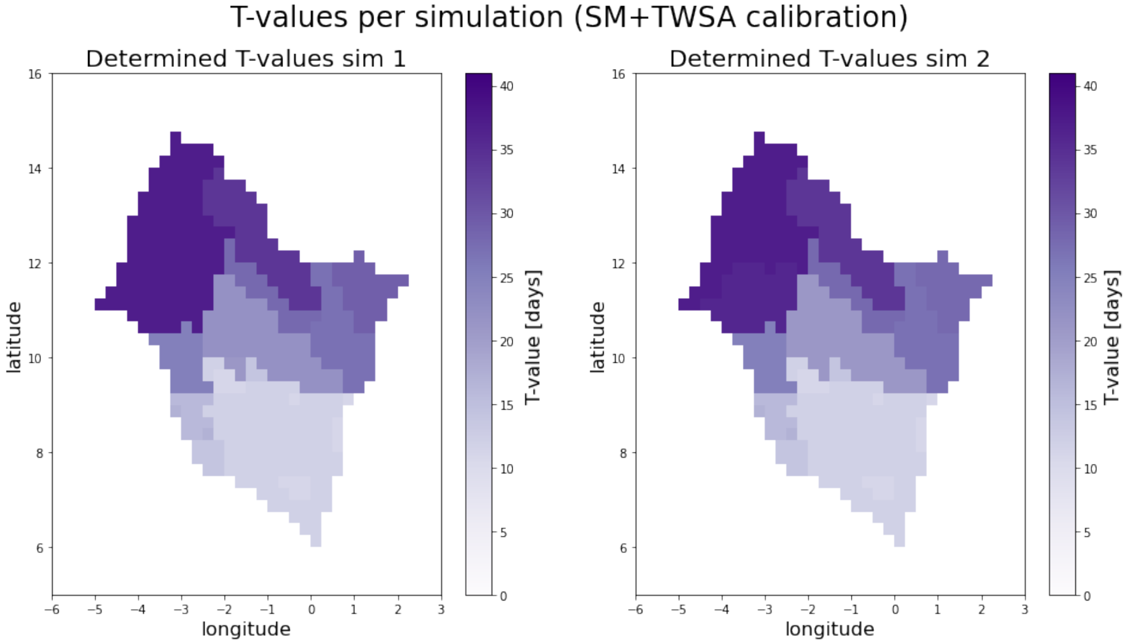

T-values per simulation (SM+TWSA calibration)
Determined T-values sim 1
Determined T-values sim 2
latitude
longitude
T-value [days]
16
14
12
10
8
6
−6
−5
−4
−3
−2
−1
0
1
2
3
40
35
30
25
20
15
10
5
0

Figure 68: The determined T-values per subcatchment / land use class for simulation 1 and 2.

Table 37: Overview of the parameter values found in the 2 calibration runs based on SM+TWSA

Parameter	ICF_{nf}	ICF_f	$CEVPF_{nf}$	$CEVPF_f$	β	LP	$PERC$	Q_{cf}	SUZ	$K0$	$K_{QuickFlow}$
Unit	$[mm]$	$[mm]$	$[-]$	$[-]$	$[-]$	$[-]$	$[mm/d]$	$[mm/d]$	$[mm]$	$[d^{-1}]$	$[d^{-1}]$
Sim 1	0.75	0.84	0.9	1.11	2.0	0.55	3.5	0.58	22.61	0.3	0.87
Sim 2	0.69	1.17	0.9	1.1	2.0	0.55	3.57	0.93	16.06	0.3	0.86
mean	0.72	1.0	0.9	1.11	2.0	0.55	3.54	0.75	19.34	0.3	0.87

