Master's Thesis in Hydrology

Value of Citizen Science Data and Limited Other Information for Hydrological Model Calibration

GEO 511 Master's Thesis

Franziska Schwarzenbach
16-728-867

Supervisor
Prof. Dr. Jan Seibert

Faculty member
Dr. Ilja van Meerveld

30 June 2022

Department of Geography

Physical Geography, H2K

Personal Motivation

In summer 2019, right before I started my Master's degree, I travelled along the Danube: I cycled to its source in Donaueschingen, and I marvelled at this beautiful little river. I saw the Danube again in Bratislava, and I regretted not coming by boat from Vienna. A few days later, I stood at the shore of the Danube in Budapest and tried to relate this – by my standards – giant river to the small Danube that I had seen in southern Germany. From that moment on, one question did not get out of my head anymore: **How much water is in this river?** I was lucky: In Budapest, the water level and discharge of the Danube are recorded, and the data are publicly available. Thus, I was able to satisfy my curiosity. For many rivers in the world, the question about the past, present and future discharge is not posed out of curiosity but because it is relevant for decision-making regarding flood protection and water resources management (Buytaert et al., 2014). However, for many rivers in the world, there is no data that can be used to answer these important questions.

Already during my Bachelor studies, I participated as a citizen scientist in the CrowdWater project. The easy way of contributing to science by collecting data that can help to provide a remedy to the lack of data fascinated me. This fascination grew even more when I started to work as an assistant in the CrowdWater project in late 2019. With a deeper insight to the project, I realised that citizen science has the potential to help answer the burning question about the discharge in a river in places where there is no data available otherwise. Thus, I decided to explore the value of the data collected in this citizen science project for hydrological model calibration in my Master's thesis. The overarching goal of this thesis was to test low-cost approaches using citizen science data and other data types for a reliable model calibration that can be applied in regions that lack an official hydrological measurement network.

Content

1 Summary ... 1
2 Introduction ... 2
2.1 Relevance of and challenges in hydrological modelling ... 2
2.2 Citizen science as a data source for hydrological modelling ... 3
2.3 Research questions and hypotheses ... 4
3 Background ... 6
3.1 Literature review ... 6
3.1.1 Hydrological modelling with limited calibration data ... 6
3.1.1.1 Model calibration with a limited number of discharge measurements ... 6
3.1.1.2 Model calibration with water level data ... 7
3.1.2 Citizen science data for hydrological modelling ... 8
3.1.2.1 Citizen science in hydrology ... 8
3.1.2.2 Model calibration with data from the CrowdWater project ... 8
3.1.2.3 Other citizen science approaches for hydrological model calibration ... 9
3.2 The CrowdWater project ... 11
3.2.1 The virtual staff gauge ... 12
3.2.2 The CrowdWater game ... 13
4 Methods ... 15
4.1 Study catchments ... 15
4.2 Meteorological data ... 21
4.2.1 Precipitation ... 21
4.2.2 Temperature ... 21
4.2.3 Evaporation ... 22
4.3 Discharge data ... 23
4.4 Citizen science data ... 23
4.4.1 App data and pen and paper data ... 25
4.4.2 Quality-controlled water level classes from the CrowdWater game ... 28
4.5 The HBV model ... 30
4.5.1 Model description ... 30
4.5.1.1 Snow routine ... 31
4.5.1.2 Soil moisture routine ... 31
4.5.1.3 Response function ... 32
4.5.1.4 Routing routine ... 32

4.6 Model settings ... 32
4.6.1 Catchment settings ... 32
4.6.2 Calibration, validation, and warm-up period ... 33
4.6.3 Parameter ranges ... 33
4.6.4 Model calibration methods ... 34
4.6.4.1 Objective functions ... 34
4.6.4.2 Monte Carlo simulation ... 35
4.6.4.3 GAP simulation ... 36
4.6.5 Evaluation ... 36
4.6.5.1 Model validation ... 36
4.6.5.2 Upper and lower benchmark ... 36
4.7 Definition of scenarios ... 37
4.7.1 Model calibration per scenario ... 38
4.8 Implementation of additional knowledge ... 39
4.8.1 Mean discharge ... 39
4.8.2 Water levels instead of water level classes ... 39
4.8.3 Water level classes checked by citizen scientists ... 40
4.9 Data analysis ... 40
5 Results ... 41
5.1 Benchmarks ... 41
5.2 Basic approach ... 43
5.3 Additional mean discharge ... 56
5.4 Water levels instead of water level classes ... 66
5.5 Water level class data checked by citizen scientists ... 71
5.5.1 Adjusted benchmarks ... 71
5.5.2 Resulting model performances ... 72
6 Discussion ... 77
6.1 Value of only water level class observations ... 77
6.2 Value of additional discharge measurements ... 78
6.3 Value of estimating the mean discharge ... 79
6.4 Value of water levels instead of water level classes ... 80
6.5 Value of water level class data checked by citizen scientists ... 81
6.6 Limitations of this study ... 81
6.6.1 Study catchments and model ... 81
6.6.2 Data used for calibration ... 82

6.7 Outlook ... 83
7 Conclusions ... 84
8 Acknowledgements ... 85
9 Bibliography ... 86
10 Appendix ... 95
10.1 Flashiness and baseflow indices ... 95
10.2 Temperature measurement stations ... 95
10.3 Details on discharge data ... 96
10.4 Links to CrowdWater spots ... 96
10.5 Form used at the pen and paper stations ... 97
10.6 Elevation zones ... 99
10.7 Ranking of parameter sets ... 101
10.8 Shared parameter sets ... 102
10.9 Distribution of parameter values ... 103
10.10 Density plots NPE vs. volume error ... 106
10.11 Results from other filters ... 107
11 Declaration of Independence ... 113

1 Summary

In this thesis, the semi-distributed bucket-type hydrological model HBV was calibrated based on water level class data obtained by citizen scientists of the CrowdWater project and a limited number of discharge measurements distributed over the hydrological year. The value of these data types was investigated for eleven catchments in Switzerland and Austria. The results showed that accurate water level class observations are informative for hydrological model calibration, especially if they are combined with at least one discharge measurement per year. Furthermore, it was tested if the use of a rough estimate of the mean discharge in a catchment in addition to the citizen science data and the discharge measurements improves the model performance. The additional information about the mean discharge volume led to an increased model performance, especially compared to the situation in which only citizen science data was used to calibrate the model and thus any information about the discharge volume was missing. However, also if a few discharge measurements were available, some additional constraint by using an estimate of the mean discharge improved the model performances. An estimate of the mean discharge thus increases the value of both citizen science data and a limited number of discharge measurements and is of greater value than additional discharge measurements. Thereby, it is better to use a broader interval for the estimate of the mean discharge in a catchment than to use a very precise estimate to not fine-tune the model to the mean discharge. It was tested if water level measurements instead of water level class observations lead to increased model performances to represent the situation in which citizens could precisely measure the water level instead of making an estimation of a water level class. In general, this led to higher model performances. However, in catchments already having citizen science data of a high quality, the impact of precise measurements as a replacement for the water level classes observed was very limited. Thus, the easier approach of collecting water level class data instead of precise water levels is sufficient to calibrate a model reliably. Water level class data collected in the CrowdWater app can be improved by many citizen scientists playing the CrowdWater game. When the water level class data originating from the CrowdWater app were replaced with water level data of a higher resolution resulting from the CrowdWater game, the model performance could be improved. This was not the case if information got lost in the CrowdWater game, i.e., if the data-quality was deteriorated in the CrowdWater game, thus some double-checking of the data-quality may be required even after the control process in the CrowdWater game. Based on these findings, a data collection approach including water level class observations of a high quality as well as a few discharge measurements per year and an estimate of the mean discharge can be suggested for catchments in which hydrological data is missing otherwise. This thesis showed that water level class data collected by citizen scientists combined with other limited information about the discharge in a stream have a value for the calibration of a hydrological model in regions where no other data are available for this important task.

2 Introduction

2.1 Relevance of and challenges in hydrological modelling

The hydrological sciences face data scarcity (Beven, 2012). Measurements of important variables such as discharge and water level are expensive, and even though technology has improved in the last decades, costs for maintenance, storage and quality control remain high or are too high compared to the funding pressure that many institutions face. Thus, poorer countries often cannot afford to build an extensive measurement network and richer countries rather tend to discontinue measurement time series than to expand the measurement network (Hannah et al., 2011; Mishra & Coulibaly, 2009; Ruhi et al., 2018; Sivapalan, 2003). Remote areas that are more difficult to access are often very poorly gauged (Getirana et al., 2009). Especially in countries where hydrological data is completely missing, the hydrological conditions are often unfavourable, thus, data that could be used for forecasting could make a huge difference for the development of adaption or prevention strategies (Hrachowitz et al., 2013; Walker et al., 2016; Weeser et al., 2021).

Hydrological models are used for different purposes: Extreme events such as droughts and floods that need to be expected in a certain catchment can be modelled such that protection measures can be taken as a preparation (Brunner et al., 2021; Davids et al., 2017; Engeland et al., 2004). Also on a shorter timescale, the forecasting of floods is a crucial element in hydrological modelling. Based on the weather forecast, it is possible to calculate the amount of water to be expected in a stream using a hydrological model. With these calculations, it is possible to decide if for example people need to be evacuated or if other short-term measures need to be taken to avoid fatalities and damage (Addor et al., 2011). Similarly, by forecasting droughts or low flows, hydrological models help to decide on the management strategies of water resources (Fung et al., 2020). This is crucial, especially in locations where water scarcity is a major problem. The number of locations where this is the case is growing due to increasing population and climate change (Kundzewicz, 1997).

Furthermore, hydrological models can be used to model the impacts of climate change and changes in land use and thus allow to act by implementing adaption and prevention measures (Dwarakish & Ganasri, 2015). They are used in planning of hydropower plants to calculate the expected energy supply as well as the infrastructure required for the hydropower plant to be effective (Fasipe et al., 2021; de Oliveira Serrão et al., 2021). To ensure the cooling of nuclear power plants that usually relies on a sufficient water supply, hydrological models are used too (Kirkwood, 1982). Furthermore, decision making in the industry and the agriculture can be supported by hydrological models (Haberlandt, 2010). Hydrological modelling is also used when it comes to the distribution of water resources among several stakeholders (Savic et al., 2009).

On top of all these purposes, hydrological models are also an important tool in science. They allow to express a hydrological system, or the understanding of that hydrological system by the hydrologist, in the form of mathematical expressions. Thus, hydrological models help to draw proofs of concept and provide the hydrologist with a formal tool to describe her or his hydrological system of interest (Solomatine & Wagener, 2011). Hydrological models help to understand hydrological systems better and allow to improve our knowledge about hydrological processes. This improved knowledge is important to be able to support the decision-making process in many different areas, as described above.

To make use of a hydrological model, some form of calibration of the hydrological model is required. Calibration means that the parameters used in the formulae that make up the hydrological model are adjusted such that the model can simulate the hydrological behaviour of the modelled catchment, i.e., such that the simulated discharge fits the observed discharge (Bergström, 1991; Perrin et al., 2007).

In the optimal case, a long time series of several years of discharge data is available with which the hydrological model can be calibrated. If that is not possible, the availability of at least some discharge data can also be helpful to calibrate the model (see for example Brath et al., 2004; Perrin et al., 2007; Pool et al., 2017). Using the calibrated model and meteorological or climatological data, the model can then be used to model the discharge in time periods for which no hydrological data is available (Bergström, 1992).

However, as mentioned earlier, the availability of hydrological data is limited. Therefore, hydrologists are always confronted with a lack of data: Usually, there are no long time series of discharge measurements or other hydrological variables available, and thus it remains difficult to calibrate a hydrological model (Perrin et al., 2007; Pool et al., 2017). In many cases, there is even no data at all available and it is not possible to calibrate a model in order to predict discharge. In the hydrological decade starting in 2003, the Predictions in Ungauged Basins (PUB) initiative tried to explore alternative prediction methods in order to ease this obstacle (Sivapalan, 2003). Even though many successes could be reached in this initiative, the prediction of water resources in ungauged catchments is still a major challenge in hydrology (Hrachowitz et al., 2013). One approach that is often used is the regionalization of parameters, i.e., the expansion in space rather than in time (Seibert, 1999). In regionalization approaches, parameters calibrated for one catchment are transferred to other catchments based on spatial proximity and similar catchment characteristics (Merz & Blöschl, 2005).

Aside regionalization, modelling attempts have been made in catchments in which some data is available but has gaps, consists of only a short time series of measurements, or is attached with a large uncertainty. These data can still be valuable for hydrological modelling, especially if wet hydrological conditions are covered by the available data points (as for example found by Kim & Kaluarachchi, 2009; Melsen et al., 2014; Perrin et al., 2007; Seibert & Beven, 2009; Sun et al., 2017). Especially conceptual models such as the HBV model (see section 4.5) can deal with missing input data as their parameters do not represent one quantifiable physical property of a catchment. In return, these models demand more calibration efforts based on the data that is available (Bergström, 1991).

2.2 Citizen science as a data source for hydrological modelling

As some data is required to calibrate a model, less traditional ways of getting hydrological information and especially information about the amount of water in a stream are becoming more popular with a good reason. Aside for example the extraction of hydrological data from satellite imagery (Elmi et al., 2015), citizen science is another of these less traditional ways to gather hydrological data: Even if there are no measurement devices available in some region, oftentimes there are people who can collect hydrological data. New technologies and especially the internet offer easy tools to make the data available quickly and bring many different sources such as the smartphones of different people together (Davids et al., 2017; Lowry & Fienen, 2013; Silvertown, 2009; Starkey et al., 2017). Citizen science approaches (see section 3.1.2) can therefore be used to generate data where official measurement networks are deficient or not existent (Buytaert et al., 2014). Data collected in collaboration with the public are usually of low cost (Assumpção et al., 2018) and have shown to be useful for hydrological modelling (for examples, see Etter et al., 2020b; Mazzoleni et al., 2017; Starkey et al., 2017; Weeser et al., 2019). Therefore, citizen science is a promising method to help reducing the lack of data in the hydrological sciences.

This thesis sought to further investigate on the value of hydrological citizen science data, more specifically water level class data collected by citizen scientists in the CrowdWater project (see section 3.2). In contrary to discharge estimates by untrained citizen scientists (Etter et al., 2018), water level class observations provide a valuable data base for otherwise ungauged catchments (Etter et al.,

2020b) as they are way more accurate than discharge estimates (Strobl et al., 2020a). In addition to the use of these water level class data, other data that could be obtained with a limited effort by trained personnel (that could potentially consist of citizen scientists, as it was shown for discharge measurements by Davids et al. (2019)) was used to calibrate a hydrological model. While there is usually no reference data available when citizen science is used as a data source to answer a research question (Aceves-Bueno et al., 2017), official discharge measurement time series as a reference were used here to be able to judge the model performances obtained with the aforementioned model calibrations. The resulting insights in the reliability of the model calibrations allowed to suggest low-cost approaches for data collection by citizen scientists in regions where official hydrological data is missing.

2.3 Research questions and hypotheses

Following the overarching goal of providing low-cost approaches for reliable model calibrations, this thesis had the goal to calibrate a semi-distributed bucket-type model with limited data from different sources. To do so, a limited number of discharge measurements and water level class data obtained by citizen scientists in the CrowdWater project were used. Furthermore, it was investigated if using an estimate of the mean discharge, water levels instead of water level classes and citizen science data that has been checked by many citizen scientists in the CrowdWater game (see section 3.2.2) for the model calibration improves the model performance. More specifically, the thesis sought to answer the following main research question and three sub-questions:

Does the calibration of a hydrological model based on citizen science data and a limited number of discharge measurements lead to a reliable simulation of discharge?

1. *Does an additional estimate of the mean discharge improve the model performance?*
2. *Do water level data instead of water level class data improve the model performance?*
3. *Does citizen-based quality control of the citizen science data improve the model performance?*

To answer the main research question, eleven study catchments were selected (see section 4.1). These catchments were assumed to be (almost) ungauged. In reality, there was discharge data with an hourly resolution available for these catchments. These discharge data time series were used to extract the limited number of discharge measurements, and to determine the performance of the model calibration by comparison of the simulated and observed discharge. The citizen science data were the water level class data collected by citizen scientists of the CrowdWater project (see section 3.2). The number of observations and the quality of these differed for each catchment (see section 4.4).

For each study catchment, a combination of a limited number (0, 1, 3, 6, 12) of regular discharge measurements per hydrological year and a certain percentage (0%, 25%, 50%, 75%, 100%) of the available citizen science data were combined. This resulted in 24 different data availability sets (or scenarios) that were used to calibrate the model. Each of the 24 different data availability scenarios resulted in calibrated model parameter sets and corresponding model performances that were used to judge the value of the data used for calibration.

One can in general expect more reliable model performances when using more water level class data (Etter et al., 2020b) and when using more discharge measurements (Pool et al., 2017; Seibert & Beven, 2009) for the calibration of the model. Thus, the expectation was that there would be an improvement in model performance if more citizen science data and a larger number of discharge measurements are used for model calibration. The impact of the citizen science data on the model performance was expected to depend on the number of available observations, as well as on the accuracy

of the observations in terms of the correlation between the water level class data and the actual measurements.

For the first sub-question, it was assumed that an estimate of the mean discharge is available additionally. The water level class data from CrowdWater contained no information about the discharge volume. Thus, additional knowledge of the mean discharge was assumed to improve the model performance when mainly citizen science data was used for model calibrations. In similar studies, the model calibration with water level data could be improved by adding some volume information as an additional constraint (Seibert & Vis, 2016, Weeser et al., 2019).

To answer sub-question 2, the citizen science data was assumed to be perfectly correlated with the discharge in the stream. Instead of water level classes, exact water level measurements were thus used for the model calibration. This represented the situation when citizen scientists read the exact water level from a staff gauge installed in the stream and record this value instead of a water level class and without any errors. The impact of perfectly correlated citizen science observations on model performance was expected to be larger for catchments for which there was a low correlation between the water level class data submitted by citizen scientists and the measured discharge than for catchments for which the correlation of the water level class data and the measured discharge was already high. However, because the resolution of the perfectly correlated citizen science data was higher than the resolution of the water level classes, an improve in the model performance was also expected for the catchments that already had a high correlation.

It is not realistic that perfectly correlated water level observations are obtained with a citizen science approach. However, some of the errors contained in water level class observations can be filtered out by careful data quality control. Therefore, for sub-question 3, instead of using the water level class data submitted by a single citizen scientist, quality-controlled citizen science data was used for the calibration. The quality-control was done by (other) citizen scientists in the CrowdWater game (see section 3.2.2). Thanks to the "wisdom of the crowd" (Surowiecki, 2004), this should lead to an improvement in the accuracy and resolution of the data after enough players estimated the water level classes on the pictures uploaded in the CrowdWater app (Strobl et al., 2019).

As data from the CrowdWater game was not available for all study catchments, the use of quality-controlled data from the CrowdWater game was limited to four catchments. In these four catchments, at least one quarter of the water level class observations available was already checked by enough citizen scientists in the CrowdWater game. Analogous to the hypothesis for sub-question 2, an improvement of the model performance for all catchments was expected because the data points from the CrowdWater game were assumed to be higher resolved and more accurate than the data points from the CrowdWater app. However, the improvement of the model performance was expected to be especially pronounced for catchments for which there was a poor correlation between the original citizen science data and the discharge measurements, as the potential room for improvement was assumed to be large in these cases.

3 Background

3.1 Literature review

3.1.1 Hydrological modelling with limited calibration data

3.1.1.1 Model calibration with a limited number of discharge measurements

Regionalization is a common approach to model the ungauged catchment. However, a traditional regionalization approach without any hydrological information about the target catchment comes with uncertainties. A limited number of discharge measurements can help to overcome this issue by constraining the parametrization found by regionalization with data that actually describes the behaviour of the target catchment (Drogue & Plasse, 2014; Pool et al., 2019; Viviroli & Seibert, 2015). A study covering more than 600 catchments in France showed that as little as ten discharge measurements over a period of ten years close more than half of the performance gap between a parametrization obtained by a traditional regionalization approach and the calibration of a model against a complete discharge time series measured in the catchment of interest (Rojas-Serna et al., 2016).

However, the alternative to a regionalization approach is to directly calibrate a hydrological model based on a limited number of discharge measurements or a comparably short calibration period. This is reasonable since the information content of a complete discharge time series may be partly redundant as a catchment reacts similarly to similar conditions (Juston et al., 2009). The approach seems to have potential to perform better than traditional regionalization: Rojas-Serna et al. (2006) found that a model calibration based on thirty measurements at random days during a period of three to five years outperforms the traditional regionalization.