L Results Scenarios Compared

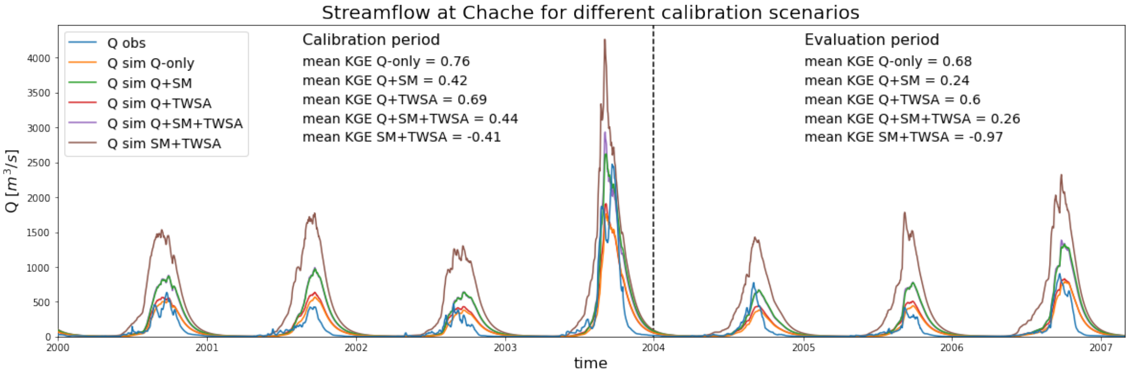


Figure 69: Hydrographs of the observed and simulated streamflow timeseries for all calibration scenarios at Chache (Black Volta). Chache is a calibration station. For each calibration scenario, the mean of the 2 simulations is shown. The mean KGE values given in the plots are mean values of the 5 objective functions used for calibration/evaluation as was explained in subsubsection 3.3.2.

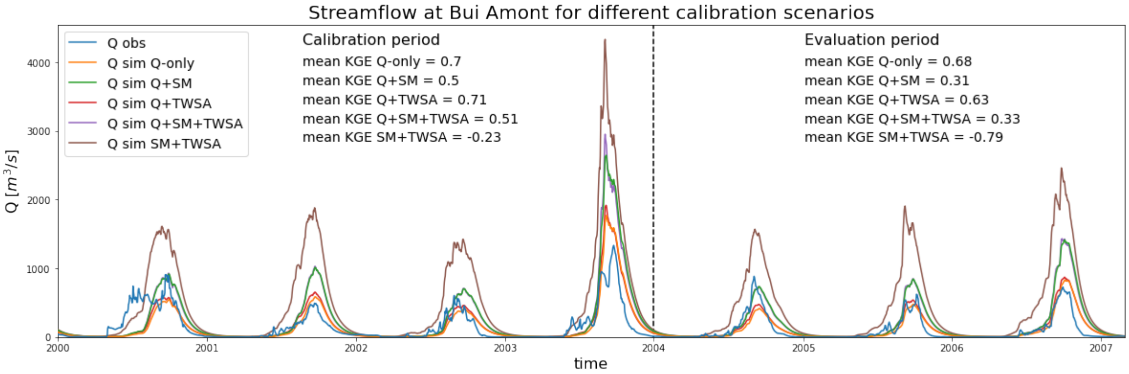


Figure 70: Hydrographs of the observed and simulated streamflow timeseries for all calibration scenarios at Bui Amont (Black Volta). Bui Amont is a calibration station. For each calibration scenario, the mean of the 2 simulations is shown. The mean KGE values given in the plots are mean values of the 5 objective functions used for calibration/evaluation as was explained in subsubsection 3.3.2.

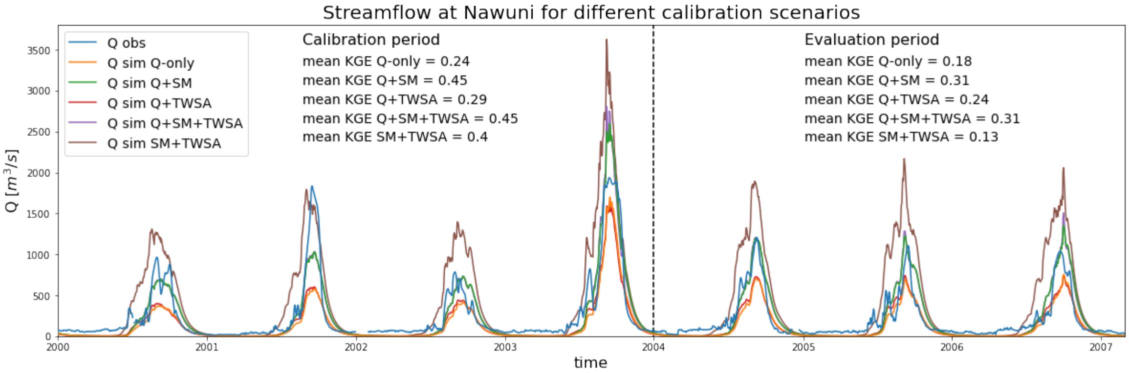


Figure 71: Hydrographs of the observed and simulated streamflow timeseries for all calibration scenarios at Nawuni (White Volta). Nawuni is an evaluation station. For each calibration scenario, the mean of the 2 simulations is shown. The mean KGE values given in the plots are mean values of the 5 objective functions used for calibration/evaluation as was explained in subsubsection 3.3.2.

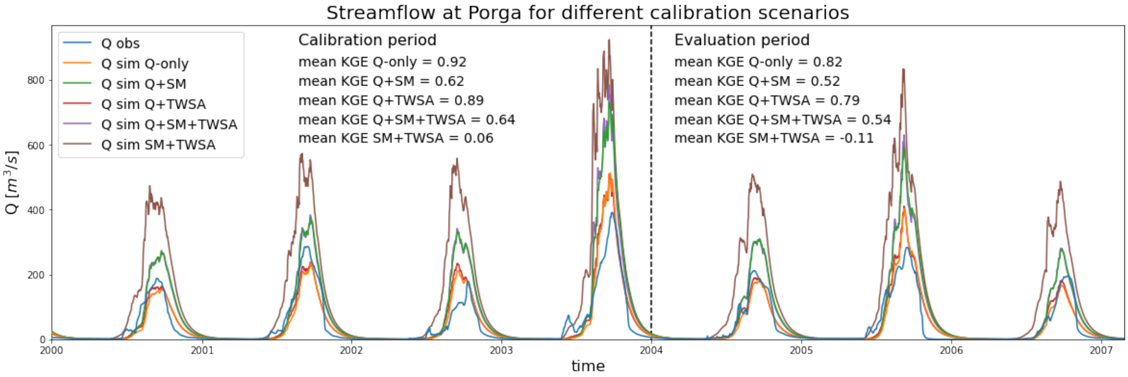


Figure 72: Hydrographs of the observed and simulated streamflow timeseries for all calibration scenarios at Porga (Oti). Porga is a calibration station. For each calibration scenario, the mean of the 2 simulations is shown. The mean KGE values given in the plots are mean values of the 5 objective functions used for calibration/evaluation as was explained in subsubsection 3.3.2.

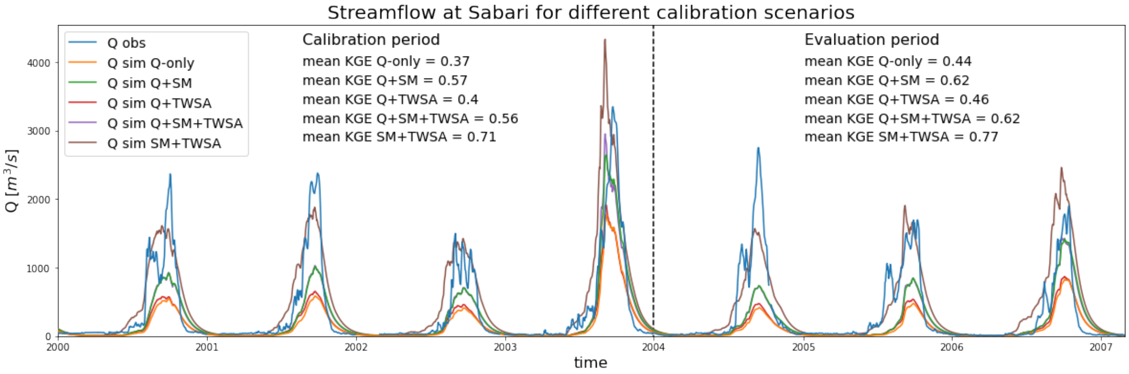


Figure 73: Hydrographs of the observed and simulated streamflow timeseries for all calibration scenarios at Sabari (Oti). Sabari is a calibration station. For each calibration scenario, the mean of the 2 simulations is shown. The mean KGE values given in the plots are mean values of the 5 objective functions used for calibration/evaluation as was explained in subsubsection 3.3.2.

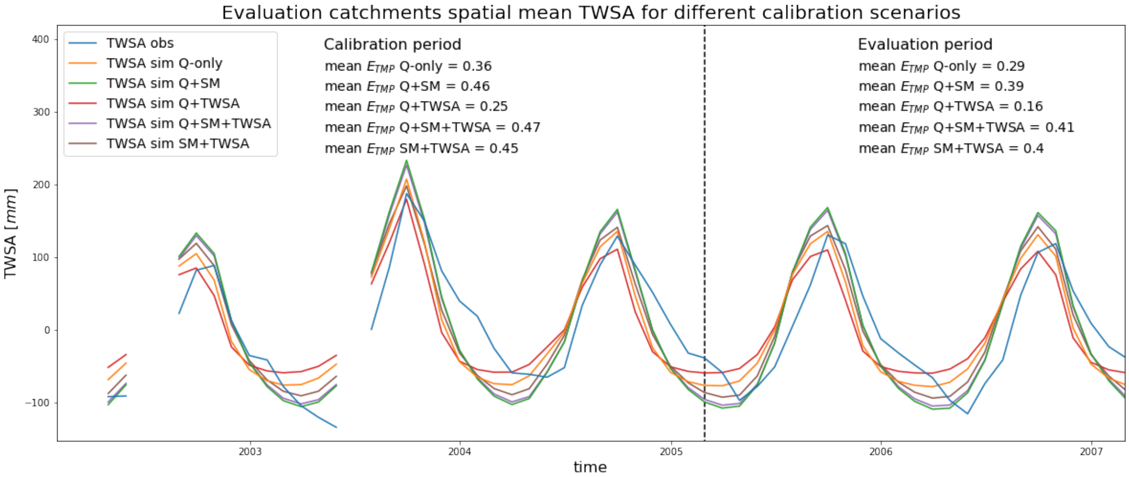


Figure 74: Timeseries of the mean TWSA observations and simulations within the evaluation catchments for all calibration scenarios. For each calibration scenario, the mean of the 2 simulations is shown.