Several other researchers have investigated on the value of a limited number of discharge measurements: Eng & Milly (2007) were able to characterise base flow recession by considering two to twenty discharge measurements per year under low flow conditions. Thereby, a higher number of measurements led to a shrinking error (Eng & Milly, 2007). Perrin et al. (2007) found that the use of 350 random discharge measurements spanning different hydrological conditions during a period of 39 years for model calibration was enough to reach a plateau of parameter variability and model performance, thus additional information did not provide any added value. They concluded that especially for simple models, little data from different conditions is already enough to get stable and robust parameter sets (Perrin et al., 2007). The importance of sampling during different hydrological conditions, especially during high flow conditions, to be able to simulate the hydrograph accurately was also highlighted by Seibert & Beven (2009) as well as by Pool et al. (2017).

Seibert & Beven (2009) found that a plateau in model performance was reached when 32 observations per year over a 10-year period were used. They showed that less than four observations per year can act disinformatively if the selection of sampling days is poor. In those cases, parametrizing the model randomly led to better results than constraining the parametrization with these observations. They furthermore found that even though an intelligent sampling strategy is more challenging than just random or regular sampling, model performance can be increased by applying such a strategy (Seibert & Beven, 2009). In a recent study by Pool & Seibert (2021), two to six discharge measurements per hydrological year were shown to be informative in most cases, especially if taken during high flow conditions. However, as already mentioned by Seibert & Beven (2009), the authors also state that too few discharge observations may also act disinformatively in some cases (Pool & Seibert, 2021).

Different sampling strategies of varying complexity have been investigated by Pool et al. (2017). They investigated the value of different sets of twelve discharge measurements per year that resulted

from the different sampling strategies. As mentioned earlier, they found that sampling during high flow conditions improved the simulation of the hydrograph. On contrary, the flow duration curve could best be simulated when the sampling took place during medium and low flow conditions. Little performance differences were obtained among strategies that contain discharge measurements from a variety of hydrological conditions (Pool et al., 2017).

Tada & Beven (2012) investigated the value of coherent measurement days during a 10-year period, whereby these data started on a random day and had a length of four to 512 days. They found that a longer observation period resulted in better model performance and that at least 32 to 64 days of observations are required to avoid timing errors in the simulated hydrograph (Tada & Beven, 2012). The finding of Seibert & Beven (2009) that the ensemble mean built from the top-performing 1% of all parameter sets considered outperformed the best parameter set was confirmed for all cases with observation periods shorter than 16 days by Tada & Beven (2012).

Also Melsen et al. (2014) were able to show that five months of data are sufficient to reliably calibrate a model. Thereby, periods with a lot of precipitation were valuable to simulate high flows while modelling the recession was more successful when a period with low evapotranspiration was used (Melsen et al., 2014). This is in line with the findings of Pool et al. (2017) that were described before. However, the authors highlight that longer time series are still valuable since they decrease the effect of short-term disinformation that may be contained in a limited observation period. Based on that, they note that in regions were data quality may be more of an issue than in Switzerland where they conducted their study, it could be valuable to consider longer observation periods for calibration to avoid large uncertainties (Melsen et al., 2014).

Further work using short calibration periods was done by Brath et al. (2004) using a distributed model. They found that a stable parametrization could be reached with an observation period of three months (Brath et al., 2004). Short periods of data availability were also shown to be valuable for physically based models: In a study by Sun et al. (2017), already one month of observations from wet catchments led to similar performances as three years of observations. For drier and headwater catchments, a total of six months was required to reach these performances (Sun et al., 2017).

Seibert & McDonnell (2015) compared the value of a continuous discharge time series of three months and the value of ten flexibly chosen observations during high flow events and found that the latter was almost as large as the first. The value of the flexibly chosen observations decreased if fewer observations were made or if they were not done during high flow events but for example each week. Furthermore, they found that a continuous sampling during one event was more valuable than making single observations within a fixed time interval (Seibert & McDonnell, 2015). The advantages of event-based sampling compared to a sampling strategy at fixed points in time or randomly was also shown by Juston et al. (2009), Singh & Bárdossy (2012) as well as Correa et al. (2016). However, a different study showed that also randomly sampled data can be valuable as long as at least one third of the data is obtained during high flow conditions (Kim & Kaluarachchi, 2009).

3.1.1.2 Model calibration with water level data

Discharge measurements are more expensive and challenging than measuring the water level in a stream. Therefore, in continuous discharge time series, it is often the water level that is measured and then converted to discharge via a rating curve. As rating curves are interpolations between several point measurements that come with uncertainties themselves, they are subject to uncertainty (Le Coz et al., 2014; McMillan et al., 2010). Seibert & Vis (2016) thus tested a model calibration using water level data only. Water level data contains all information on the discharge dynamics but lacks any

information about the discharge volume. Especially in wet catchments, the approach led to a good simulation of discharge time series. Additional knowledge of the annual discharge volume led to an improvement in model performance, whereby especially catchments that did not perform well when only water level data was available profited from this volume information (Seibert & Vis, 2016).

Similar findings were made by Jian et al. (2017): The calibration using water level data only was good regarding the correlation but could be improved in terms of volumes when a few discharge measurements were available additionally (Jian et al., 2017). In their study about the value of water level time series and a limited number of discharge observations, Pool & Seibert (2021) found that the combination of both data types led to more reliable model calibrations than if water levels or a few discharge measurements only were used to calibrate the model (Pool & Seibert, 2021).

3.1.2 Citizen science data for hydrological modelling

3.1.2.1 Citizen science in hydrology

The Oxford English Dictionary defines citizen science as "scientific work undertaken by members of the general public, often in collaboration with or under the direction of professional scientists and scientific institutions" (OED, 2021). In a shorter way, citizen science can be defined as "the participation of the general public (i.e., non-scientists) in the generation of new scientific knowledge" (Buytaert et al., 2014: 1). The term "citizen science" may cause problems, for example since citizenship is usually (as for example in the case of the CrowdWater project) not a condition for contributing to citizen science, and since the actual involvement of the public may vary strongly between different citizen science projects. Still, it was used here to describe the collaboration of scientists and volunteers in a common project. For a detailed discourse on the terminology in citizen science, the reader is referred to the review and synthesis paper by Eitzel et al. (2017).

In hydrology, citizen science has already been used for several purposes. In their review paper, Buytaert et al. (2014) give examples in which the data scarcity in hydrology was tackled with the help of citizen science. Thereby, citizen scientists obtained data about precipitation, discharge, water quality, soil moisture, vegetation dynamics and water use (Buytaert et al., 2014). Even though the data obtained like this may differ in type and quality from traditional hydrological data, new opportunities open with the help of citizen scientists. Thus, citizen science is a promising method not only in hydrological modelling, but in environmental science in general. This is especially true since many projects or research questions require large datasets with a high spatial resolution over large regions (Silvertown, 2009). In the literature review presented here, the focus was on hydrological data obtained by citizen scientists that can be used for hydrological modelling. Especially the recognition of the value of water level data for model calibration (see section 3.1.1.2) is of interest for citizen science approaches in hydrological modelling since water level data can more easily be obtained by untrained citizen scientists than discharge data (Strobl et al., 2020a).

3.1.2.2 Model calibration with data from the CrowdWater project

Within the CrowdWater project, the value of hydrological data obtained by citizen scientists for hydrological model calibration has been investigated in different studies. One study showed that discharge estimates by untrained people are not a feasible data source: Etter et al. (2018) built a synthetic data set of discharge estimates based on the errors in such estimates obtained in surveys with untrained people along streams. They used these data to calibrate the HBV model and found that the discharge estimates are not accurate enough to be used in hydrological model calibration if the error contained in the data cannot be reduced. However, with a reduced error thanks to training or filtering of the data (which could be simulated with corresponding changes in the synthetic data set), discharge estimates were found to be valuable for hydrological model calibration. Thereby, more observations and a more

even distribution of the observations over the year led to better model performances. Especially in catchments that showed a seasonal variable discharge behaviour, the model performance could be increased when data from different discharge conditions (i.e., different seasons) were available compared to the situation when the observations were more clustered (Etter et al., 2018).

Already earlier, van Meerveld et al. (2017) showed that not only water level measurements but also water level class observations are of value for the calibration of a hydrological model, even if only two water level classes with a boundary at high flow conditions are used. Especially for drier catchments, the use of up to five water level classes led to an improvement of the model performance. However, for these drier catchments, the calibration against discharge data instead of water level classes was found to be clearly advantageous, whereby the difference when comparing the value of those two data types was rather small for wet catchments (van Meerveld et al., 2017). In this study, daily information about the water level in a stream was used, thus not directly data that can be expected from citizen scientists

In a subsequent study by Etter et al. (2020b), the authors used a synthetic data set of water level class data in irregular intervals and investigated the value of these data for the calibration of the HBV model. The data set was built based on water level class estimates by people comparing the water level in a stream to a picture including the virtual staff gauge used in the CrowdWater project (see section 3.2.1). For the majority of the water level class estimates, the deviation of the correct water level class was one water level class at most (Strobl et al., 2020a) and the authors found that these errors had a clearly smaller effect on the model calibration than the errors in the discharge estimates described above. The main finding of the study was that the availability of one water level class observation per week over one year is enough to improve the model performance compared to the situation without any data. Furthermore, if at least four water level classes were used, the number of water level classes did not have an influence on the model performance. Another important finding was that the replacement of the water level class observations with water level measurements of the same temporal resolution resulted in similar model performances, i.e., the lower resolution of the water level classes was found to be less important than the temporal resolution with which the observations were made (Etter et al., 2020b).

3.1.2.3 Other citizen science approaches for hydrological model calibration

The value of high-resolution water level read from a physical staff gauge for hydrological model calibration has been investigated by Weeser et al. (2018, 2019). In a citizen science project in western Kenya, they asked untrained people to record the water level of a stream in a headwater catchment by sending a text message to the research team (Weeser et al., 2018). They found a very high accuracy of the observations by citizen scientists when compared to the water level recorded by a radar. Still, the calibration of the model on the water level observations led to a drop in model performance compared to the calibration of the model on discharge data. Part of this drop could be explained by the missing volume information in the water level data. To reduce this drop in model performance, a water balance filter was applied: Each parameter set had to match the simplified water balance calculated from the difference of the observed precipitation and the actual evapotranspiration to be considered relevant. The use of this filter led to a strong increase in model performance, especially since parameter sets that resulted in an overestimation of the discharge volume could be removed this way. However, due to the added uncertainties that come with the application of the water balance filter, the authors recommended using a wide range for the filter. There was no evidence found that the irregularity of the data resulted in a worse model calibration than this would have been the case for regular observations (Weeser et al., 2019).

The method used by Weeser et al. (2018, 2019) is based on the Social.Water technology (Fienen & Lowry, 2012), invented as a part of the Crowdhydrology project. Crowdhydrology is a long-lasting hydrological citizen science project in the United States of America. It started with nine locations at which citizen scientists can read the water level from a ruler installed in a stream and send their observation via text message to the research team (Lowry & Fienen, 2013). During the 2010s, the observation network was expanded to a national network (Lowry et al., 2019). The data obtained in the Crowdhydrology project revealed a high quality (Lowry & Fienen, 2013) and was shown to be useful for hydrological model calibration, even though the majority of observations were made during low flow and medium flow conditions while only few citizen scientists contributed observations during rain events (Avellaneda et al., 2020; Lowry & Fienen, 2013).

The missing water level observations during and shortly after rain events and the irregularity of the data were found to be obstacles for the calibration of a hydrological model in a study by Luffman & Connors (2022). They investigated the value of water level observations in a flashy catchment in Tennessee: The observations made by citizen scientists were insufficient for a reliable model calibration in this case. The main reason for this could be found in the irregularity of the data. In the study catchment, precipitation events in the preceding 15 minutes have the largest influence on the amount of discharge. However, observations by citizen scientists were mainly received during the day and in sunny weather conditions. Thus, the authors found that an increased amount of observations during all weather conditions would be required to obtain a reliable model calibration for this flashy catchment (Luffman & Connors, 2022). The finding that synthetic citizen science data, i.e., observations with a low frequency, may be less feasible for flashy catchments than for less flashy catchments was also made in a study about the value of such data for the calculation of basic discharge statistics (Davids et al., 2017).

To overcome potential difficulties that arise when only one type of data is used for the calibration of a hydrological model, several authors such as Avellaneda et al. (2020) or Seibert & McDonnell (2015) highlighted the importance of using different types of data for the calibration of hydrological models. In a study by Starkey et al. (2017) for example, trained participants were asked to observe rainfall amounts, the water level in a stream as well as floods if they occurred. The authors found that these observations were only valuable in their case if they were combined with traditional data, i.e., if the citizen science observations were used to fill gaps and characterise the catchment of interest on a local scale. Furthermore, similar as Seibert & McDonnell (2015) who included soft data in their modelling framework, this study also highlighted the value of combining quantitative data with qualitative knowledge (Starkey et al., 2017). A similar conclusion was also drawn by Le Coz et al. (2016), who summarised three hydrological citizen science projects in Argentina, France and New Zealand: The authors recommended combining data obtained explicitly for such a project with similar information (e.g., videos and photos from high flow events) that is available from other sources such as social media (Le Coz et al., 2016).

Walker et al. (2016) showed that observations of rainfall, discharge and groundwater levels by citizen scientists are valuable to reduce uncertainties in traditional data. Mazzoleni et al. (2017) stressed that new data types such as citizen science observations should be used to compensate for the lack in traditional hydrological data, even though these may be irregular in their temporal resolution and variable in their accuracy. Other than Etter et al. (2020b), these authors found that the temporal variability in the observation intervals has a smaller influence than the accuracy of the data (Mazzoleni et al., 2017). Just as Juston et al. (2009) described the redundancy of traditional discharge measurement time series, these authors found that additional observations of citizen scientists become redundant at some point, but that it is difficult to define the amount of data that is required in a specific situation in advance (Mazzoleni et al., 2017).

3.2 The CrowdWater project

The CrowdWater project (www.crowdwater.ch) is a hydrological citizen science project at the University of Zurich that has been running since 2016. In the CrowdWater project, citizen scientists collect hydrological data all around the world and help to check and improve the data quality of parts of the data in an online game. Furthermore, CrowdWater aims to explore the value of hydrological data that can be obtained by citizen scientists. The overarching goal of CrowdWater is to develop a tool for the collection and control of hydrological data by citizen scientists, whereby these data can be used for hydrological modelling purposes. The methodology developed in the CrowdWater project should be applicable in data-scarce and remote regions and provide a low-cost opportunity to improve hydrological forecasts, water management and decision-making (see also the project descriptions on the homepage of the Swiss National Science Foundation: www.p3.snf.ch/Project-163008 for the first phase of the project and www.p3.snf.ch/Project-192125 for the second phase of the project). The CrowdWater project is supported by the Swiss National Science Foundation.

The CrowdWater app (Seibert et al., 2019), developed in collaboration with the SPOTTERON GmbH in Vienna (www.spotteron.net), serves as a tool for the collection of hydrological observations by citizen scientists since February 2017. Using the CrowdWater app, citizen scientists can observe hydrological variables in several categories:

- Water levels of streams can be read from virtual staff gauges (see section 3.2.1) as well as from physically installed staff gauges.
- The flow state of temporary (i.e., intermittent) streams can be assessed qualitatively.
- The soil moisture of any unsealed surface can be assessed qualitatively.
- General information about water bodies, including information about the observed water quality, can be entered.
- The pollution of waterbodies with macroplastics can be recorded.

For all these categories, no measurement devices are required to do an observation. Thus, the approach is arbitrarily scalable. Observation stations, so called "CrowdWater spots", at which observations in one of these categories are conducted can be started everywhere in the world. Anyone having the CrowdWater app installed on a smartphone or using the web-version of the CrowdWater app (www.spotteron.com/crowdwater) can start new and update all available CrowdWater spots.

This thesis focuses on the observation of stream levels using virtual staff gauges. This category was the main focus of the first PhD students in the CrowdWater project (Etter, 2020; Strobl, 2020). Before the CrowdWater app was available, the collection of water level data using virtual staff gauges had already started at so called "pen and paper" stations (Etter et al., 2020a). At these pen and paper stations, the virtual staff gauge is shown on a printed picture of the stream and citizen scientists can add an observation by filling in a form. Here, data from the app as well as data from these pen and paper stations are used (see section 4.4).

Since the launch of the CrowdWater app, a total of more than 33'000 observations was collected in 68 different countries and at almost 6000 individual CrowdWater spots. In the virtual staff gauge category, more than 13'000 observations were collected at more than 1600 CrowdWater spots. At 153 of these spots, 10 and more updates were made (all numbers from April 2022).

3.2.1 The virtual staff gauge

The so called "virtual staff gauge" builds the centrepiece of the water level class observations in the CrowdWater project. A virtual staff gauge is a virtual sticker containing ten water level classes (Figure 1). Depending on the flow conditions at the time of setting up a new CrowdWater spot, the green (low flow), the yellow (medium flow) or the red (high flow) sticker should be chosen. This sticker is then virtually glued onto the photo of the stream at the time of the first observation. Thereby, the wavy line at the water level class zero should be adjusted to the horizontal line formed by the water surface when the photo is taken perpendicular to the stream. Each time the spot gets updated, the current water level is compared to the virtual staff gauge on the reference picture and the corresponding water level class is determined. In an optimal reference picture, some reference object (such as a stone or a bridge pillar) is visible in the background. This makes the determination of the current water level class easier (Seibert et al., 2019).

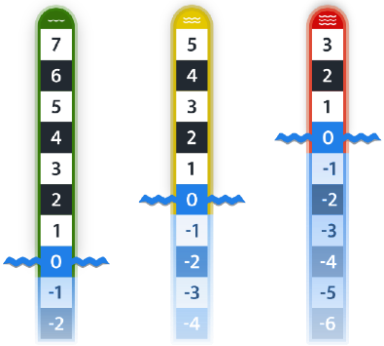


Figure 1: The three virtual staff gauges used in the CrowdWater project (depending on the flow conditions at the time of starting a new CrowdWater spot). Design by Spotteron Citizen Science GmbH.

The left side of Figure 2 shows an example of a virtual staff gauge that has been "installed" at the Ova dal Fuorn in Zernez in July 2017. The picture on the right side shows the same location in September 2021. In the left part of this picture, some reddish parts of the rock are visible which are not visible on the picture with the virtual staff gauge. These rocks can be used as a reference to determine the water level class of the stream. They show that the water level on the picture on the right side is lower than on the picture on the left side. This resulted in a water level class estimate of -1.

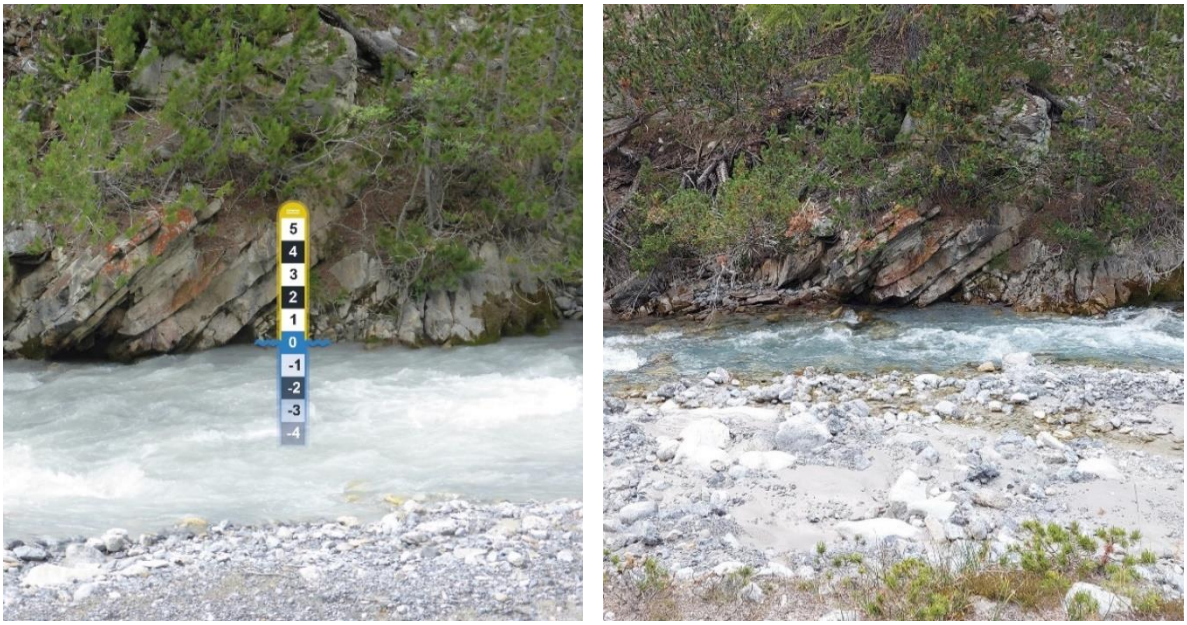


Figure 2: CrowdWater spot at the Ova dal Fuorn in Zernez. Left side: original image with the virtual staff gauge (Simon Meili-Etter, 25.07.17); right side: update for which the water level class -1 was estimated (Mirjam Scheller, 14.09.21).