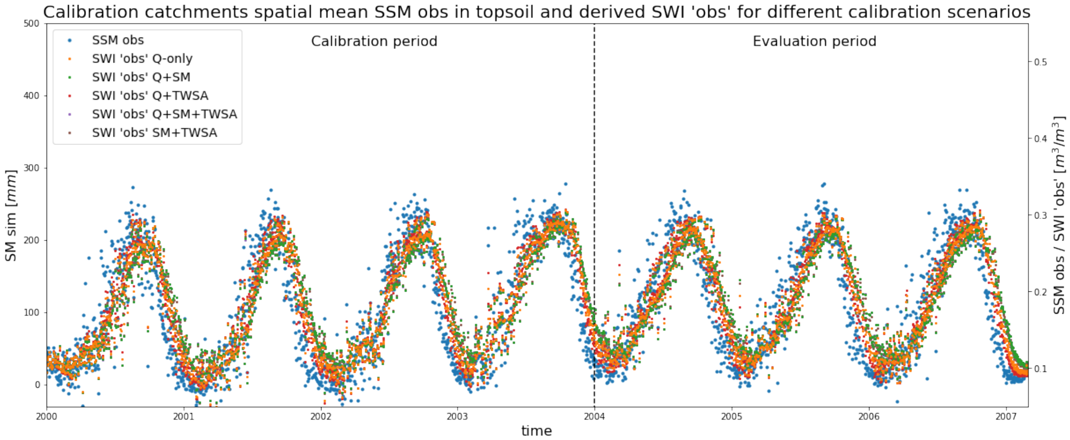


Figure 75: Timeseries of the mean surface soil moisture observations (SSM) and thereof derived soil water index 'observations' (SWI) within the calibration catchments for all calibration scenarios. For each calibration scenario, the mean of the 2 simulations is shown.

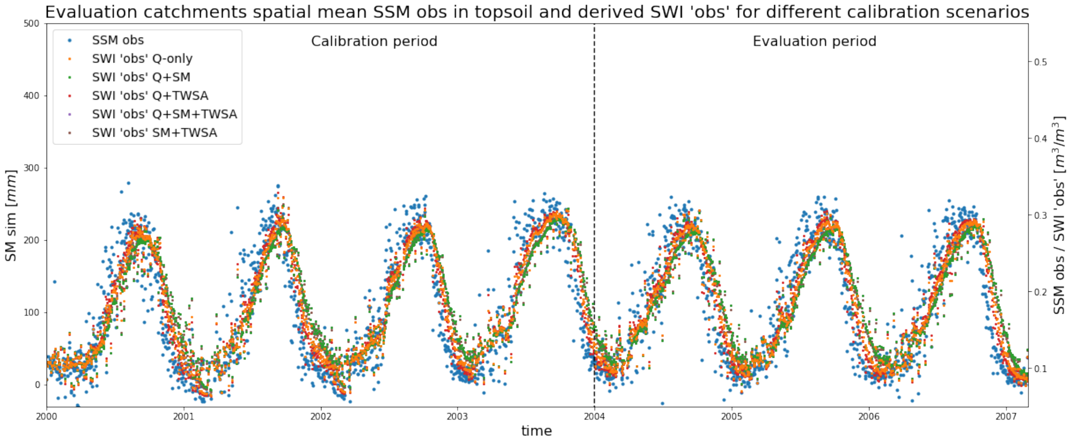


Figure 76: Timeseries of the mean surface soil moisture observations (SSM) and thereof derived soil water index 'observations' (SWI) within the evaluation catchments for all calibration scenarios. For each calibration scenario, the mean of the 2 simulations is shown.

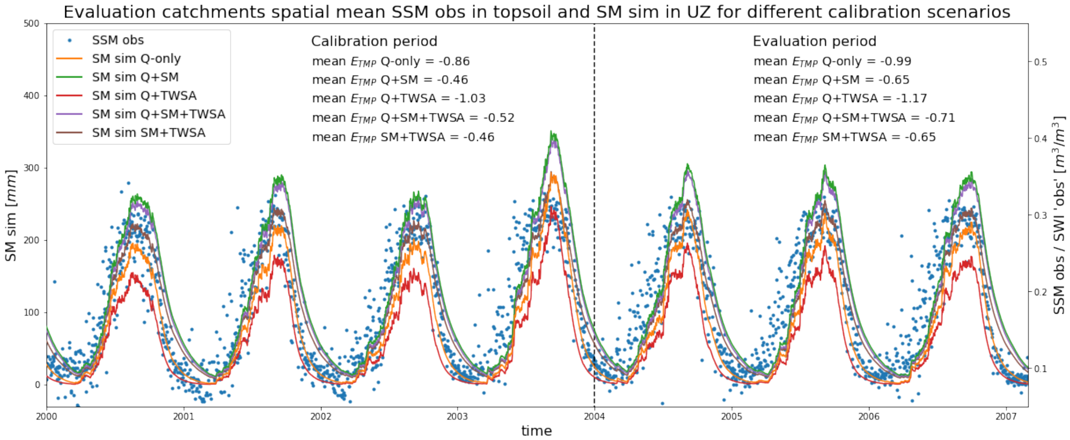


Figure 77: Timeseries of the mean surface soil moisture observations (SSM) and the simulated mean amount of water in the soil moisture reservoir (SM) within the evaluation catchments for all calibration scenarios. For each calibration scenario, the mean of the 2 simulations is shown.