Several of these water level class observations at the same CrowdWater spot result in a time series of relative water level classes. The absolute water levels are site specific and cannot be determined from the observations using the virtual staff gauge (Seibert et al., 2019). However, relative values are enough to model the discharge behaviour in a stream (van Meerveld et al., 2017; Seibert & Vis, 2016). The dynamics of the water level can be recorded well using the virtual staff gauge (Figure 3).

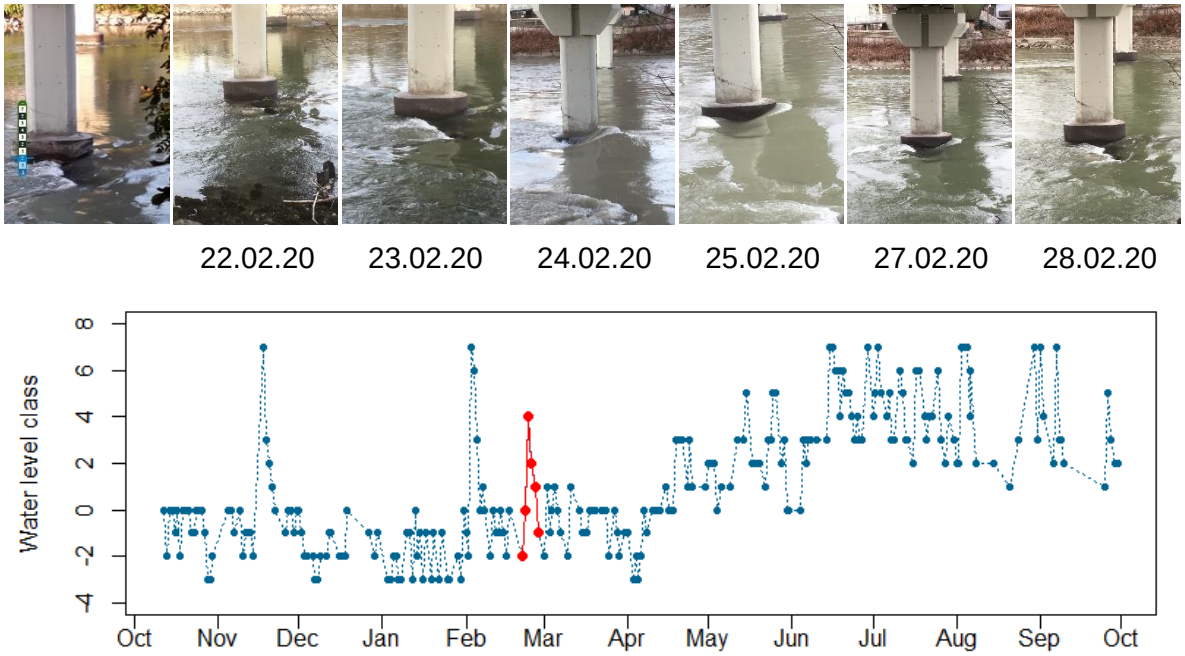


Figure 3: Example time series of relative water level classes. Top: Picture with the virtual staff gauge at the Salzach in Salzburg (by Elisabeth Strobl) and update pictures from February 2020 (by Karin Ebermann). Bottom: All water level class values observed at this CrowdWater spot between October 2019 and September 2020. The observations shown on the pictures are indicated in red. Note that the tick marks mark the beginning of the month.

The virtual staff gauge ensures that water levels can be observed at each stream. A new observation site can be set up quickly and easily, there is no maintenance of the station required and the approach can be applied all over the world (Seibert et al., 2019; Strobl et al., 2020a). Even though the vertical resolution of the virtual staff gauge is limited and the time step between the observations is irregular, it was shown that water level class data can be useful for hydrological model calibration (Etter et al., 2020b; van Meerveld et al., 2017). Furthermore, while the direct estimation of discharge by untrained citizen scientists is inaccurate (Strobl et al., 2020a) and useless for hydrological model calibration (Etter et al., 2018), the water level class observations using the virtual staff gauge show a high accuracy, especially for streams in which the location of the virtual staff gauge is relatively close to the shore from which the observation is conducted (e.g., if the opposite shore is not too far away or if some bridge pillar or rock in the stream is used) (Strobl et al., 2020a).

3.2.2 The CrowdWater game

To control and if necessary improve the quality of the water level class data uploaded to the CrowdWater app, the CrowdWater game (Strobl et al., 2019) is part of the CrowdWater project. In the CrowdWater game, citizen scientists are asked to determine the water level class on a picture compared to the corresponding picture containing the virtual staff gauge. The CrowdWater game is played online (www.cwgame.spotteron.net) and can be joined independently from contributing data to the CrowdWater app. There is a championship in the CrowdWater game each month, consisting of 28 rounds that last for one day each. In each round, a player compares twelve pictures and decides on the water level class on the picture without the virtual staff gauge (Figure 4). Players can collect points by playing the CrowdWater game and win prizes in each championship. For more details on the gamification aspects of the CrowdWater game, the reader is referred to Strobl et al. (2019).

The "wisdom of the crowd" (Surowiecki, 2004) generally leads to an improvement of the quality of the water level class data in the CrowdWater game. This could be shown earlier in the project by

considering all picture pairs with at least 15 votes from the game. The value resulting from the trimmed mean (10% on each side) of all votes for this picture pair was usually better in accuracy and resolution compared to the original value entered by a citizen scientist in the app. The threshold of 15 votes was shown to be reasonable as the resulting value was stable then and did not change significantly anymore when more votes were added (Strobl et al., 2019).

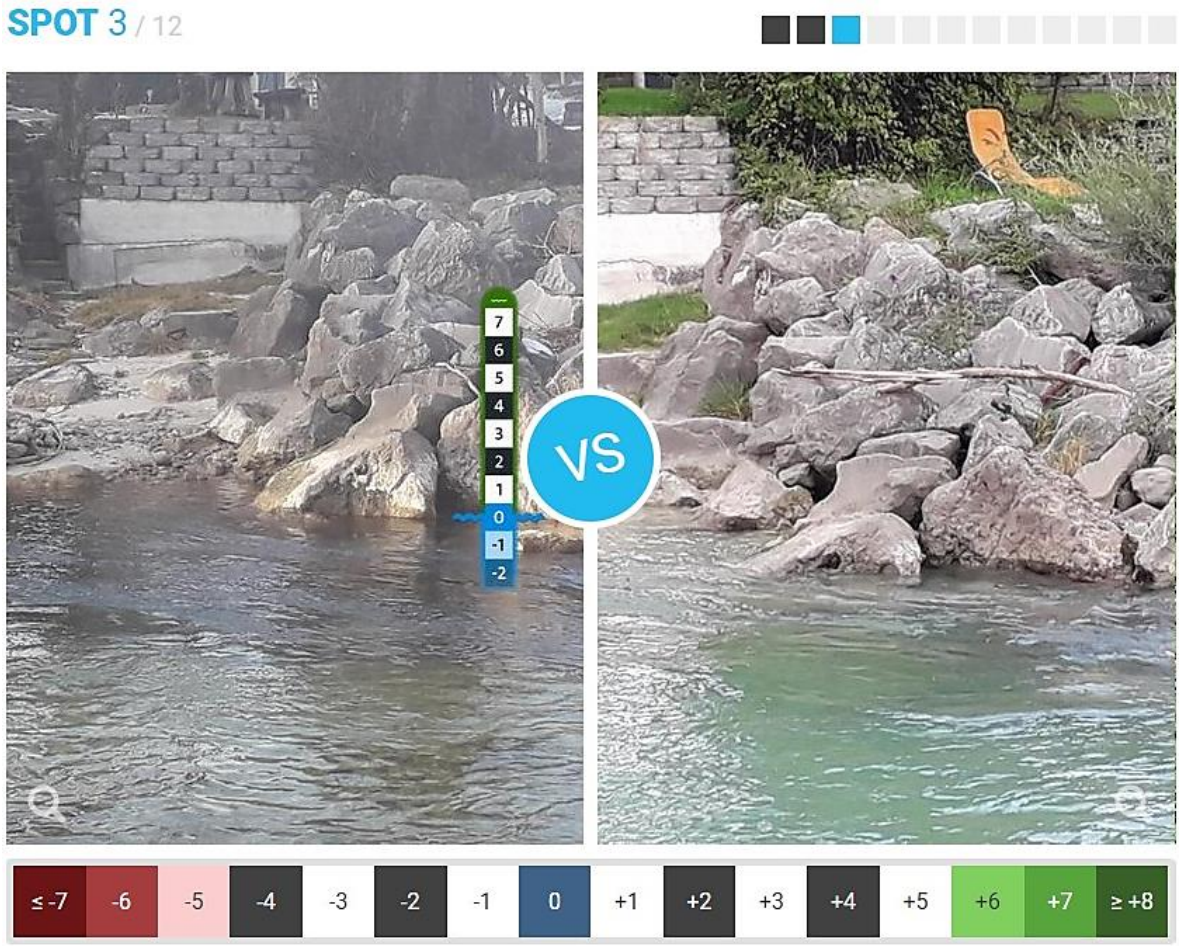


Figure 4: Screenshot from the CrowdWater game. The original image with the virtual staff gauge is shown next to an image of the same spot at a later point in time. The player chooses the water level class on the number bar shown on the bottom of the screenshot. Example from the spot at the Koenigsseeache. Both pictures by Elisabeth Strobl.

With the CrowdWater game, unusable contributions (e.g., spots in which the virtual staff gauge is not inserted correctly or update pictures from a different site) can be filtered out using a report function (Strobl et al., 2019). Furthermore, the CrowdWater game can serve as a training opportunity for new citizen scientists in the CrowdWater project, before they start using the app: Citizen scientists that play the CrowdWater game make less errors when inserting the virtual staff gauge in the app for the first time (Strobl et al., 2020b).

Most pictures uploaded in the virtual staff gauge category of the CrowdWater app were not part of the CrowdWater game yet: Compared to the citizen scientists that upload observations to the app, the number of CrowdWater game players is relatively small. From the 11'650 pictures that could potentially be part of the CrowdWater game, only just below 3000 pictures have reached at least 15 votes (until May 2022).

4 Methods

4.1 Study catchments

The study catchments used for investigating on the research questions were chosen according to two criteria:

- Availability of at least hourly discharge data from an official measurement station. Furthermore, availability of meteorological data such as precipitation and temperature measurements.
- Availability of citizen science data from a CrowdWater spot relatively close to the official measurement station.

According to these two criteria, nine catchments in Switzerland and two catchments in Austria were chosen as study catchments. Three of these catchments are subcatchments of a larger study catchment: The Koenigsseeache is a subcatchment of the Salzach, the Alp is a subcatchment of the Sihl and the Kleine Emme serves as a research catchment twice, whereas the measurement station in Werthenstein is located about 15 km upstream from the measurement station in Emmen (Figure 5).

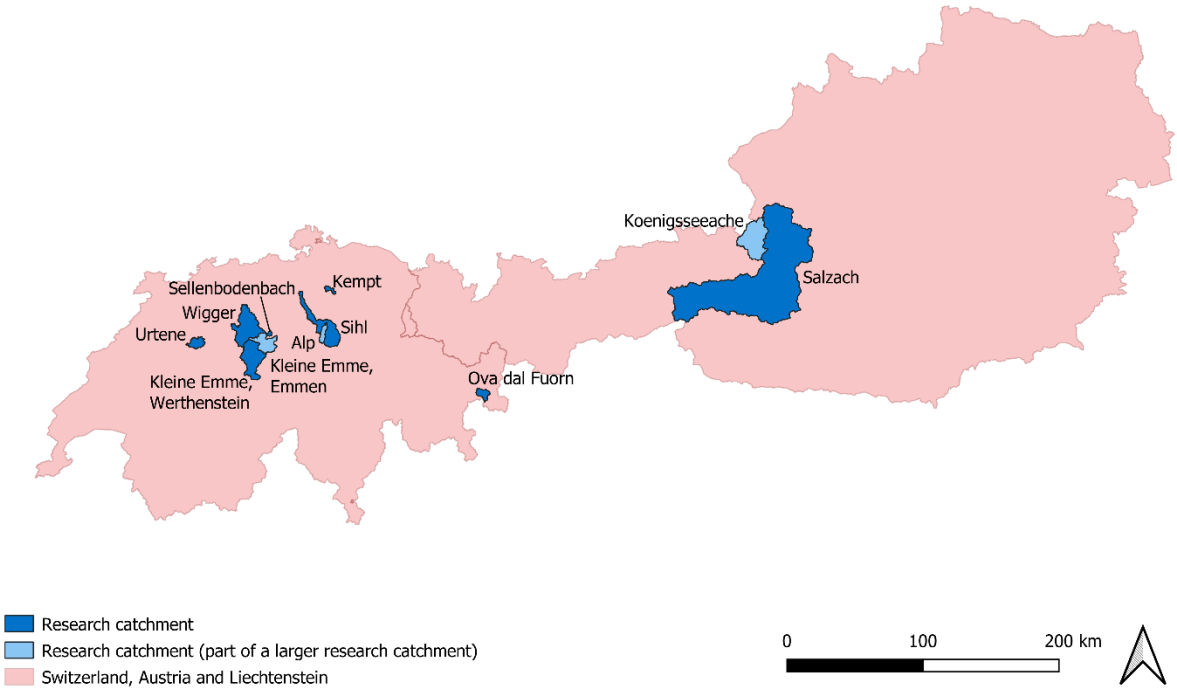


Figure 5: Map showing the locations of the eleven study catchments in Switzerland and Austria. Data sources: Swisstopo (outline of Switzerland and Liechtenstein), GADM (outline of Austria), Hydrological Atlas of Switzerland (Swiss catchments), Open Data Austria (Austrian catchments).

Table 1 gives an overview of the properties of the eleven catchments:

- The catchment areas and the elevation ranges were extracted from the *Hydrological Atlas of Switzerland* for the Swiss catchments. For the Austrian catchments, the catchment areas were provided by the *Hydrological Service Salzburg* and the elevation ranges were extracted from the EU-DEM provided by the *Copernicus Land Monitoring Service at the European Environment Agency*.

- Mean annual temperature, precipitation and mean discharge values were calculated from the meteorological time series used for modelling in each catchment. For a detailed explanation on the sources of these data, see section 4.2.
- The distances between the CrowdWater spots and the corresponding official discharge measurement stations represent the beeline and were calculated using the coordinates stored in the CrowdWater app and given by the authorities, respectively.

Table 1: Main characteristics of all study catchments, sorted by the number of citizen science data points available.

Catchment	Area [km^2]	Elevation range [m a. s. l.]	Mean annual temperature [°C]	Mean annual precipitation [mm]	Mean discharge [m^3/s]	Distance to station [m]
Koenigsseeache, Niederalm	429	495-2642	6.2	1075	12.4	1690
Salzach, Salzburg	4394	421-3395	5.1	1500	181	3830
Kempt, Fehraltdorf	22.5	489-932	9.7	1167	0.4	630
Urtene, Kernenried	73.9	487-896	9.8	1052	0.8	5290
Alp, Einsiedeln	46.7	660-1783	6.9	1535	2.0	2590
Kleine Emme, Werthenstein	311	525-2290	7.2	1313	10.3	20
Ova dal Fuorn, Zernez	55.3	1666-3114	0.2	749	1.0	280
Kleine Emme, Emmen	478	425-2290	7.8	1284	14.3	70
Wigger, Zofingen	366	419-1393	9.5	1075	4.9	0
Sellenbodenbach, Neuenkirch	10.4	510-832	10.2	1064	0.2	30
Sihl, Zurich	343	402-2223	7.6	1480	6.8	200

The four catchments printed in bold and italics are the catchments for which the third sub-question (citizen science data checked in the CrowdWater game) was answered additionally to the other three research questions. Wherever the catchments are listed in this thesis, they are sorted according to the number of available citizen science data points, i.e., as in Table 1. The number of citizen science data points per catchment are given in Table 2. More details on the citizen science data in general are given in section 4.4.

All study catchments can be found in the humid region of Central Europe. However, the altitudes where the catchments are located at vary among the catchments and so do the temperatures and precipitation amounts (see Table 1 and Figure 6). The Ova dal Fuorn catchment, which shows the highest mean elevation, is located in the dry region of the Engadine and thus does not have precipitation amounts as high as it could be expected from the elevation alone.

The different conditions of the study catchments lead to different discharge regimes that are well reflected in the mean discharge behaviour averaged over the years of interest (Figure 7). The nival discharge regime (snow dominated) can be found in the Koenigsseeache, the Salzach and the Ova dal Fuorn. The Alp, the Kleine Emme (both stations) and the Sihl show a mixed signal of rain and snowmelt in their discharge behaviour and are thus catchments with a nivo-pluvial discharge regime. The pluvial discharge regime (rain dominated) can be found in the Kempt, the Urtene, the Wigger and the Sellenbodenbach. The flow duration curves indicate how often a certain discharge is equalled or exceeded. Here, the flow duration curves for all study catchments over the eight years of interest are given normalized and in logarithmic representation, thus they indicate how often a certain discharge in relation to the median discharge in a catchment is equalled or exceeded (Figure 8).

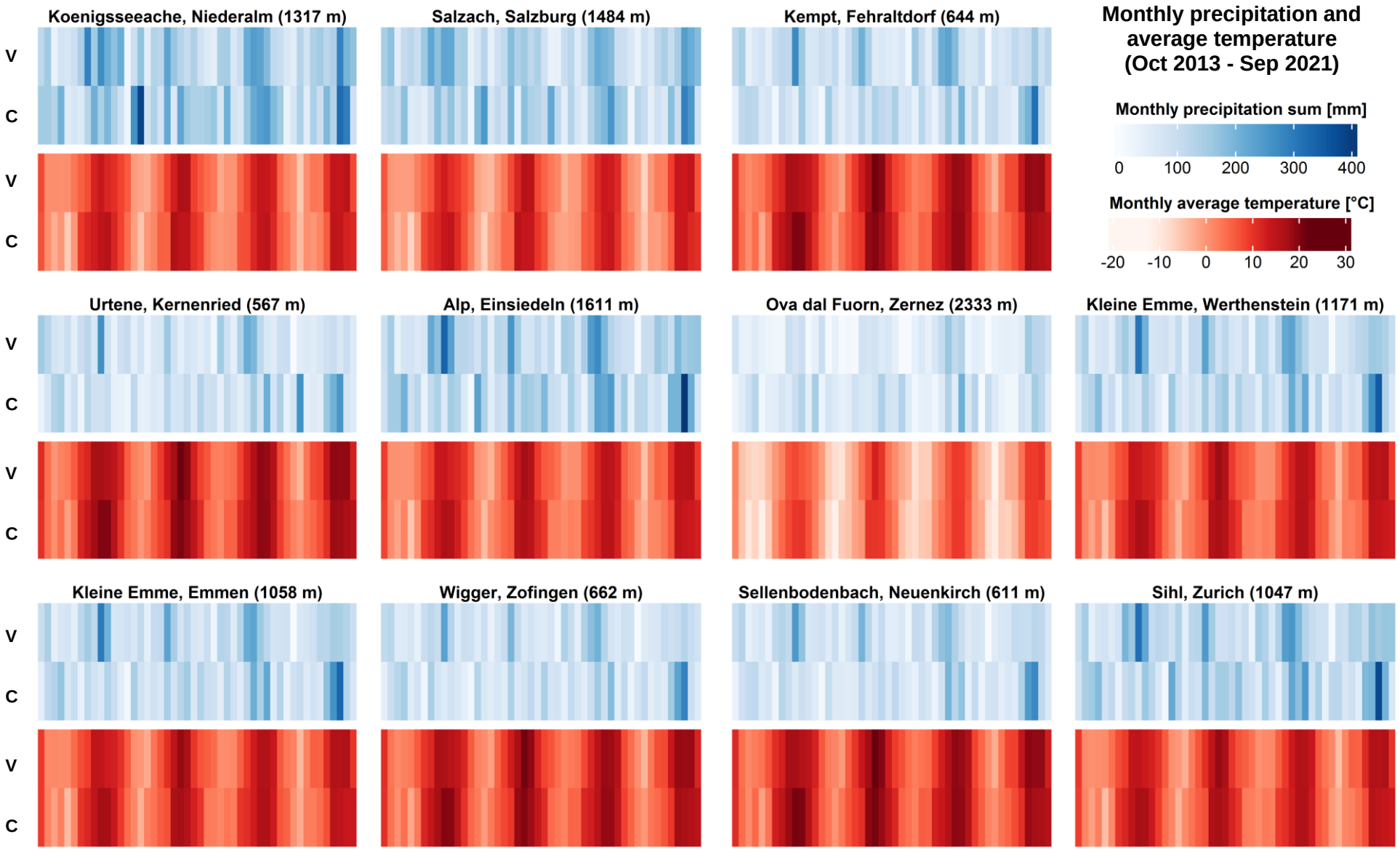


Figure 6: Overview of the meteorological conditions during the years used for validation (top row) and calibration (bottom row) for all catchments. Each stripe represents one month (Oct 2013 at the top left, Sep 2017 at the top right, Oct 2017 at the bottom left, Sep 2021 at the bottom right). The number after the catchment name is the mean elevation of the catchment in meters above sea level.