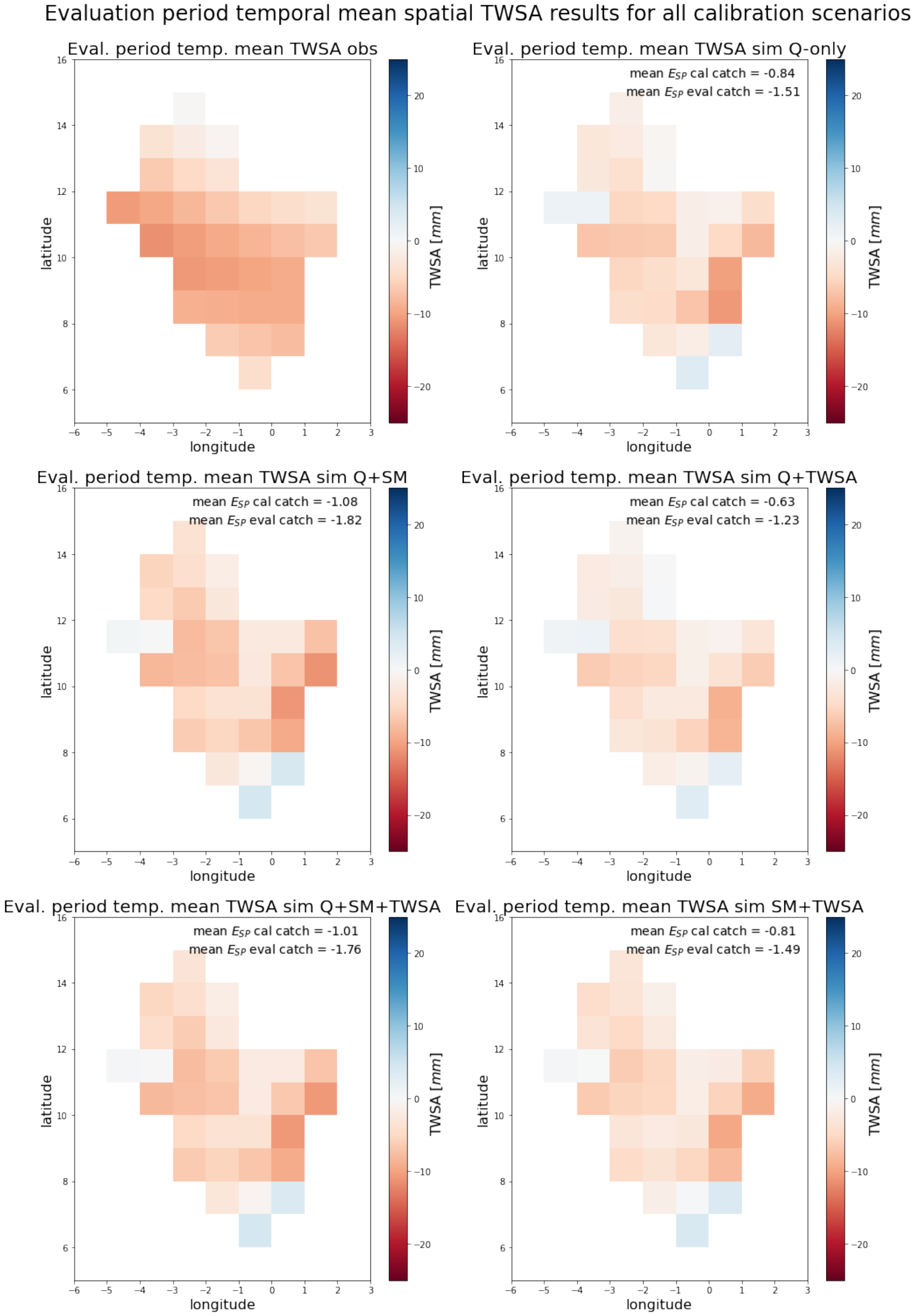

Evaluation period temporal mean spatial TWSA results for all calibration scenarios
Eval. period temp. mean TWSA obs
Eval. period temp. mean TWSA sim Q-only
mean E_{SP} cal catch = -0.84
mean E_{SP} eval catch = -1.51
Eval. period temp. mean TWSA sim Q+SM
mean E_{SP} cal catch = -1.08
mean E_{SP} eval catch = -1.82
Eval. period temp. mean TWSA sim Q+TWSA
mean E_{SP} cal catch = -0.63
mean E_{SP} eval catch = -1.23
Eval. period temp. mean TWSA sim Q+SM+TWSA
mean E_{SP} cal catch = -1.01
mean E_{SP} eval catch = -1.76
Eval. period temp. mean TWSA sim SM+TWSA
mean E_{SP} cal catch = -0.81
mean E_{SP} eval catch = -1.49
latitude
longitude
TWSA [mm]
16
14
12
10
8
6
−6
−5
−4
−3
−2
−1
0
1
2
3
20
10
0
−10
−20

Figure 78: Spatial plots of the mean TWSA observations and simulations in the evaluation period for all calibration scenarios. For each calibration scenario, the mean of the 2 simulations is shown.