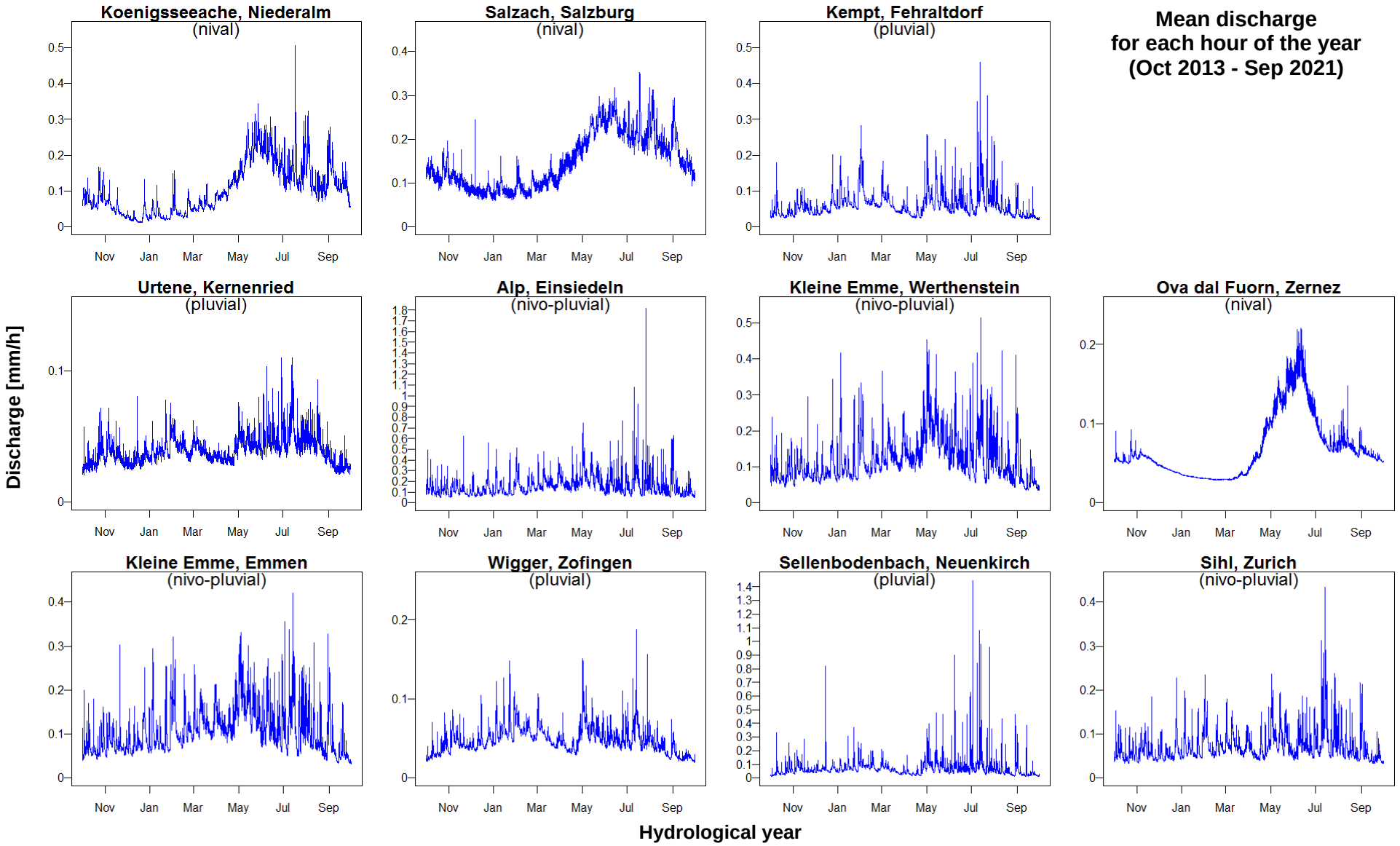


Figure 7: Average discharge for each hour of the year for each study catchment (calculated for the eight hydrological years of the validation and calibration period). The discharge regime type is given in brackets below the catchment name. Note that the y-axis differs for each catchment. The tick marks on the x-axes mark the beginning of the month.

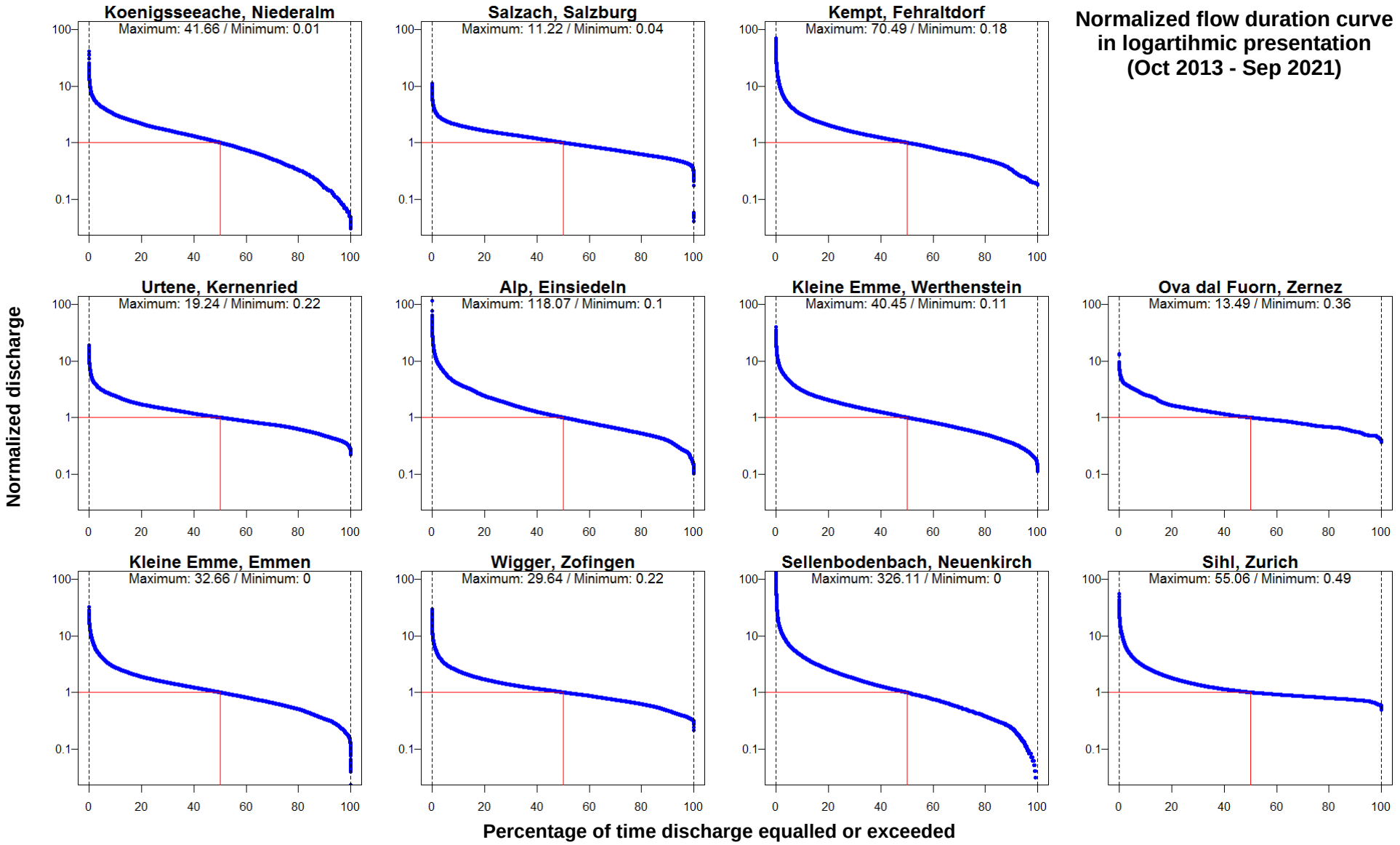


Figure 8: Flow duration curves of all study catchments, with flow normalized by the median flow (i.e., the median flow has a value of 1). The maximum and the minimum normalized flows during the eight years of interest are written below the catchment name. Note that the y-axis is logarithmic.

For all study catchments, the Richards-Baker Flashiness Index (I_{RB}) was calculated using the formula

$$I_{RB} = \frac{\sum_{i=1}^{n} |q_i - q_{i-1}|}{\sum_{i=1}^{n} q_i} \quad (1)$$

given on page 506 in Baker et al. (2004). Thereby, daily discharge values for the eight hydrological years ranging from 1 October 2013 to 30 September 2021 were used for calculation. According to the authors, the index has a low interannual variability and measures the oscillation of the discharge in a stream compared to the total discharge in the same stream (Baker et al., 2004). In the formula given above, q_i is the discharge on a certain day, whereas q_{i-1} is the discharge on the day before. Thus, the index reflects the changes in discharge from one day to another in relation to the total discharge over the period of interest (here, eight years). Streams with a flashy discharge behaviour, thus fast and frequent changes in their discharge during rain events, show a higher value of this index. A low value results for streams in which the discharge does not change very quickly and very often but the discharge behaviour is more stable.

Furthermore, the Baseflow Index (BFI) was calculated according to the procedure given in Gustard et al. (1992) on the pages 21 to 23. Again, daily discharge values from the eight hydrological years of interest were used to do so. As recommended, the BFI was calculated for the whole period of interest and not averaged from yearly BFI calculations (Gustard et al., 1992). The BFI gives the ratio between the area under the baseflow line and the area under the hydrograph. Thus, a BFI close to 1 represents the situation in which the hydrograph does not deviate strongly from the baseflow line. On the other hand, a small value of the BFI represents the situation in which the baseflow line and the hydrograph differ strongly, thus, the catchment shows a rather flashy behaviour.

The values of the two indices for each study catchment are given in appendix 10.1. Note that the two indices are negatively correlated, as a flashy catchment results in a high I_{RB} and a low BFI and vice-versa (Figure 9).

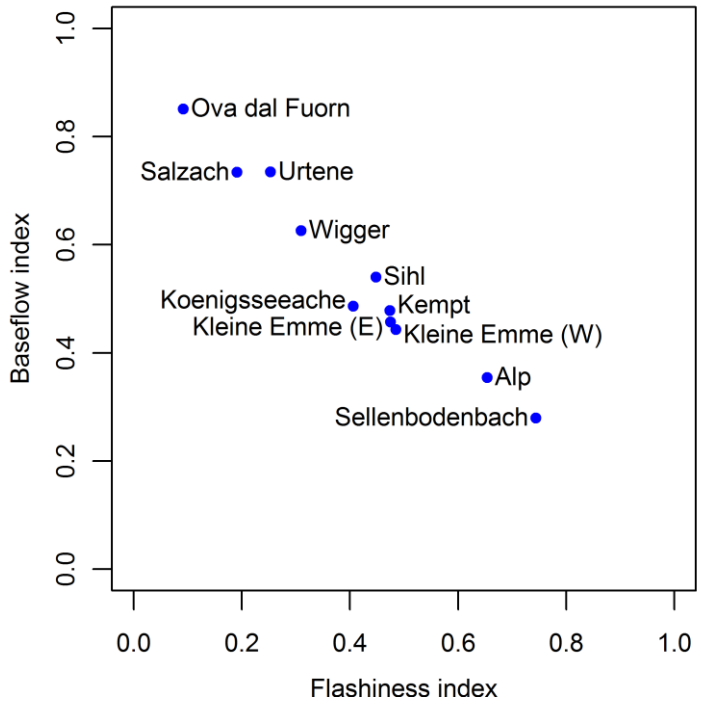


Figure 9: Richards-Baker flashiness index and baseflow index of all study catchments.

4.2 Meteorological data

The time series of precipitation and temperature as well as the values used to account for evaporation should represent the average conditions in the catchment and are used as input data when simulating the catchment discharge in a hydrological model. When using gridded data for these variables, no decision on the interpolation method needs to be made since the data is already interpolated using the best standards. Thus, wherever possible, gridded data and not station data was used to calculate the time series with an hourly granularity for precipitation and temperature as well as the evaporation values for the study catchments.

4.2.1 Precipitation

In Switzerland as well as in Austria, a gridded data set with hourly granularity containing the best precipitation estimates over the whole area of the country is available. The data set *CombiPrecip* by MeteoSwiss (Sideris et al., 2014) was used to determine hourly areal precipitation sums for the study catchments in Switzerland. *CombiPrecip* is a grid data set with a resolution of 1 km^2 per pixel containing hourly precipitation data for Switzerland since 2005. The precipitation values are calculated from rain gauge and radar backscatter data (Sideris et al., 2014).

For the Austrian study catchments, the grid data set *INCA* by the ZAMG was used. The spatial resolution of *INCA* is also 1 km^2 per pixel and the data set and corresponding system is developed for nowcasting of weather conditions, especially in mountainous terrain (Haiden et al., 2011), where the Koenigsseeache and the Salzach are located.

The areal precipitation sums calculated from the gridded data sets were directly used in the precipitation time series of the study catchments, covering the time from 1 October 2013 to 30 September 2021. The mean elevation of the catchment was assigned to the time series as measurement elevation.

4.2.2 Temperature

The gridded temperature data available from MeteoSwiss has a daily granularity. Thus, a combination of the gridded data and hourly station data was used to obtain hourly temperature data for all catchments. The station data was downloaded from IDAWEB, a data portal of MeteoSwiss. The following workflow conducted for each study catchment led to the hourly temperature time series:

1. Using Thiessen polygons for all temperature measurement stations in Switzerland, the relevant temperature measurement stations for each catchment were determined.
2. All relevant temperature measurement stations within and around the catchment were assigned a weight according to the proportion of catchment area located within the Thiessen polygon of the temperature measurement station. Temperature measurement stations for which the proportional area was smaller than 3% of the catchment area were excluded. The small parts of the catchment area corresponding to these temperature measurement stations were distributed among the remaining temperature measurement stations, thus these weights slightly increased. The resulting stations per catchment as well as their weights are listed in appendix 10.2.
3. The measured hourly temperature time series of the stations were shifted according to the difference between the mean elevation of the catchment and the elevation of the temperature measurement stations with a lapse rate of -0.6°C per 100 m (Bergström, 1992).
4. A time series with an hourly granularity was calculated as the weighted average of the shifted temperature time series. Whenever single values were missing in one time series, the weight of this station was equally distributed on the other stations to calculate a value for the time step. For the special case with only one temperature measurement station, see the explanation below.

5. Using the gridded temperature data, the daily average temperature over the whole catchment was determined.
6. The hourly temperature values were shifted such that the average temperature for the whole day was in accordance with the daily average temperature determined in step 5.

For the Ova dal Fuorn in Zernez, only the temperature measurement station in Buffalora (BUF) was determined to be relevant. The measurement time series of this station contained a lack of data of 53 hours in July 2015. This lack was filled with the data measured at this station 95 hours earlier to reach a realistic and smooth temperature fluctuation for the missing time steps.

At the stations in Sattel (SAG) and Fluehli (FLU), the measurements started later than on 1 October 2013 (22 October 2013 at SAG and 19 March 2015 at FLU). Thus, the time series needed to be extrapolated for these stations. To do so, the spatially closest temperature measurement stations (Einsiedeln (EIN) for SAG and Schuepfheim (SPF) for FLU) were used: The mean difference in temperature between the two stations was calculated using all time steps for which data was available from both stations. These mean differences in temperature were used to shift the measurements at the stations with the full time series such that they could be used as extrapolated temperature time series for the stations SAG and FLU.

For the Austrian catchments, gridded data was available and was used to calculate the hourly mean annual temperature. As for precipitation, the *INCA* dataset by the ZAMG was used to do so. The mean catchment elevation was assigned to the resulting temperature time series for all study catchments.

4.2.3 Evaporation

Evaporation data can be put into the HBV model by indicating one value per month, one value per day of the year or one value per modelled time step. In this thesis, the second option was used.

For the Swiss catchments, it was assumed that the potential evaporation E_{pot} $[mm\ d^{-1}]$ follows a sinus curve with its minimum on 21 December of each year. The minimum was set to 0.5 mm per day for all catchments based on the data about actual evaporation in the *Hydrological Atlas of Switzerland* (Menzel et al., 1999). Furthermore, the *Hydrological Atlas of Switzerland* contains values of the net radiation R_n $[W\ m^{-2}]$ in each square kilometre in Switzerland for the years 1983 to 1994 (Z'graggen & Ohmura, 2000). The average of all the net radiation data points within each catchment was taken to be the net radiation value for the catchment. To get from net radiation to potential evaporation, the formula below (Z'graggen & Ohmura, 2000) was used:

$$E_{pot} = \frac{R_n \cdot 8.64 \cdot 10^4\ s/d}{2.256 \cdot 10^6\ J/kg} \tag{2}$$

The sinus curve assumed in the beginning was then scaled, such that the potential evaporation matched the total value calculated from the net radiation over the whole year. This resulted in 365 daily evaporation values for each study catchment.

For the Austrian catchments, the evaporation values of the *WINFORE* data set (Haslinger & Bartsch, 2016) were used to obtain an evaporation average for each day of the year in both catchments. This data set contains information for each square kilometre in Austria from 1961 to today. To be as consistent with the Swiss catchments as possible, the years 1983 to 1994 were used to calculate the average evaporation values for the Austrian catchments, too.

The HBV model corrects the evaporation values by comparing the input temperature data with a long term mean of the temperature. According to the years used to calculate evaporation, the years 1984 to 1993 were used to obtain long term means for all days of the year. February 29 in leap years was skipped, such that 365 values resulted. For the Swiss catchments, daily grid data from MeteoSwiss was used to do so. For the Austrian catchments, the mean temperature values were estimated as the average of the minimal and maximal daily temperature which is contained in the *SPARTACUS* data set (Hiebl & Frei, 2016).

4.3 Discharge data

As "ground truth" for the model, a time series with hourly discharge measurements at the station of interest was required for each study catchment. For all catchments except for the Kempt and the Urtene, hourly mean values were ordered directly by the authorities: The remaining seven Swiss discharge measurement stations are maintained by the confederation (FOEN). The Austrian discharge data was made available by the hydrographic service Salzburg. For the Kempt and the Urtene, the discharge measurement stations are maintained by the cantonal authorities of Zurich (AWEL) and Berne (AWA), respectively. For these catchments, the hourly mean discharge values were calculated from measurement time series with a (partly irregular) interval of up to 10 minutes.

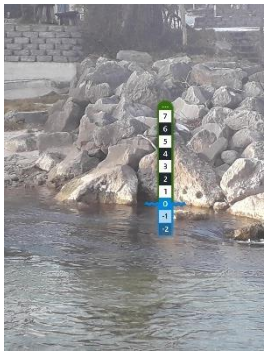
For the simulations, measurements from recent years were used. Therefore, some of the data had not been validated by the authorities yet. For most catchments, this was the case for the data from the years 2019-2021. However, the measurements done by the authorities were the best approximation available, so these data were used as if they had already been validated. See appendix 10.3 for more details on the source of the discharge time series and the amount of unvalidated data per catchment.

4.4 Citizen science data

All citizen science data originate from the CrowdWater project (see section 3.2). The water level class data are collected based on a photo in which a virtual staff gauge is inserted (Figure 10). The virtual staff gauges have different sizes relative to the sizes of the streams. A larger staff gauge makes it easier to choose the correct water level class, while a smaller staff gauge allows to collect data with a higher resolution. This and the observation quality leads to different Spearman rank correlations between the water level class estimates (citizen science data) and the actual discharge. The total number of observations and the correlations are given in Table 2 which is colour-coded according to the value of correlation (blue referring to the highest correlations, followed by yellow, orange, and red).