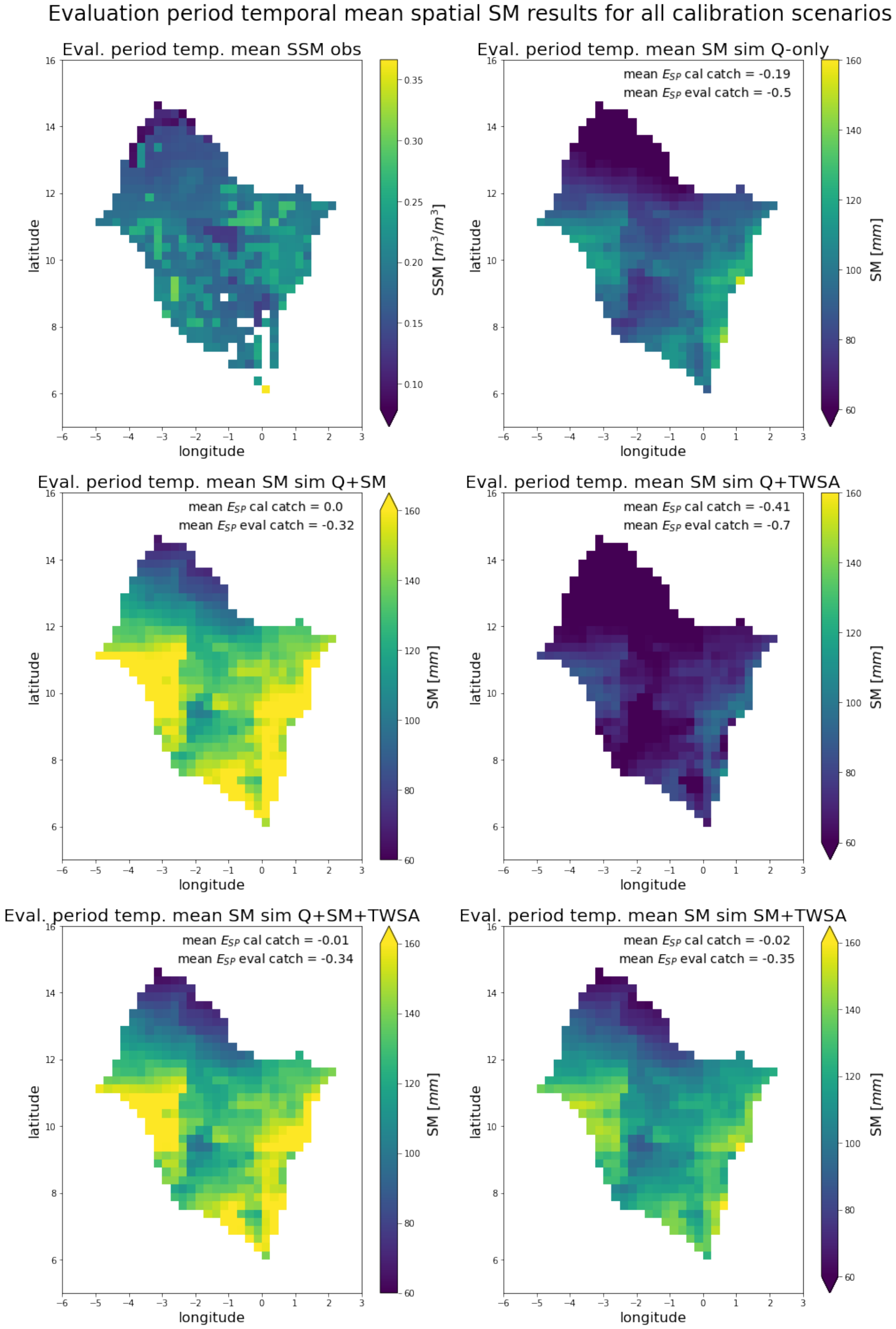

Evaluation period temporal mean spatial SM results for all calibration scenarios
Eval. period temp. mean SSM obs
SSM [m^3/m^3]
0.35
0.30
0.25
0.20
0.15
0.10
Eval. period temp. mean SM sim Q-only
mean E_{SP} cal catch = -0.19
mean E_{SP} eval catch = -0.5
Eval. period temp. mean SM sim Q+SM
mean E_{SP} cal catch = 0.0
mean E_{SP} eval catch = -0.32
Eval. period temp. mean SM sim Q+TWSA
mean E_{SP} cal catch = -0.41
mean E_{SP} eval catch = -0.7
Eval. period temp. mean SM sim Q+SM+TWSA
mean E_{SP} cal catch = -0.01
mean E_{SP} eval catch = -0.34
Eval. period temp. mean SM sim SM+TWSA
mean E_{SP} cal catch = -0.02
mean E_{SP} eval catch = -0.35
SM [mm]
160
140
120
100
80
60
latitude
16
14
12
10
8
6
longitude
−6
−5
−4
−3
−2
−1
0
1
2
3

Figure 79: Spatial plots of the mean surface soil moisture observations (SSM) and the simulated amount of water in the soil moisture reservoir (SM) in the evaluation period for all calibration scenarios. For each calibration scenario, the mean of the 2 simulations is shown.