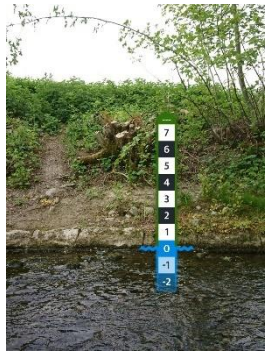
Table 2: Number of CrowdWater observations and Spearman rank correlation between water level classes and discharge measurements for all study catchments. Colour-coded according to the value of correlation.

Catchment	Total number of observations	Spearman rank correlation
Koenigsseeache, Niederalm	1113	0.964
Salzach, Salzburg	632	0.964
Kempt, Fehraltdorf	395	0.907
Urtene, Kernenried	380	0.678
Alp, Einsiedeln	293	0.727
Kleine Emme, Werthenstein	139	0.568
Ova dal Fuorn, Zernez	74	0.470
Kleine Emme, Emmen	69	0.303
Wigger, Zofingen	51	0.699
Sellenbodenbach, Neuenkirch	49	0.113
Sihl, Zurich	41	0.217



Koenigsseeache, Niederalm
25.12.2017
Elisabeth Strobl

Salzach, Salzburg
17.08.2018
Elisabeth Strobl

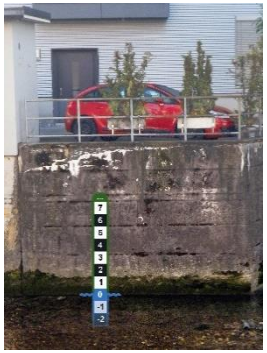


Kempt, Fehraltdorf
22.05.2020
Monika Dietschi Hanselmann

First image of each spot, containing the virtual staff gauge

With date of acquisition and acquiring CrowdWater user

Urtene, Kernenried
19.06.2018
Auria Buchs



Alp, Einsiedeln
26.10.2017
Simon Meili-Etter

Kleine Emme, Werthenstein
23.02.2018
Simon Meili-Etter

Ova dal Fuorn, Zernez
25.07.2017
Simon Meili-Etter

Kleine Emme, Emmen
23.02.2018
Simon Meili-Etter

Wigger, Zofingen
23.02.2018
Simon Meili-Etter

Sellenbodenbach, Neuenkirch
23.02.2018
Simon Meili-Etter

Sihl, Zurich
01.04.2017
Simon Meili-Etter

Figure 10: First photo of each spot, showing the virtual staff gauge as it is used for the water level estimation at this spot. The date of acquisition and the acquiring CrowdWater user are given with each photo. Some of the images were uploaded to the CrowdWater app in a different format and were cropped due to the limited space available.

4.4.1 App data and pen and paper data

All eleven spots exist as virtual observation stations, i.e., observations can be added using the CrowdWater app. The links to the CrowdWater spots on the interactive map by Spotteron can be found in appendix 10.4. At the Alp, at both stations at the Kleine Emme, at the Ova dal Fuorn, at the Wigger, at the Sellenbodenbach and at the Sihl, there were letterboxes and boards installed additionally, so that observations could also be made using a form that has been designed by Barbara Strobl and Simon Meili-Etter (see appendix 10.5). In the app, it is only possible to choose one water level class (integer number). On the forms, citizen scientists sometimes indicated a number in between two classes, for example 0.5. These estimates in between classes were accepted and used as they were, since they represented the best estimate of the water level class at that point in time.

All data collected using the CrowdWater app and the forms at the pen and paper stations were compiled to one time series per catchment. The total number of observations and thus the temporal resolution of these time series varied heavily among the catchments (Figure 11 and Table 2). As it is in the nature of citizen science data, the observations were not made in regular time steps.

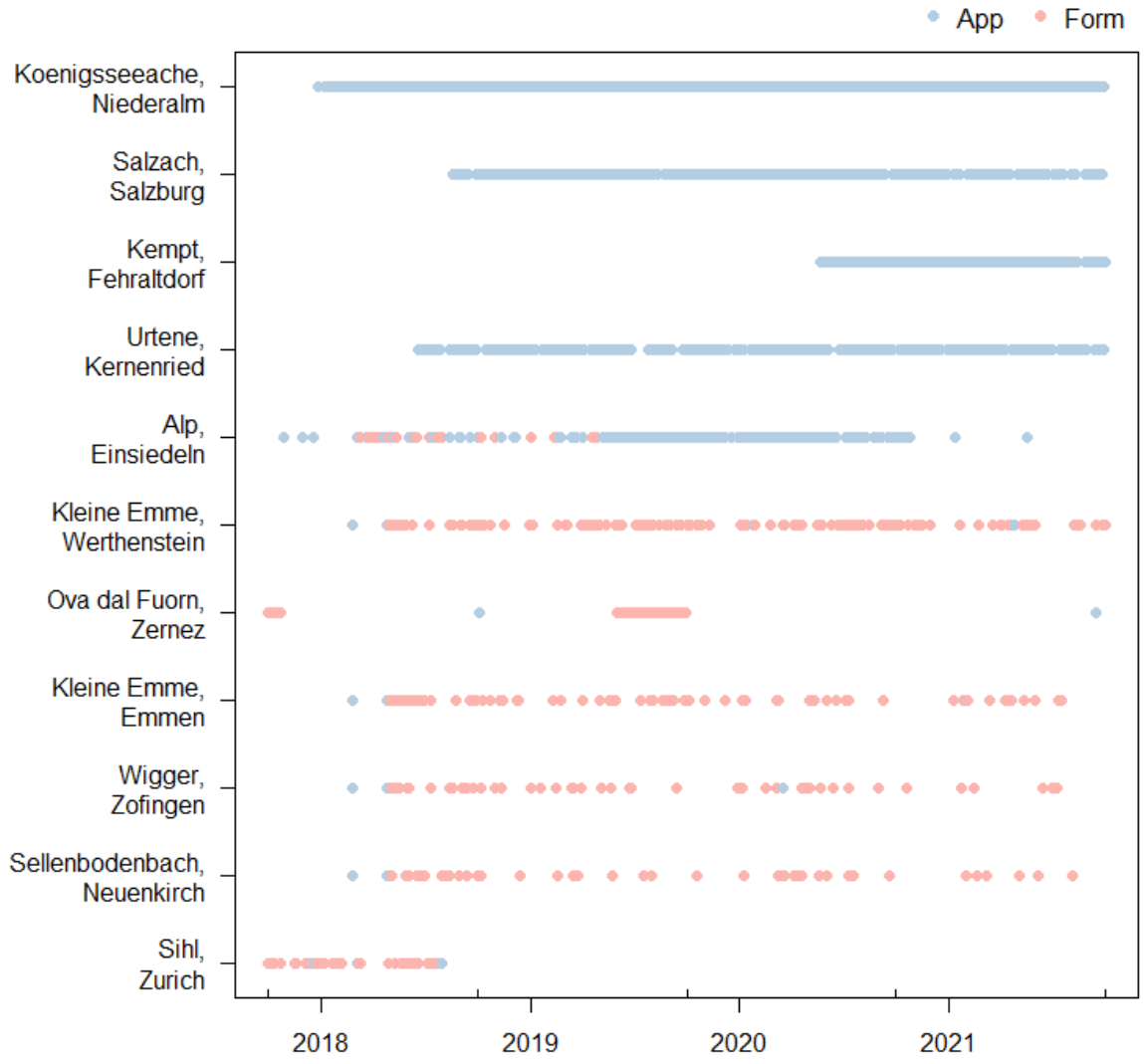


Figure 11: Distribution of the citizen science data collected using the app or the form at the pen and paper stations for each catchment during the calibration period. Note that the tick marks on the x-axis mark the beginning of the calendar years 2018-2021 and the small additional tick marks within the drawing area of the plot mark the beginning of the hydrological years.

The station at the Ova dal Fuorn is hardly accessible in winter. Therefore, the station was only installed during the summer months of the years 2017 and 2019. At the Sihl and at the Alp, the pen and paper stations were removed earlier than the other pen and paper stations. Therefore, the time series ended early at the Sihl. At the Alp, a citizen scientist continued the time series using the CrowdWater app. For the other catchments in which the data was collected using the CrowdWater app (Koenigsseeache, Salzach, Kempt and Urtene), it was also mainly one citizen scientist per station that collected all data. At the remaining (pen and paper) stations, the form was filled in by different people.

The timestamp of each citizen science observation was rounded to one hour, such that each observation could be linked to the discharge measurement at that point in time. This was required to calculate the Spearman rank correlation between the discharge time series and the water level classes (see Table 2) as well as for the calibration of the model using these data (see section 4.7). If there were two observations at the same point in time, the average water level class was used as observed water level class. Thus, aside the estimates in between two classes on the forms, this is a second reason why not only integer numbers were obtained in the water level class time series (Figure 12).

The Spearman rank correlation of the water level class observations (Table 2) as well as the distribution of the water level classes in dependence of the actual discharge measurements (Figure 12), differed among the catchments. Etter et al. (2020a) found that the observations in the app that were mainly done by one citizen scientist were of a higher quality (as the perception bias was consistent) than the observations that were done by many different people using the forms. Furthermore, they found that observations were made during a variety of flow conditions, especially if one person felt somewhat responsible for a station, as it was the case for the app-dominated study catchments in this thesis. These findings were also reflected in the data points available for the eleven study catchments.

Another difference among the study catchments was the different range of water level classes and discharge conditions that was covered by the citizen science data: This has a lot to do with the size of the virtual staff gauge (Figure 10) as well as with the variation that can be observed in the different streams. The ranges of the x-axes in Figure 12 represent the ranges of the discharge conditions during which a citizen scientist made a water level observation. As a comparison, the mean discharge calculated over the eight years of interest, as well as the maximal and minimal discharge value during this period are given in the same unit in Table 3.

Table 3: Mean, maximum and minimum discharge for each study catchment for the eight years used for calibration and validation (October 2013 to September 2021).

Catchment	Mean discharge [mm/h]	Maximum discharge [mm/h]	Minimum discharge [mm/h]
Koenigsseeache, Niederalm	$1.04 \cdot 10^{-1}$	3.02	$9.24 \cdot 10^{-4}$
Salzach, Salzburg	$1.48 \cdot 10^{-1}$	1.40	$5.00 \cdot 10^{-3}$
Kempt, Fehraltdorf	$6.18 \cdot 10^{-2}$	2.67	$6.83 \cdot 10^{-3}$
Urtene, Kernenried	$3.99 \cdot 10^{-2}$	$5.96 \cdot 10^{-1}$	$6.87 \cdot 10^{-3}$
Alp, Einsiedeln	$1.58 \cdot 10^{-1}$	10.06	$8.71 \cdot 10^{-3}$
Kleine Emme, Werthenstein	$1.19 \cdot 10^{-1}$	3.15	$8.66 \cdot 10^{-3}$
Ova dal Fuorn, Zernez	$6.79 \cdot 10^{-2}$	$7.04 \cdot 10^{-1}$	$1.89 \cdot 10^{-2}$
Kleine Emme, Emmen	$1.08 \cdot 10^{-1}$	2.41	$1.51 \cdot 10^{-4}$
Wigger, Zofingen	$4.81 \cdot 10^{-2}$	1.08	$7.91 \cdot 10^{-3}$
Sellenbodenbach, Neuenkirch	$7.22 \cdot 10^{-2}$	11.06	0
Sihl, Zurich	$7.14 \cdot 10^{-2}$	2.45	$2.19 \cdot 10^{-2}$

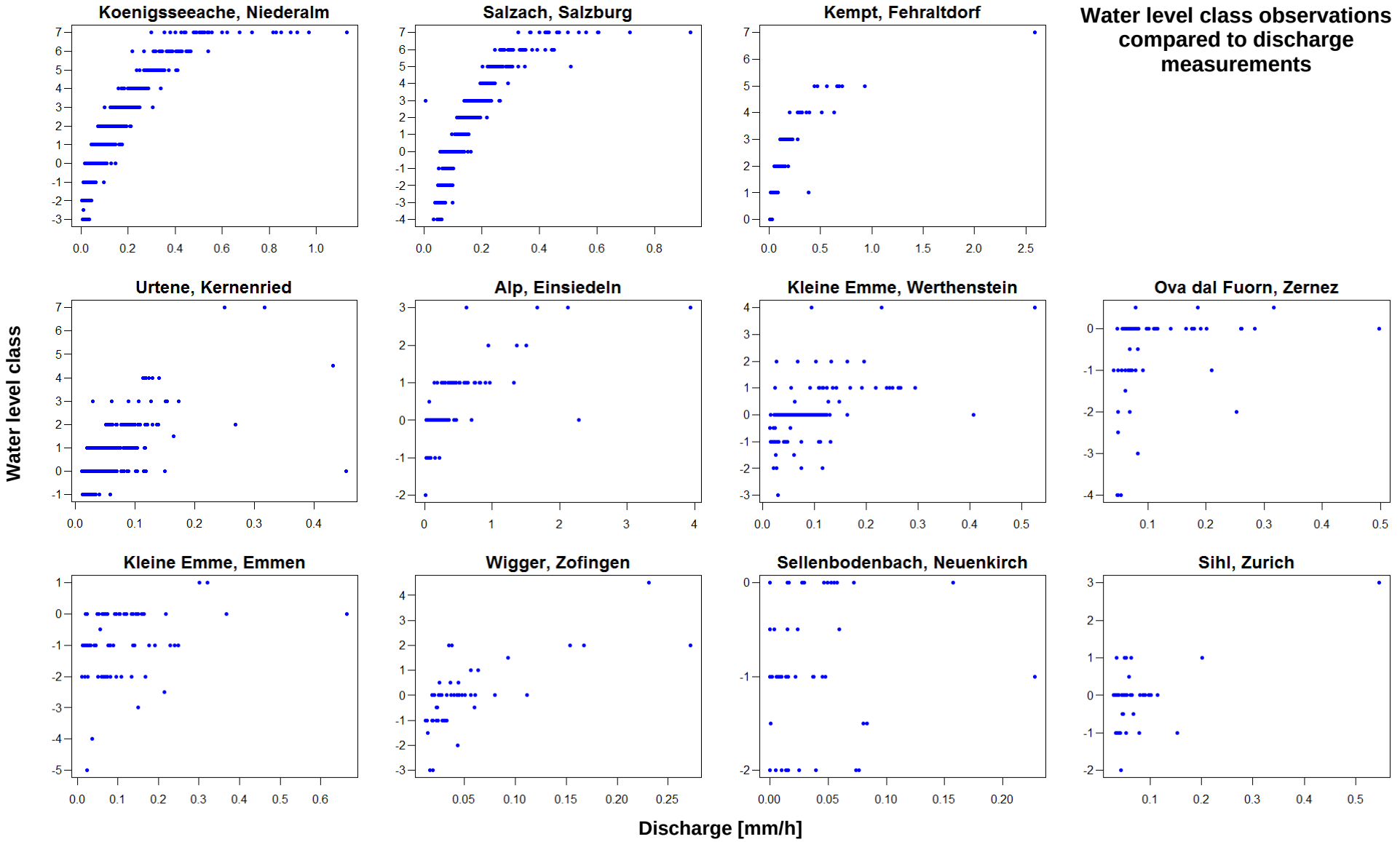


Figure 12: Water level class observations by citizen scientists plotted against the measured discharge at the same time for each study catchment. Note that the scales of the x- and the y-axes (i.e., the range of discharge conditions and water level classes covered) differ for each catchment.

4.4.2 Quality-controlled water level classes from the CrowdWater game

For four catchments, namely the Koenigsseeache, Salzach, Urtene and Alp, the last sub-question was answered using data from the CrowdWater game. For each data point collected in the app at one of these stations, the number of votes in the CrowdWater game on 1 April 2022 was checked. If this number was at least 15, the data point was considered as classified (Strobl et al., 2019). The original water level class was then replaced with the trimmed mean (10% on each side) of the water level class votes in the CrowdWater game. If there were fewer than 15 votes for a data point, the original value from the app was left unchanged. This resulted in a modified water level class time series for these four catchments (Figure 13). Note that the calibration period was shortened by one year (see section 4.6.2) for the approach using game data, therefore only the data points that were obtained before the end of September 2020 are shown in the plots.

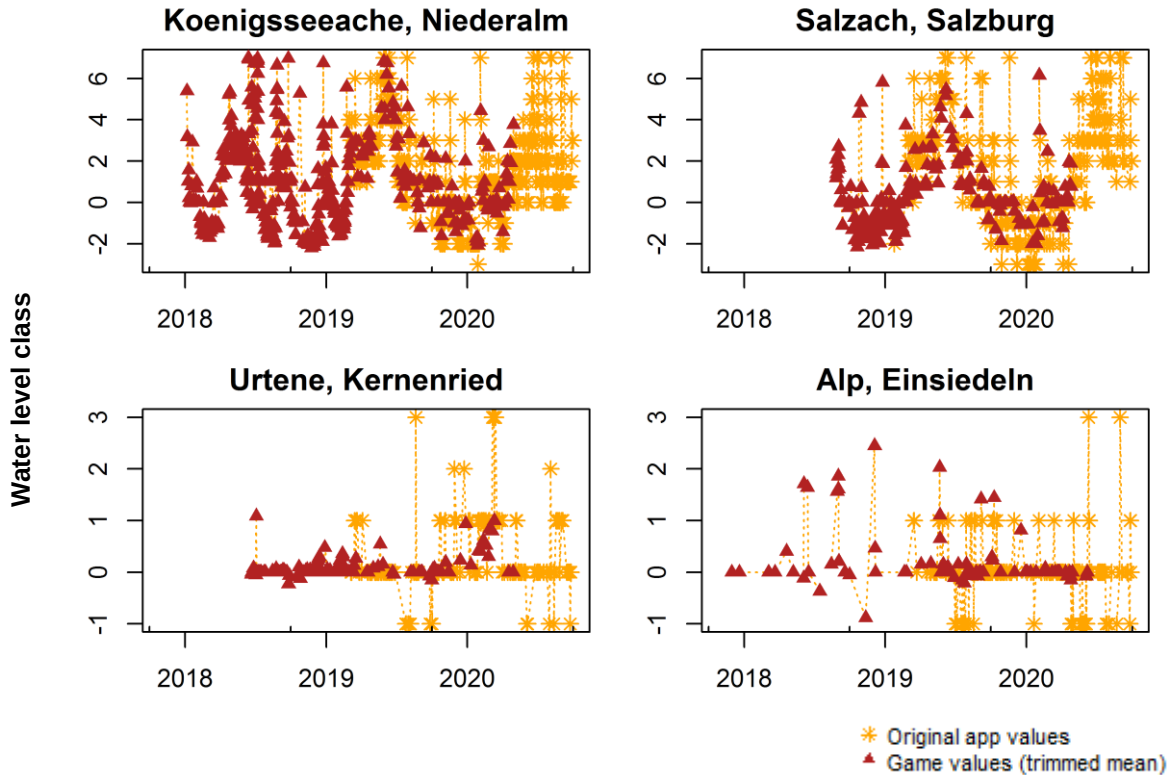


Figure 13: Values from the CrowdWater app that were not classified in the game yet and classified values with higher resolution for all four catchments for which quality-controlled data was available. Note that the scale on the y-axis differs for the two rows. Note that the tick marks on the x-axis mark the beginning of the calendar years 2018-2020 and the small tick marks within the drawing areas of the plots mark the beginning of the corresponding hydrological year.

The amount of data points that could be replaced with the value resulting from the game differed for the four catchments. At the Koenigsseeache 443 out of 861 observations (51%) were classified, at the Salzach, 209 out of 501 observations (42%) were classified, at the Urtene, 113 out of 257 observations (44%) were classified and at the Alp, 82 out of 261 observations (31%) were classified. Note that for none of the catchments the amount of classified data points was below 25%. This is important as the data was used in four steps: If only 25% of the available citizen science data was used to calibrate the model, classified data only (i.e., only red data points) were used. If only 50% of the available citizen science data was used to calibrate the model, the data set still mainly consisted of classified data and was filled up (in all catchments except the Koenigsseeache) with unclassified data. For the cases in which 75% and 100% of the citizen science data was used, all classified data was included and then supplemented with unclassified citizen science data.

For each partition of citizen science data, the Spearman rank correlation between the discharge measurements and the data from the app and the game, respectively, was calculated (Figure 14). Dark blue symbols represent the values for the water level classes collected with the app, light blue symbols represent the values if this data was (partially) replaced by values from the game. The water level class data collected with the app at the Koenigsseeache and the Salzach showed a high Spearman rank correlation with the measured discharge in these streams. The replacement of 443 and 209 data points respectively with values resulting from the game did not have a large influence on the Spearman rank correlation between the citizen science data and the discharge time series for these two catchments. If not the full data set but only a part of it (25%, 50% or 75%) was considered and thus a higher percentage of the data was replaced with checked data from the game, this still did not have any influence on the Spearman rank correlation in these two catchments. That was different at the Urtene, where a clear drop of the Spearman rank correlation could be observed. This drop was even stronger if not the full data set was used but only a part of it (and thus a larger percentage of data was already modified in the game). Note that in the plot corresponding to the Urtene in Figure 13, it is visible that most values are set to 0 or close to 0 when they are changed in the game, while there is a larger variation of water level classes where the data is still original app data. The opposite case could be observed at the Alp: There, the game had the expected influence. This means that in general, higher Spearman rank correlation values were reached if a higher percentage of the considered data set consisted of classified values.

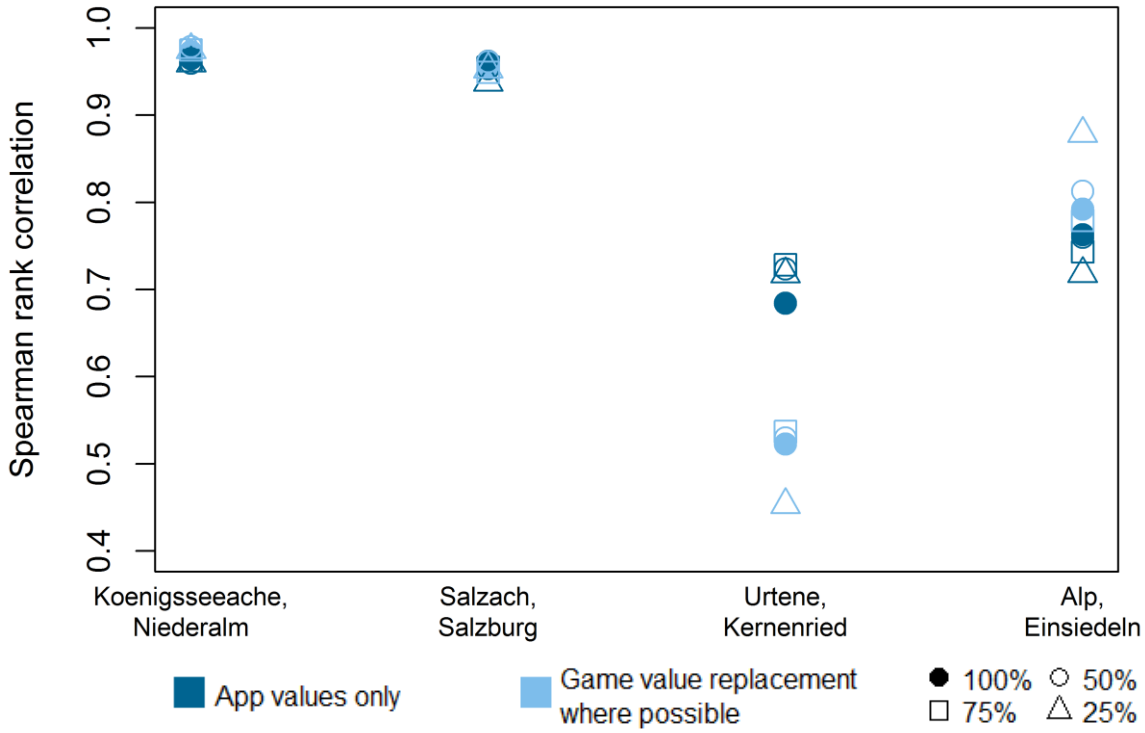


Figure 14: Spearman rank correlation between the citizen science data and the discharge data measured by the authorities. For each set of citizen science data (25%, 50%, 75% and 100% of all available observations), the value of the correlation is shown if only water level class data from the app is considered and if as many data points as possible were replaced with the trimmed mean values from the game.

Note that the Kempt could not be included in this part because there were no data points that were classified by at least 15 votes in the CrowdWater game. The other six catchments were excluded because a major part of the data collected at these spots was pen and paper data and therefore not included in the CrowdWater game. Furthermore, note that the pen and paper data collected at the Alp in Einsiedeln was excluded from this part to have similar conditions at the Alp as in the other three catchments.

4.5 The HBV model

4.5.1 Model description

For all simulations done in this thesis, the HBV model was used. The HBV model is named after the Hydrologiska Byråns Vattenavdelning unit at the Swedish Meteorological and Hydrological Institute. The model is semi-distributed, meaning that several vegetation and elevation zones can be defined for each catchment. Furthermore, the model is conceptual, meaning that its processes are transparent while the requirements for the input data are relatively low (Seibert & Vis, 2012).

In comparison to other hydrological models, the HBV model is of a rather simple structure. The complexity of the model was held low in order to not formulate a too complex model just because the increasing computer capacity allows to do so (Bergström, 1992). While developing the model and increasing its complexity, changes were only accepted when they led to a significant improvement of the model performance. This does further explain the simplicity of the HBV model (Bergström, 1991). More on the history of the HBV model can be found in the retrospective paper by Seibert & Bergström (2022). Most parameters used in the HBV model cannot directly be related to some physical properties of a catchment, thus they can vary heavily for different catchments (Bergström, 1992). As the HBV model is working on a catchment scale, the parameters represent an average value for the whole catchment and should be interpreted as an index representation and not as the true representation of some physically measurable value (Bergström, 1991).

In this thesis, the version HBV light (Seibert & Vis, 2012) was used. In this version, it is possible to run simulations with different than daily time steps and several subcatchments (Seibert & Vis, 2012). However, since citizen science data is only available at the catchment outlets and thus the application of subcatchments is not possible, only the first of these improvements was used in this thesis. Furthermore, following the approach of assessing a good model performance without much knowledge about a catchment and its discharge characteristics, only elevation zones were defined for each catchment (as they can rather easily be determined using remote sensing data), but no vegetation zones. Therefore, the distinguishing between different vegetation zones is not mentioned in the descriptions of the model equations below.

The next four paragraphs give an overview about the different routines the model consists of and introduces the abbreviations for the parameters. These abbreviations and the four routines are also given in the schematic sketch in Figure 15. Note that the parameters used

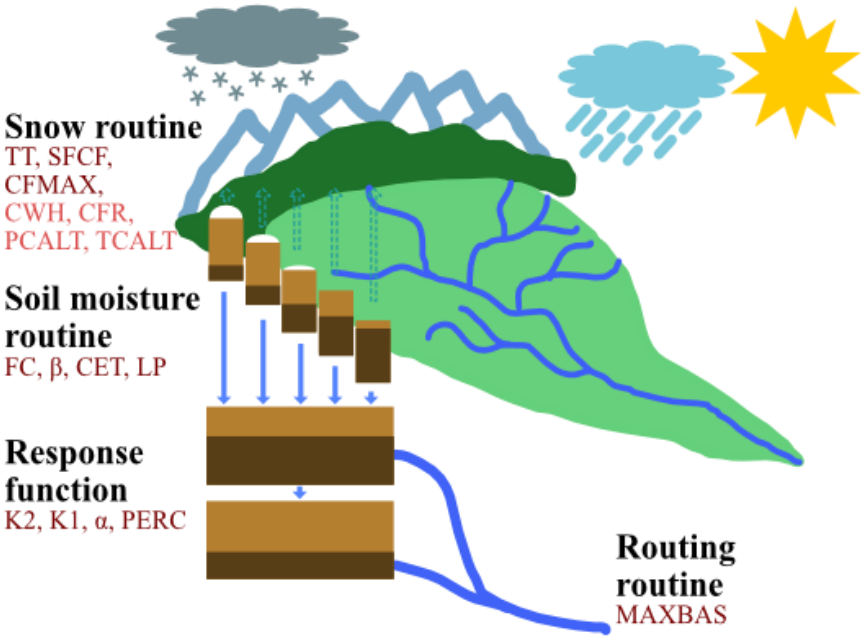


Figure 15: Sketch of the HBV model. Adapted from a drawing by Petra Seibert, available in Seibert & Vis (2012). Parameters shown in light red were fixed to a certain value, parameters shown in dark red were used for calibration.

for calibration in this thesis are printed in dark red while the parameters that were fixed are printed in light red. However, this can be different in other applications of the HBV model. The following description is a summary of the information given in the help section of the HBV model, whereby the units of the parameters are given according to the hourly time step used in this thesis instead of the daily time step that serves as default. Moreover, only the parameters used for the simulations in this thesis are described here. More detailed descriptions of the model routines and the influence of the different parameters can be found in the help section of the HBV model or in other publications (e.g., Bergström, 1995; Lindström et al., 1997; Seibert, 1999; Seibert & Vis, 2012).

4.5.1.1 Snow routine

The snow routine takes the precipitation and temperature data as input and results in the calculation of the snowpack and the snowmelt. The calculations of the snow routine are carried out for each elevation zone separately, whereby the temperature data is corrected with **TCALT** [°C/100m] and the precipitation data with **PCALT** [%/100m].

If the temperature is below the threshold temperature **TT** [°C], the precipitation is modelled to be accumulated as snow. In this case, the precipitation amount is multiplied by a snowfall correction factor **SFCF** [-]. As soon as the temperature rises above TT again, snowmelt starts. For the calculation of the amount of snow melted, the degree-Δt-factor **CFMAX** [mm $°C^{-1}$ h^{-1}] is multiplied with the temperature difference between the measured temperature and TT. The meltwater is retained in the snowpack until the amount of meltwater stored in the snowpack gets larger than the water holding capacity, defined as a certain portion **CWH** [-] of the snow water equivalent. If there is meltwater stored in the snowpack when temperatures fall below TT again, this meltwater refreezes again. This process is modelled as a product of the difference between TT and the temperature, CFMAX and a refreezing coefficient **CFR** [-]. Furthermore, the parameter **SP** [-] can be used to lower CFMAX in winter.

4.5.1.2 Soil moisture routine

In the soil moisture routine, the precipitation falling as rain, the snowmelt calculated in the snow routine, and the potential evaporation serve as input data based on which the soil moisture, the groundwater recharge and the actual evaporation are calculated. The calculations of the soil moisture routine are done individually for each elevation zone.

The water entering the soil from rain or snowmelt is distributed into groundwater recharge and additional water for the soil box. The proportion of the entering water that ends up as groundwater flux is depending on the ratio between the amount of water already in the soil box and the maximal water content the soil box can hold **FC** [mm], as well as on the shaping parameter **β** [-], which is used as an exponent. The remainder of the entering water is stored in the soil box.

The soil box is exposed to evaporation. To obtain the potential evaporation at a given time step, the long-term mean values of potential evaporation are corrected by the deviation of the temperature at this time step from the long-term mean and a correction factor **CET** [$°C^{-1}$]. If the soil moisture is higher than a certain proportion **LP** [-] of FC, the actual evaporation is assumed to be equal to the potential evaporation. If the soil moisture is lower than this, the actual evaporation is linearly reduced and thus depends on how dry the soil is.

4.5.1.3 Response function

In the response function, the groundwater levels in an upper and a lower groundwater box as well as the resulting discharge are calculated under the use of the potential evaporation and the groundwater recharge coming from the soil moisture routine.

For the lower groundwater box, a simple linear storage model is assumed. In each time step, a certain portion **K2** [h^{-1}] of the amount of water stored in the lower groundwater box contributes to discharge. In the upper groundwater box, the outflow is non-linear: Depending on which of the following water amounts is smaller, either the whole content of the groundwater box contributes to discharge or only a portion, calculated as a product of **K1** [h^{-1}] and the water content to the power of (1+**α** [-]).

The percolation of water from the upper to the lower groundwater box is regulated by **PERC** [mm/h], whereby this parameter gives the maximum value of percolation that is possible. The discharge outflowing from the lower groundwater box can never exceed PERC and the maximum storage content of the lower groundwater box is given by the ratio of PERC and K2.

4.5.1.4 Routing routine

In the routing routine, the delay between the generation of discharge in the catchment and the water reaching the catchment outlet is considered. To model this delay, the discharge generated in one time step is distributed over the following few time steps, whereby the last time step affected is determined by **MAXBAS** [h]. The generated discharge of time step 1 is distributed to time steps 1 to MAXBAS following an equilateral triangular weighing function.

4.6 Model settings

4.6.1 Catchment settings

As the discharge and water level class data were available at one point per catchment, the catchments were not partitioned into subcatchments. One can expect that the parameter variability would not be too large anyways between different subcatchments, as it is the case for most applications of the HBV model (Bergström, 1992).

The catchments were partitioned in several elevation zones, each covering a band of 100 to 200 meters in altitude. To determine the elevation zones, the elevation data provided by the *Hydrological Atlas of Switzerland* was used for the Swiss catchments. For the Austrian catchments, the elevation zones were determined from the EU-DEM provided by the *Copernicus Land Monitoring Service at the European Environment Agency*. The elevation zones were kept regular wherever possible. Especially the upper and lower elevation zones sometimes covered larger elevation ranges, because very small elevation zones were merged and there is a limit of 20 elevation zones in HBV. The elevation zones used for all study catchments can be found in appendix 10.5. The measurement elevation for temperature and precipitation was set to the mean elevation of each catchment since gridded data was used to determine these timeseries (see section 4.2).

The catchments were not partitioned into vegetation zones. Lake properties were neglected since the only lake contained in one of the study catchments made up for only 3% of the catchment area (Sihlsee in the Sihl catchment). Thus, all study catchments were treated as if they did not contain any lakes.

4.6.2 Calibration, validation, and warm-up period

The choice of the hydrological years (duration in Switzerland from 1 October to 30 September) used for the calibration period was made according to the availability of citizen science data (see 4.4) at the time of writing this thesis. Thus, the calibration period covered the four years from 1 October 2017 to 30 September 2021. The independent validation period covered the four preceding hydrological years, namely the time from 1 October 2013 to 30 September 2017. A calibration period of four years is comparably short. However, in a case where continuous data was used for the calibration of the HBV model, a longer calibration period than two years did not improve the model performance significantly (Harlin, 1991). Thus, one can expect that the limited length of the calibration period was sufficient also in this case.

The meteorological data of the two hydrological years between 1 October 2015 and 30 September 2017 served as a warm-up period for the model. For the calibration period, this was the natural warm-up period, i.e., the period that really preceded the one used for calibration. The meteorological data of the two years have also been used as warm-up period before the validation period because the measurement time series at several temperature measurement stations started later than on 1 October 2011 and thus the calculation of the temperature time series would have been inconsistent.

It takes a while until at least 15 citizen scientists playing the CrowdWater game rate an observation made in the CrowdWater app. Thus, the modified water level class estimates resulting from the CrowdWater game were mainly available for data points in the beginning of the calibration period. To increase the percentage of quality-controlled citizen science data in the corresponding approach, the calibration period was chosen to be one year shorter, i.e., ranged from 1 October 2017 to 30 September 2020. The validation period however was the same as for the other approaches, i.e., ranged from 1 October 2013 to 30 September 2017.

4.6.3 Parameter ranges

Table 4 shows the parameter ranges and fixed parameter values used to calibrate the model in this thesis. For fourteen parameters, a range was defined, the other five parameters were fixed. The 0.6°C/100m is a common used value for the temperature change with altitude (TCALT) (e.g., Bergström, 1992) and can be measured rather easily. The change in precipitation with altitude (PCALT) was fixed at 5%/100m in accordance with the value used in the EXAR project (Kauzlaric et al., 2021). To keep the number of variable parameters low, the rather unsensitive parameters CWH and CFR were fixed at their default values. Furthermore, since no seasonal variability in CFMAX was expected, the value of SP was fixed at 0.

Table 4: Parameter ranges and fixed parameter values, respectively.

Parameter	Unit	Range/Value
TCALT	°C/100m	0.6
PCALT	%/100m	5
TT	°C	[-3; 2.5]
SFCF	-	[0.4; 1.6]
CFMAX	mm $°C^{-1}$ h^{-1}	[0.001; 0.5]
CWH	-	0.1
CFR	-	0.05
SP	-	0
FC	mm	[50; 550]
β	-	[1; 6]
CET	$°C^{-1}$	[0; 0.3]
LP	-	[0.3; 1]
K2	h^{-1}	[10^{-7}; 0.05]
K1	h^{-1}	[10^{-5}; 0.1]
α	-	[0; 1]
PERC	mm/h	[0; 0.125]
MAXBAS	h	[1; 100]

The parameter ranges for TT, SFCF, CFMAX, FC, β, CET, LP and PERC were chosen according to Seibert & Vis (2012). Thereby, the ranges were adjusted such that they fitted the hourly time step used in this thesis and rounded to a close number. The same was done for MAXBAS too, whereby the upper limit of the parameter range was set to the maximal value that is allowed for MAXBAS in HBV (100 hours).

For the parameters K1 and K2, the choice of the parameter range was also based on the choice made by Seibert & Vis (2012) as well as Etter et al. (2018). However, these ranges were slightly enlarged in order to be on the safe side and to not exclude slightly larger or smaller parameter values than would be in the range used by these authors. For the non-linearity parameter α, all values from 0 up to 1 were allowed, since no improvement was expected as soon as α+1 reaches values larger than 2.

4.6.4 Model calibration methods

For all parameters described above, parameter values were estimated during the model calibration process. The calibration of a model describes the process of adjusting parameters to reach a sufficiently high similarity between simulated and observed discharge in a catchment (Solomatine & Wagener, 2011). Here, each model calibration resulted in 100 "best" parameter sets. This number was chosen to account for the fact that there is not one optimal parameter set: In hydrological modelling, we face the concept of equifinality, describing the situation that different parameter sets lead to similarly good results (Beven, 2012). Even though those 100 parameter sets were the best that could be found as a result of the model calibration process applied, there may be other parameter sets in the parameter space that would lead to even better model performances. Therefore, the parameter sets can be called "best" in quotation marks only.

4.6.4.1 Objective functions

To judge the similarity between the input data and the simulated discharge during the calibration process and to rate the model performance, objective functions are used in hydrological modelling. In this thesis, the NPE (Non-Parametric Efficiency) introduced by Pool et al. (2018) was used whenever possible. Additionally, Spearman's Rank Correlation Coefficient (Spearman, 1904), also called Spearman rank correlation, was used for those cases where the NPE was not applicable (namely the cases in which water level class data only were used for the calibration of the model).

The NPE is an objective function developed to improve hydrological model calibrations and is a modification of the widely used KGE (Kling-Gupta-Efficiency) introduced by Gupta et al. (2009). When calculating the KGE, one assumes that the data does not contain any outliers, as well as that it is linear and normal. However, this is never the case for discharge data that is usually right skewed, since high flow events are rare (Pool et al., 2018). To avoid this discrepancy between data requirements and data properties, the NPE is based on the normalized flow-duration curve (FDC), describing how often a certain magnitude of discharge values is reached (Vogel & Fennessey, 1995), the Spearman rank correlation and the mean discharge (Pool et al., 2018). As mentioned before, the Spearman rank correlation served as an alternative objective function when the input data consisted of water level classes only. The Spearman rank correlation does not make any statement about the fit of the observed and simulated discharge volume, but only about the dynamics of the hydrograph. As there is no volume information contained in the water level class data, this objective function is an option that still allows calibration against such input data, as done by Etter et al. (2020b).

The NPE as well as the Spearman rank correlation are included as objective functions in the HBV model. The external calculations required to evaluate the model performances were done using a script available from Pool et al. (2018), or the function included in the base package of R, respectively. The calculations can be formulated as follows, with n as the length of the time series, Q_{sim} as the simulated discharge time series and Q_{obs} as the observed discharge time series:

$$\alpha = 1 - \frac{1}{2}\sum_{i=1}^{n}\left|R_{FDC,sim,i} - R_{FDC,obs,i}\right|, \text{ with } R_{FDC,x,i} = \frac{Q_{x,i}|i=|\{Q_{x,j}|Q_{x,j}<Q_{x,i}\}|+1}{\overline{Q_x}\cdot n} \quad (3)$$

$$\beta = \overline{Q_{sim}} \Big/ \overline{Q_{obs}} \quad (4)$$

For the Spearman rank correlation, the rank $R(Q_{x,i})$ with $x \in \{sim, obs\}$ of each datapoint is used to calculate the coefficient (Spearman, 1904):

$$r_S = \frac{\sum_{i=1}^{n}(R(Q_{sim,i}) - \overline{R(Q_{sim})})(R(Q_{obs,i}) - \overline{Q_{obs}})}{\sqrt{\sum_{i=1}^{n}\left(R(Q_{sim,i}) - \overline{Q_{sim}}\right)^2 \cdot \sum_{i=1}^{n}\left(R(Q_{obs,i}) - \overline{Q_{obs}}\right)^2}} \quad (5)$$

The final objective function R_{NPE} (or NPE) is calculated based on these three components:

$$R_{NPE} = 1 - \sqrt{(\alpha - 1)^2 + (\beta - 1)^2 + (r_S - 1)^2} \quad (6)$$

For a perfect match between simulated and observed discharge, the NPE as well as the Spearman rank correlation have the value 1.