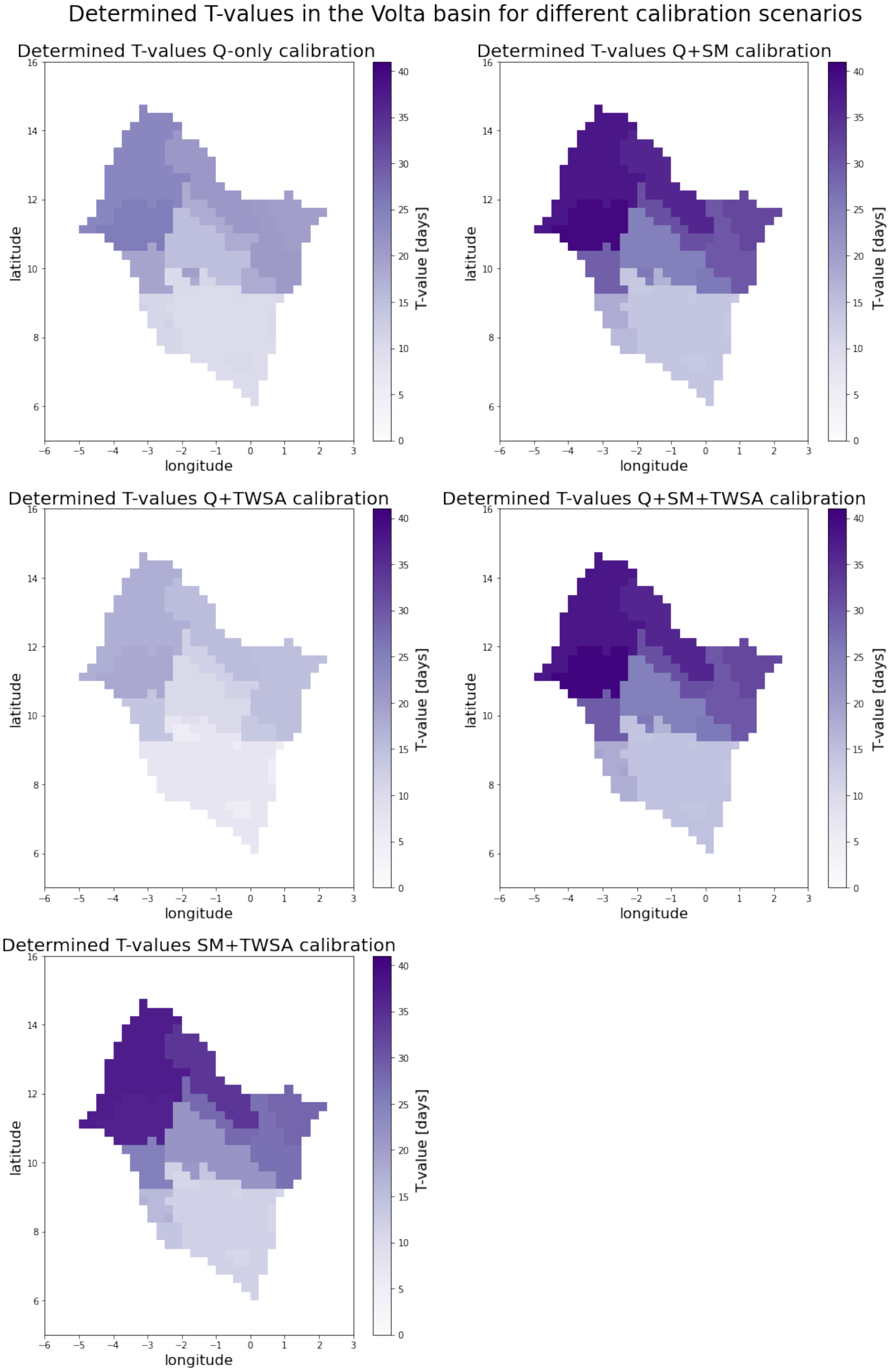

Determined T-values in the Volta basin for different calibration scenarios
Determined T-values Q-only calibration
Determined T-values Q+SM calibration
Determined T-values Q+TWSA calibration
Determined T-values Q+SM+TWSA calibration
Determined T-values SM+TWSA calibration
latitude
longitude
T-value [days]
16
14
12
10
8
6
−6
−5
−4
−3
−2
−1
0
1
2
3
0
5
10
15
20
25
30
35
40

Figure 80: The determined T-values per subcatchment / land use class for all calibration scenarios. For each calibration scenario, the mean of the 2 simulations is shown.

Table 38: Overview of the parameter values found in all calibration scenarios of this study

Parameter values	**Parameter**	ICF_{nf}	ICF_f	$CEVPF_{nf}$	$CEVPF_f$	β	LP	$PERC$	Q_{cf}	SUZ	$K0$	$K_{QuickFlow}$
Scenario	**Unit**	$[mm]$	$[mm]$	$[-]$	$[-]$	$[-]$	$[-]$	$[mm/d]$	$[mm/d]$	$[mm]$	$[d^{-1}]$	$[d^{-1}]$
Q-only	Sim 1	0.31	1.09	1.01	1.1	3.21	0.40	4.43	2.50	25.0	0.3	0.45
	Sim 2	0.52	1.1	1.08	1.1	2.83	0.37	5.6	1.83	24.36	0.3	0.85
	mean	0.41	1.09	1.04	1.1	3.02	0.39	5.02	2.16	24.68	0.3	0.65
Q+SM	Sim 1	0.3	0.83	0.9	1.1	3.5	0.55	4.33	0.6	21.01	0.16	0.45
	Sim 2	0.3	0.8	0.9	1.1	3.43	0.55	5.37	2.5	14.95	0.29	0.52
	mean	0.3	0.81	0.9	1.1	3.46	0.55	4.85	1.55	17.98	0.22	0.48
Q+TWSA	Sim 1	0.8	0.95	1.1	1.75	2.7	0.39	3.51	0.23	11.87	0.3	0.75
	Sim 2	0.8	1.0	1.1	1.75	2.11	0.26	3.5	0.48	22.07	0.28	0.51
	mean	0.8	0.98	1.1	1.75	2.41	0.32	3.51	0.36	16.97	0.29	0.63
Q+SM+TWSA	Sim 1	0.48	1.18	0.9	1.11	3.42	0.55	3.51	1.17	16.5	0.3	0.9
	Sim 2	0.74	1.19	0.9	1.1	3.29	0.55	3.5	2.1	17.8	0.3	0.45
	mean	0.61	1.18	0.9	1.1	3.36	0.55	3.5	1.64	17.15	0.3	0.68
SM+TWSA	Sim 1	0.75	0.84	0.9	1.11	2.	0.55	3.5	0.58	22.61	0.3	0.87
	Sim 2	0.69	1.17	0.9	1.1	2.	0.55	3.57	0.93	16.06	0.3	0.86
	mean	0.72	1.	0.9	1.11	2.	0.55	3.54	0.75	19.34	0.3	0.87