To implement an estimation of the mean discharge (see section 4.8.1), the volume error was used. The volume error V_{err} is implemented as an objective function in HBV and is defined as

$$V_{err} = 1 - \frac{\left|\sum_{i=1}^{n} Q_{obs,i} - Q_{sim,i}\right|}{\sum_{i=1}^{n} Q_{obs,i}} \quad (7)$$

This means that a perfect fit between the simulated and the observed discharge results in a volume error of 1. However, for a perfect fit, there is no error in volume resulting, thus the error itself is 0. Therefore, what was used as volume error here is only the ratio in equation (7), so the deviation of the simulated discharge volume from the observed discharge volume. The closer the simulated discharge fits with the observed discharge, the closer to 0 the volume error.

4.6.4.2 Monte Carlo simulation

The model calibration method mainly used in this thesis was the Monte Carlo simulation. With a Monte Carlo simulation, the parameter space is sampled as thoroughly as possible in order to find the parameter sets that lead to good model performances. Thereby, the parameter sets are built randomly, i.e., for each parameter, a random value within its range is chosen and combined with random values of all the other parameters (Harrison, 2009). Here, one million parameter sets were sampled from the multidimensional parameter space. The resulting model performances of all these parameter sets were evaluated regarding the Spearman rank correlation or the NPE of the simulated discharge with the input data in order to find out which parameter sets performed well and should further be used.

4.6.4.3 GAP simulation

A genetic algorithm was used to calibrate the model with the full discharge time series, i.e., to find the upper benchmark (see section 4.6.5.2). The algorithm is included in the HBV model and was introduced by Seibert (2000). A genetic algorithm simulates an evolution of parameter sets. The algorithm starts with a certain number of parameter sets that are all evaluated regarding one or several objective function(s). Parameter sets leading to a good model performance are more likely to further evolve than parameter sets leading to a bad model performance. In each step, a new generation of parameter sets is built by combining the parameter sets from the preceding generation. Thereby, each parameter can either be taken from one of the "parental" parameter sets, a mutation between the two values from the parental parameter sets or a random value. Since well-performing parameter sets have higher chances to evolve, the resulting parameter sets should all lead to a rather good model performance after a sufficient number of iterations (Seibert, 2000). The NPE was used as an objective function in the GAP calibration to find the parameter sets that served as an upper benchmark.

4.6.5 Evaluation

4.6.5.1 Model validation

The calibrated model was used to simulate the discharge of each catchment during the calibration period and the model performance was evaluated using the observed discharge from this period. To test for the stability of the parameter sets, the discharge of an independent validation period was additionally simulated with the parameter sets found in the calibration process and compared to the observed discharge of this period (see 4.6.2). On one hand, the purpose of that was to see if the model parameters are valuable for forecasting approaches or other extensions of discharge time series. On the other hand, overparameterization can be identified if the model performance is much lower for the validation period compared to the model performance in the period used for calibration (Bergström, 1991).

4.6.5.2 Upper and lower benchmark

From the value of an objective function alone, it is not possible to make a statement about the value of the data used to calibrate the model. For such a statement, the model performance reached should be compared to the maximal expectations (upper benchmark) and the minimal expectations (lower benchmark) of the model performance in a catchment. Using benchmarks for comparison avoids the issue that the model performances may differ strongly between different catchments. Thereby, the benchmark is a simulated time series obtained by using a different approach (Schaefli & Gupta, 2007; Seibert, 2001; Seibert et al., 2018).

An upper and a lower benchmark were used to assess the model performances that resulted from the different scenarios (see section 4.7) in each catchment. In many common objective functions, the observed discharge implicitly serves as an upper benchmark. Since the focus here was on the value of the data compared to an optimal data availability, the upper benchmark was chosen to be the simulation resulting from the calibration with the full time series of discharge data. Therefore, for the upper benchmark, 100 GAP calibrations with 100 parameter sets and 5000 runs each were done in order to optimize the value of the NPE. For the resulting 100 parameter sets, the corresponding simulated hydrographs were calculated. These 100 hydrographs were then weighed equally to form an ensemble mean hydrograph of which the NPE value served as the upper benchmark.

To find the lower benchmark, 1000 parameter sets were chosen randomly from the parameter space, and their corresponding hydrographs were calculated. Analogous to the calculations for the upper benchmark, the ensemble mean hydrograph was then calculated using equal weights. Since there was

nothing known about the quality of any of these hydrographs, all the 1000 hydrographs were used for the ensemble mean of which the NPE served as a lower benchmark then.

The relative performance $R_{NPE,x,rel}$ of a scenario x was then set in relation to the benchmarks:

$$R_{NPE,x,rel} = \frac{R_{NPE,x} - R_{NPE,l.b.}}{R_{NPE,u.b.} - R_{NPE,l.b.}} \quad (8)$$

Note that the span between the performances of the upper and the lower benchmark thus had a large influence on the relative model performance as it defined the space available for improvement when using some data instead of no data at all or the complete measurement time series.

The upper and the lower benchmark were calculated twice for each study catchment: Once for the calibration period and once for the validation period. This accounted for the fact that the two periods may vary in difficulty to model. The separate benchmarks assure that the model performance in the two different periods can be assessed in accordance with the properties of each period.

4.7 Definition of scenarios

To determine the value of having different amounts of data available, 24 scenarios of data availability were defined by combining a certain amount of citizen science data with a certain number of discharge measurements per year (Table 5).

Regarding the citizen science data, five options were taken into account: Having no citizen science data at all and having 25%, 50%, 75% and 100% of the observations done by citizen scientists available (whereby the absolute number of observations differed among the catchments). The number of observations was not standardized to the same number in each catchment since the scenarios were meant to depict reality: It is not possible to motivate the same number of people to make the same number of observations at each observation spot. Citizen scientists collect data in their own rhythm, and it is part of a citizen science project to make use of these irregular data, even though this means a loss of comparability among different sites. For the scenarios using 25%, 50% and 75% of the citizen science data, the data points were chosen randomly out of the set of available data points. These scenarios represented the situations in which only a portion of the observations would have been received. For the scenarios using 100% of the citizen science data, the complete set of data points received from citizen scientists (see Table 2) was used.

These five options were combined with five options containing different numbers of discharge measurements per hydrological year, as listed below. Aside the situation with no discharge measurements at all, these options represented situations in which a responsible person (potentially a citizen scientist) would have conducted discharge measurements at regular time steps:

- 1 discharge measurement per year: 21 April, 12:00
- 3 discharge measurements per year: 21 April / 21 August / 21 December, 12:00
- 6 discharge measurements per year: 21st of each even month, 12:00
- 12 discharge measurements per year: 21st of each month, 12:00

The date for a single measurement per hydrological year, i.e., 21 April, was chosen randomly. Note that this date is about in the middle of the hydrological year. In the snow-dominated catchments, the discharge values are about to increase around this time of the year due to the snowmelt season in

spring (Figure 7). The additional discharge measurements were spread equally over the year. If a discharge measurement was used in a set of fewer discharge measurements, it was always also used in each set with more discharge measurements. In other words, to get more discharge measurements, new data points were added but none of the already available data points got replaced. The discharge measurements were chosen in regular intervals, even though this strategy may be less informative than an intelligent sampling, e.g., during high flow conditions (Seibert & Beven, 2009). Thus, it was assumed that nothing was known about the discharge behaviour. This way of sampling has the advantages that it can be planned easily and logistic difficulties to access the catchment during certain hydrological conditions can be avoided (Jian et al., 2017).

To simulate the availability of the required discharge measurements, the corresponding data points were extracted from the complete discharge time series (see section 4.3), while all the other data points of the discharge time series were assumed to be unavailable.

Table 5: Overview of the modelling scenarios created by combining a certain number of discharge measurements per year with a certain amount of citizen science data. The first number in the scenario name corresponds to the number of discharge measurements per year, the second to the percentage of citizen science data used.

	Number of discharge measurements increasing →				
Amount of citizen science data increasing ↑	0-100	1-100	3-100	6-100	12-100
	0-75	1-75	3-75	6-75	12-75
	0-50	1-50	3-50	6-50	12-50
	0-25	1-25	3-25	6-25	12-25
		1-0	3-0	6-0	12-0

In the lower left corner of Table 5, the situation without any data, i.e., the situation of the lower benchmark is shown. By definition of the relative model performance (see formula (8)), the relative performance assigned to that field in the results section was always 0 (see section 5).

4.7.1 Model calibration per scenario

For each scenario using one type of data (i.e., either water level classes or discharge measurements but not both), one million parameter sets that were randomly sampled from the parameter space were evaluated. For all scenarios and all catchments, the same set of one million parameter sets was used to do so. For the scenarios with discharge measurements only, the NPE was used as objective function. For the scenarios with citizen science data only, the Spearman rank correlation was used as objective function. For each scenario, the top 100 parameter sets showing the best results regarding the corresponding objective function were chosen as the "best" 100 parameter sets for this scenario and this catchment. Note that the objective function could only be calculated using the available data points of the input data.

For the inner scenarios, i.e., the scenarios using citizen science data and discharge measurements, no direct evaluation of the one million parameter sets was possible. The water level class data obtained

by citizen scientists were of ordinal scale level and did not have any unit while the discharge measurements were interval-scaled and given in mm/h. It was not possible to evaluate the model on a combination of these two data types simultaneously. Thus, the separate evaluations done for the scenarios with only one data type were combined by the following procedure:

1. Each of the one million parameter sets were ranked from 1 to 10^6 according to their performance in the scenarios using only one of the two data types (thus, eight times per catchment).
2. For each of the 16 combinations of citizen science data and discharge measurements, the mean rank for each parameter set was calculated.
3. The "best" 100 parameter sets regarding the mean rank for each combination and thus for each scenario with combined data types were chosen.

For all scenarios and all catchments, the ensemble mean hydrograph resulting from the 100 hydrographs simulated by the "best" parameter sets was calculated. The absolute NPE obtained like this was converted to a relative model performance using formula (8). This value was then used to judge the model performance of each scenario in each catchment and thus to judge the value of the data that was used to calibrate the model.

For the scenarios using one data type only, it would have been possible to use the GAP algorithm for model calibration. This was done in the beginning of the investigations and resulted in slightly higher model performances than it was the case when the model was calibrated as described above. However, in order to treat all scenarios as similar as possible, this option was neglected, and all scenarios were investigated based on the Monte Carlo approach. The application of the GAP algorithm was not possible for the scenarios with mixed input data as the model could not process the information of both input data types simultaneously and make use of them in the algorithm.

4.8 Implementation of additional knowledge

4.8.1 Mean discharge

In the second approach, it was assumed that an estimate of the mean discharge was available in addition to the discharge measurements and the citizen science data used in the basic approach. To simulate this situation, the one million parameter sets were filtered according to their resulting volume error before the 100 "best" parameter sets for each scenario were chosen. The narrowest filter (simulating a very precise estimate of the mean annual discharge) allowed only parameter sets resulting in a volume error smaller than 2.5%. Furthermore, filters allowing volume errors of 5%, 10%, 20%, 30% and 50% were applied. From this smaller number of parameter sets, 100 "best" parameter sets were chosen for each scenario in the same way as it was done in the basic approach using the original one million parameter sets (see section 4.7.1).

4.8.2 Water levels instead of water level classes

In the third approach, it was assumed that instead of water level classes, water level measurements were available. These data were assumed to not contain any errors and to have a resolution as high as the resolution of the discharge measurements obtained by the authorities. To simulate this situation, each water level class data point was replaced with the amount of discharge measured by the authorities at the time of the observation. Since these data should still represent water levels, none of the volume information contained in this value was used to calibrate the model: The ranking of the one million parameter sets was still done according to the resulting Spearman rank correlation when these data points were considered. The ranking of the parameter sets regarding the discharge measurements remained unchanged compared to the basic approach.

4.8.3 Water level classes checked by citizen scientists

To explore if the value of the citizen science data was increased in the CrowdWater game, the model performances resulting from the calibration with checked and the calibration with unchecked data were compared directly. Since the calibration period was different for this approach (see section 4.6.2), the basic approach did not serve as the case using the unchecked data. Instead, an alternative basic approach was designed separately together with the approach using data that already run through the CrowdWater game (i.e., the checked data).

A data point was classified as checked when 15 or more players in the CrowdWater game voted on the water level class in the picture. From all votes, the trimmed mean (cutting off the highest and lowest 10% of voted values) was calculated and served as a checked value of the data point. Compared to the original value observed by a single citizen scientist, this checked value should be more accurate and show a higher resolution than the original, unchecked value (Strobl et al., 2019).

As in the other approaches, the amount of citizen science data was increased in steps of 25%. For the approach using checked data, checked data only was used if possible and only when no more checked data points were available, unchecked data was used to complete the data set (cf. section 4.4.2). For the case using unchecked data, the same data points as in the checked case were used, but instead of the trimmed mean of the game votes, the original value observed by the citizen scientist was used. The resulting sets of (partially) checked and unchecked data points were used to rank the one million parameter sets as described for the water level classes in the basic approach. Again, the use of discharge measurements for the ranking of the one million parameter sets remained the same as in the basic approach.

4.9 Data analysis

All data analyses were done using R. Table 6 gives an overview of the most important packages (in alphabetic order) that were used in addition to the base package of R. Additionally, the source and main purpose of the packages is given.

Table 6: Most important R packages with source and description of their purpose.

Package	Source	Purpose
ComplexHeatmap	Gu et al. (2016)	Drawing of all heatmaps shown in the results
dplyr (tidyverse)	Wickham et al. (2019)	Handling of data frames
hydroGOF	Zambrano-Bigiarini (2020)	Calculation of hydrological goodness of fit functions
Lubridate (tidyverse)	Wickham et al. (2019)	Handling of dates and time stamps
RColorBrewer	Neuwirth (2014)	Choice of colours for most visualizations
readxl (tidyverse)	Wickham et al. (2019)	Import of excel files
viridis	Garnier et al. (2021)	Choice of colours for main heatmap

5 Results

5.1 Benchmarks

The upper benchmarks in both periods had an NPE of 0.89 on average, while the NPE values of the lower benchmarks were distributed between 0.34 and 0.75 (Figure 16). As described before (see section 4.6.5.2), the benchmarks were calculated separately for the two periods. A high upper benchmark means that the hydrograph could be modelled well when the model was calibrated with the full discharge time series. A high lower benchmark means that the hydrograph could be modelled well without any knowledge about the discharge behaviour.

In addition to the upper benchmark calibrated using the GAP approach, the results of the Monte Carlo calibration with the full discharge time series (the resulting model performance if the Monte Carlo approach would have been used to calibrate the upper benchmark) are also given in Figure 16. These values indicate the model performance of the ensemble mean, built by the 100 parameter sets out of the one million parameter sets that led to the best NPE values when compared to the full discharge time series. These values only serve as a comparison. They were not used for any calculations. The upper benchmark calibrated using the GAP approach was usually higher than the upper benchmark calibrated using the Monte Carlo approach. This demonstrates the strength of the GAP approach in finding better parameter sets in less time.

The values of the lower benchmark and the upper benchmark calibrated using the GAP approach (yellow and blue dots in Figure 16) are additionally given in Table 7. The two upper and two lower benchmarks for the calibration and the validation period differed quite strongly for some of the catchments. These comparably large differences are indicated in orange in Table 7. For the lower benchmark, the differences may partly be explained by some coincidence in the parameter sets used to calculate the ensemble mean. For the upper benchmarks of the Urtene and the Ova dal Fuorn however, the rather large differences may indicate that the calibration and the validation period were not equally easy to model, i.e., the validation period was more challenging to model than the calibration period. The upper benchmarks resulting from the Monte Carlo approach in these two catchments indicate the same.

Table 7: Upper and lower benchmarks for all catchments, for the calibration period and the validation period. Comparably large differences between the calibration and the validation period are highlighted in orange.

Catchment	Lower benchmark, cal. p.	Lower benchmark, val. p.	Upper benchmark, cal. p.	Upper benchmark, val. p.
Koenigsseeache, Niederalm	0.698	0.749	0.902	0.924
Salzach, Salzburg	0.637	0.619	0.870	0.890
Kempt, Fehraltdorf	0.648	0.659	0.921	0.932
Urtene, Kernenried	0.553	0.515	0.888	0.768
Alp, Einsiedeln	0.609	0.632	0.878	0.901
Kleine Emme, Werthenstein	0.649	0.671	0.905	0.923
Ova dal Fuorn, Zernez	0.534	0.339	0.947	0.882
Kleine Emme, Emmen	0.639	0.686	0.909	0.884
Wigger, Zofingen	0.720	0.746	0.908	0.929
Sellenbodenbach, Neuenkirch	0.430	0.510	0.841	0.878
Sihl, Zurich	0.715	0.639	0.859	0.842

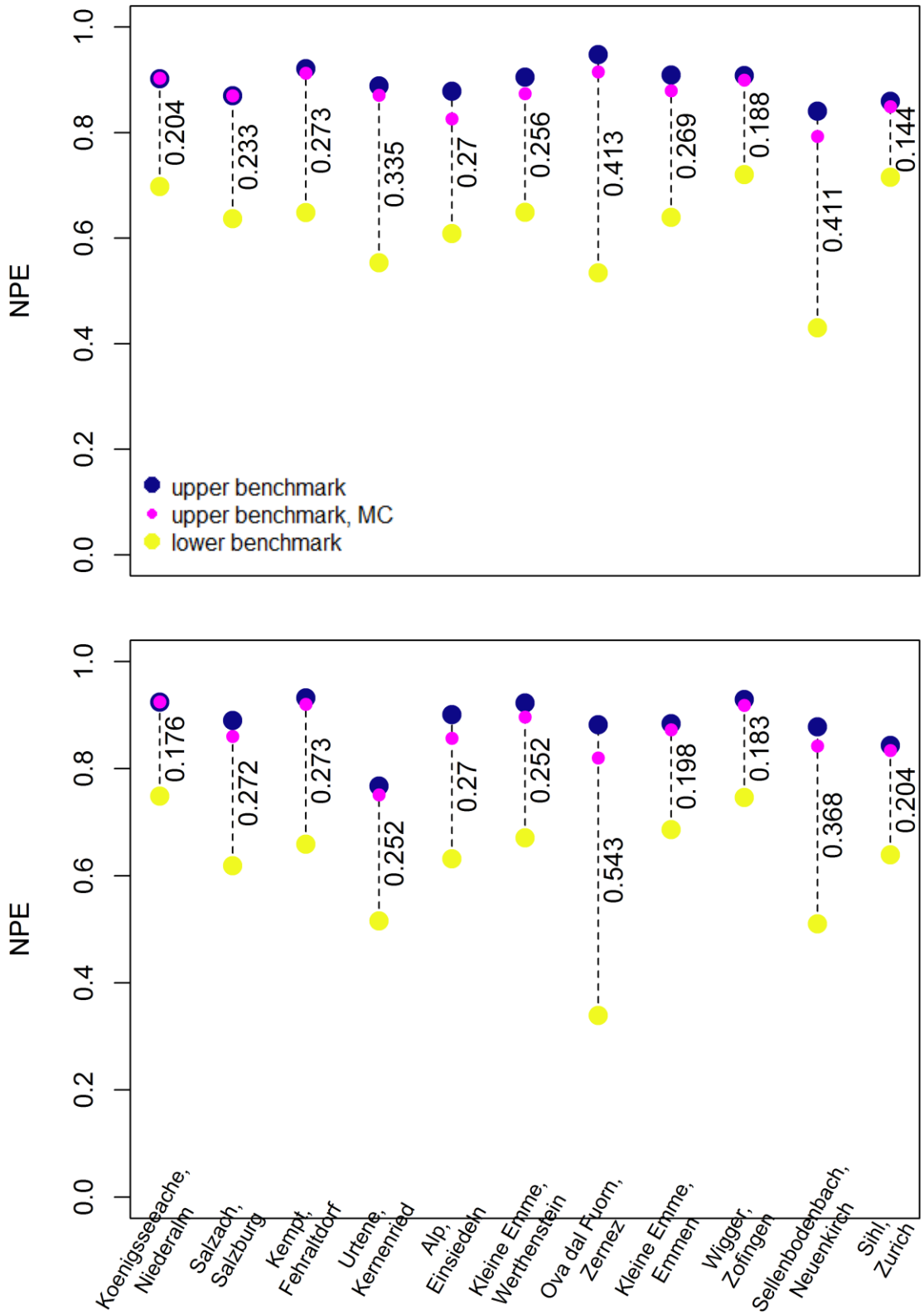


Figure 16: Benchmarks for the calibration period (upper plot) and the validation period (lower plot). The number describes the difference between the upper and lower benchmark. The upper benchmarks calibrated using the Monte Carlo approach were not used for the calculation of the relative performance and only serve as a comparison. Note that the upper benchmark for the validation period is rather low for the Urtene. Note the large difference between the upper and the lower benchmark at the Ova dal Fuorn, especially for the validation period.

5.2 Basic approach

In the basic approach, citizen science data were combined with a limited number of discharge measurements in each hydrological year. No additional knowledge was used to calibrate the model. The resulting relative performances regarding the NPE of the ensemble mean varied strongly among the different catchments and different scenarios (Figure 17 for the calibration period, Figure 18 for the validation period). A model performance resulting in the same or a worse NPE as was reached by the lower benchmark is indicated in yellow. Darker fields indicate better performances. A model performance resulting in the same or a better NPE as was reached by the upper benchmark is indicated in dark purple.

Rather poor model performances were reached if citizen science data only was used for calibration (i.e., in the first column) in most catchments, whereby the Koenigsseeache in Niederalm was an exception with good resulting model performances in the first column. In most catchments, a clear improvement of the model performance was achieved as soon as at least one discharge measurement per year was added to the input data (i.e., from the second column and onwards) and thus the choice of the parameter sets was no longer only depending on a good Spearman rank correlation between the simulated discharge and the observed water level classes. In several catchments (e.g., at the Salzach in Salzburg, the Alp in Einsiedeln and the Ova dal Fuorn in Zernez), another clear improvement of the model performances was reached between one and more than one discharge measurements per year (i.e., between the second and the third column). The best model performances were usually reached when no citizen science data at all but only discharge measurements were used to calibrate the model (i.e., in the bottom row). Thus, the best results could be reached when only the NPE of the simulated discharge and the discharge measurements used as input data was considered to choose the 100 parameter sets for the ensemble mean. In other words, the model tended to perform better if the Spearman rank correlation between the simulated discharge and the observed water level classes was not considered. This trend was most pronounced in catchments in which only a comparably small number of water level class observations which was of a rather low quality was available. The model performance thereby increased when the number of discharge measurements used for calibration per year was increased (when moving from left to right in the bottom row of each subplot).

The expected trend of better performances when using more data of any type could only be observed among the scenarios using both data types for calibration (mixed scenarios), and not even for all catchments. The better model performances that resulted when using more data of any type could be observed especially in catchments with a lot of citizen science data. These catchments were also the catchments in which the citizen science data showed a high correlation with the discharge time series (Table 2). Furthermore, the trend was more pronounced in the validation period than in the calibration period, meaning that more input data helped to find more stable parameter sets. However, also for these mixed scenarios, the discharge measurements tended to have a higher value for the calibration of the model than the citizen science data, as there was a tendency that the model performance increase was stronger when going from left to right than when going from the bottom to the top of the subplots. At the Kleine Emme in Emmen and at the Sellenbodenbach in Neuenkirch, adding more citizen science data while using three or more discharge measurements per year resulted in a decrease in model performance, thus the citizen science data acted disinformatively in these catchments. Note that these two catchments were among the catchments that showed a poor correlation between the water level class observations and the discharge time series (Table 2).

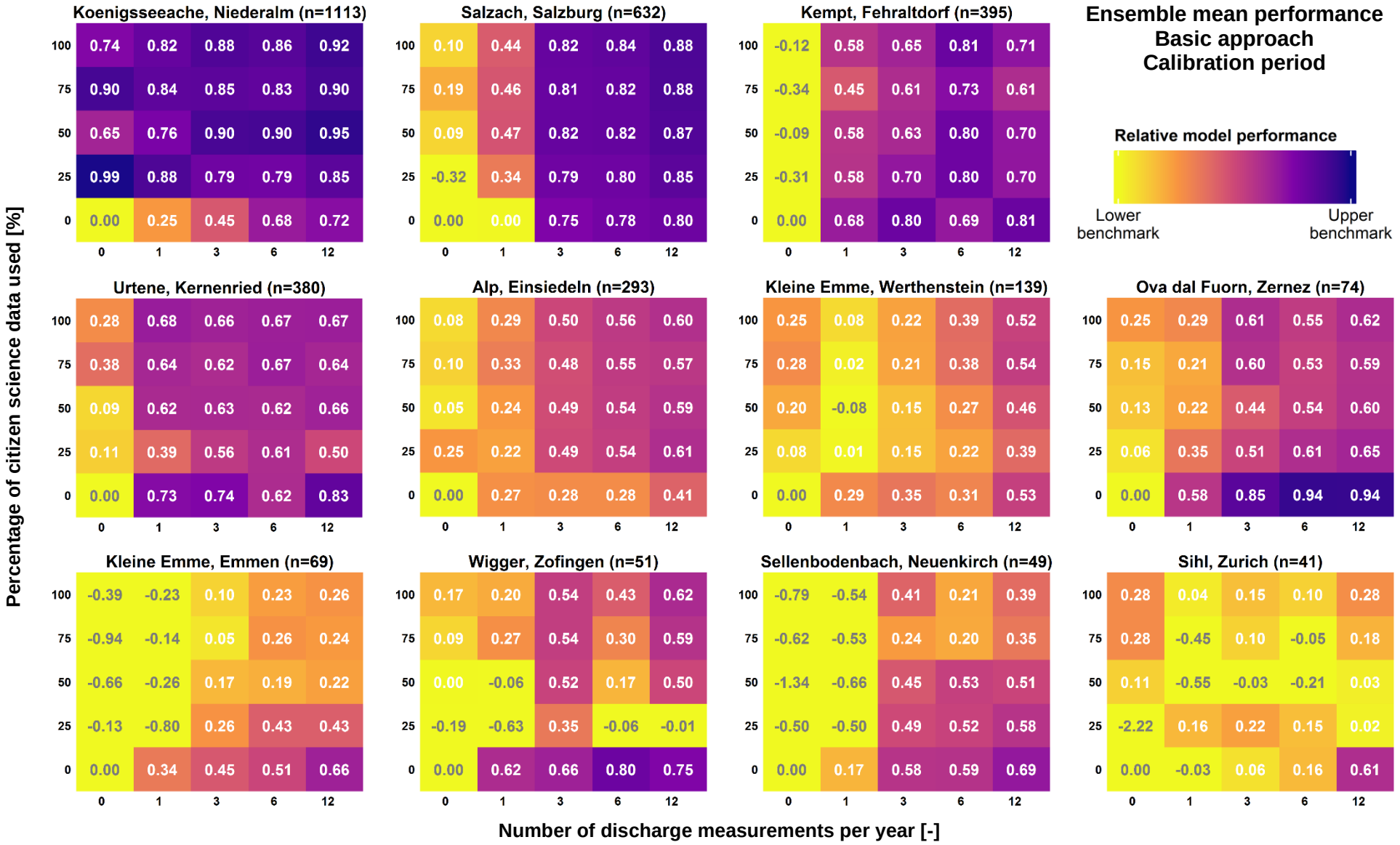


Figure 17: Relative performance of the ensemble mean for all catchments and all scenarios. Use of basic approach, results for the calibration period. The number of citizen science observations corresponding to 100% is given as n after the name of the catchment.

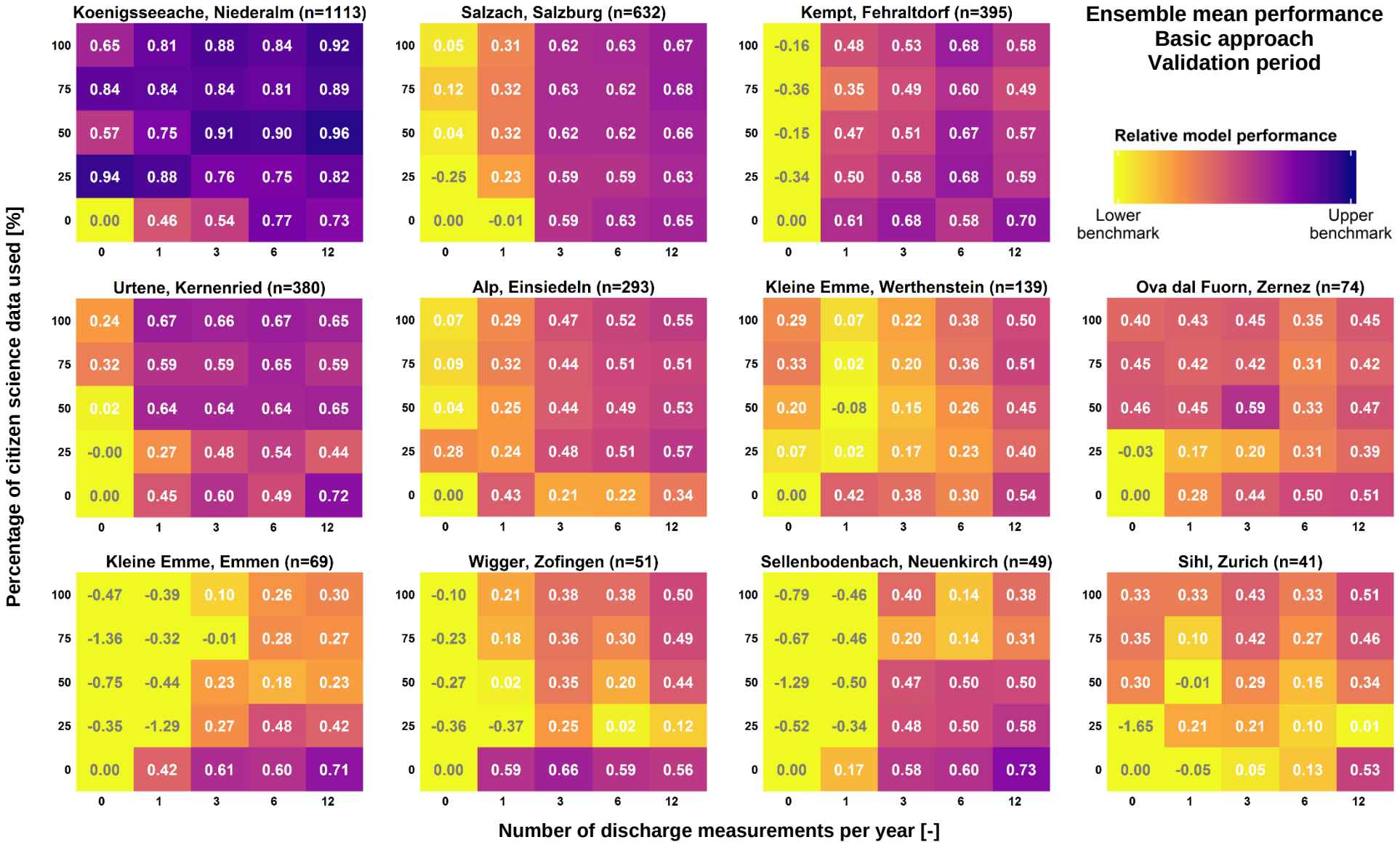


Figure 18: Relative performance of the ensemble mean for all catchments and all scenarios. Use of basic approach, results for the validation period. The number of citizen science observations corresponding to 100% is given as n after the name of the catchment.

The exemplary hydrographs for the hydrological year 2020 (Figure 19) point out more explicitly what the numbers in Figure 17 imply: The observed discharge at the Koenigsseeache (top row), at the Alp (middle row) and at the Sihl (bottom row) are shown in blue and shaded with the 100 hydrographs that were chosen to form the ensemble mean based on the data availability scenario indicated in the title of each subplot. The relative performance given in brackets is the relative performance of the ensemble mean over the whole calibration period (as stated in Figure 17) and not only over the hydrological year 2020. For each of the catchments shown, one of the worse performing scenarios is plotted on the left side and one of the better performing scenarios is plotted on the right side.

At the Koenigsseeache, the simulated hydrographs resulting from the comparably bad-performing scenario 1-0 matched the observed discharge well. Peak flows were simulated as peak flows, even if

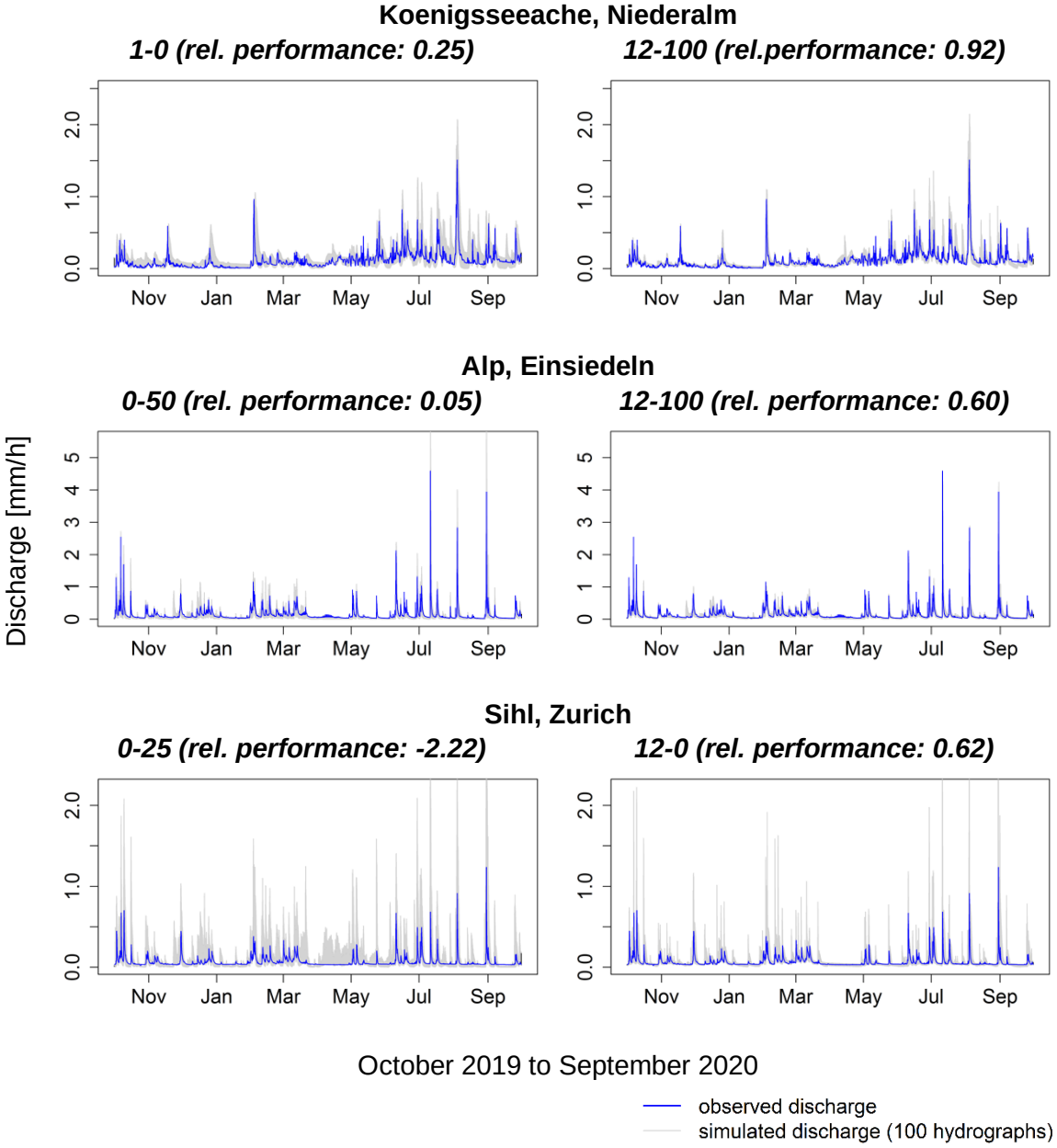


Figure 19: Observed and simulated hydrographs of the Koenigsseeache, the Alp and the Sihl for the hydrological year 2020 for one comparably bad-performing scenario (left) and one comparably well-performing scenario (right). Observed hydrograph shown in blue, 100 simulated hydrographs used to calculate the ensemble mean shown in grey.

the height of the peak was usually overestimated by some of the simulated hydrographs. Low flows were simulated as low flows, even though there was some overestimation of the discharge too (especially in winter). Partially regarding the peaks, but especially regarding the low flows, the simulated hydrographs resulting from the scenario 12-100 matched the observed hydrograph even better. At the Alp, scenario 0-50 led to a strong overestimation of the peak flows and quite some underestimation of the discharge during low flow periods. Both errors could strongly be decreased when 12 discharge measurements and all available citizen science data were used for calibration (i.e., in scenario 12-100). At the Sihl, scenario 0-25 performed worse than the lower benchmark, i.e., random guessing led to better results than calibrating the model using 25% of the citizen science data available at this site. Thus, the simulated hydrographs did not match the observed hydrograph well in this scenario. Peak flows were overestimated, and low flows were partially simulated as high flows (especially the low flow period in April 2020). Quite some improvement was reached in the best-performing scenario 12-0 at the Sihl. While peak flows were still overestimated, the model performed much better during low flow periods which were simulated as low flow periods by all 100 simulated hydrographs.

When comparing the results of the calibration period to those of the validation period, the patterns of the relative performances (Figure 17 and Figure 18) mainly remained the same, whereby the relative performance values were a bit lower in the validation period, in general. As an exemplary comparison of two hydrographs originating from the calibration and the validation period, the hydrographs of the Kempt for the hydrological years 2020 and 2016 are shown in Figure 20. In addition to the observed hydrographs shown in blue, the 100 hydrographs used for the ensemble mean of scenario 6-100 are again shown in grey.

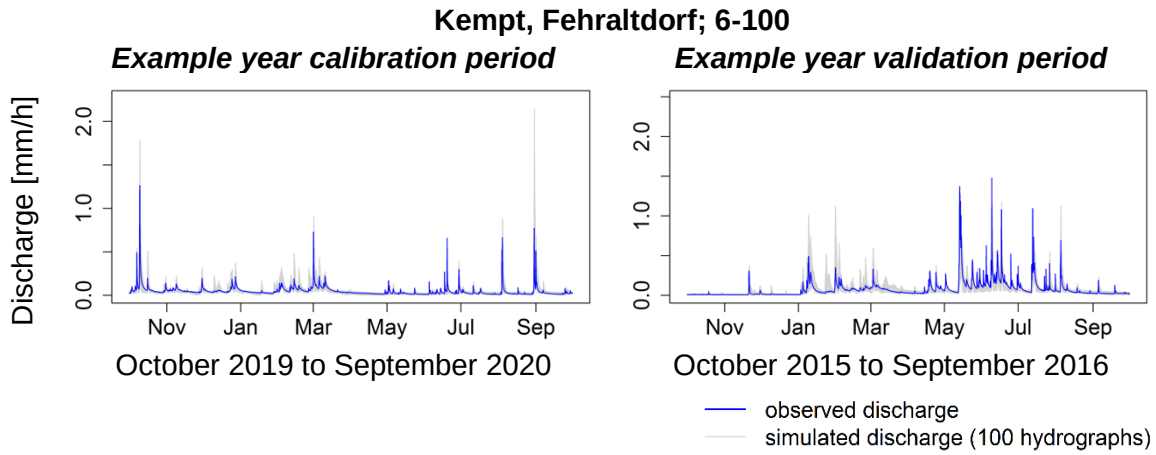


Figure 20: Observed and simulated hydrographs of the Kempt for the well-performing scenario (6-100) for one year of the calibration period (2020) and one year of the validation period (2016). Observed hydrograph shown in blue, 100 simulated hydrographs building the ensemble mean shown in grey.

For all catchments except the Ova dal Fuorn and the Urtene, the absolute values of the NPE did hardly differ between the calibration and the validation period (Figure 21). The parameter sets seem to be stable and there seems to be no fine-tuning to the input values used for model calibration. Note that for the Ova dal Fuorn and the Urtene, the validation period was more difficult to model with little input information than it was the case for the calibration period: The lower benchmark of these catchments showed low performance values during the validation period. However, at the Urtene also the upper benchmark showed a lower performance and thus the relative performances are not too different. At the Ova dal Fuorn, the upper benchmark remained high also for the validation period. Therefore, considering the relative performances shown in Figure 17 and Figure 18, there was even an improvement for those data availability scenarios of the Ova dal Fuorn that only show a small drop in absolute performance when moving from the calibration to the validation period.

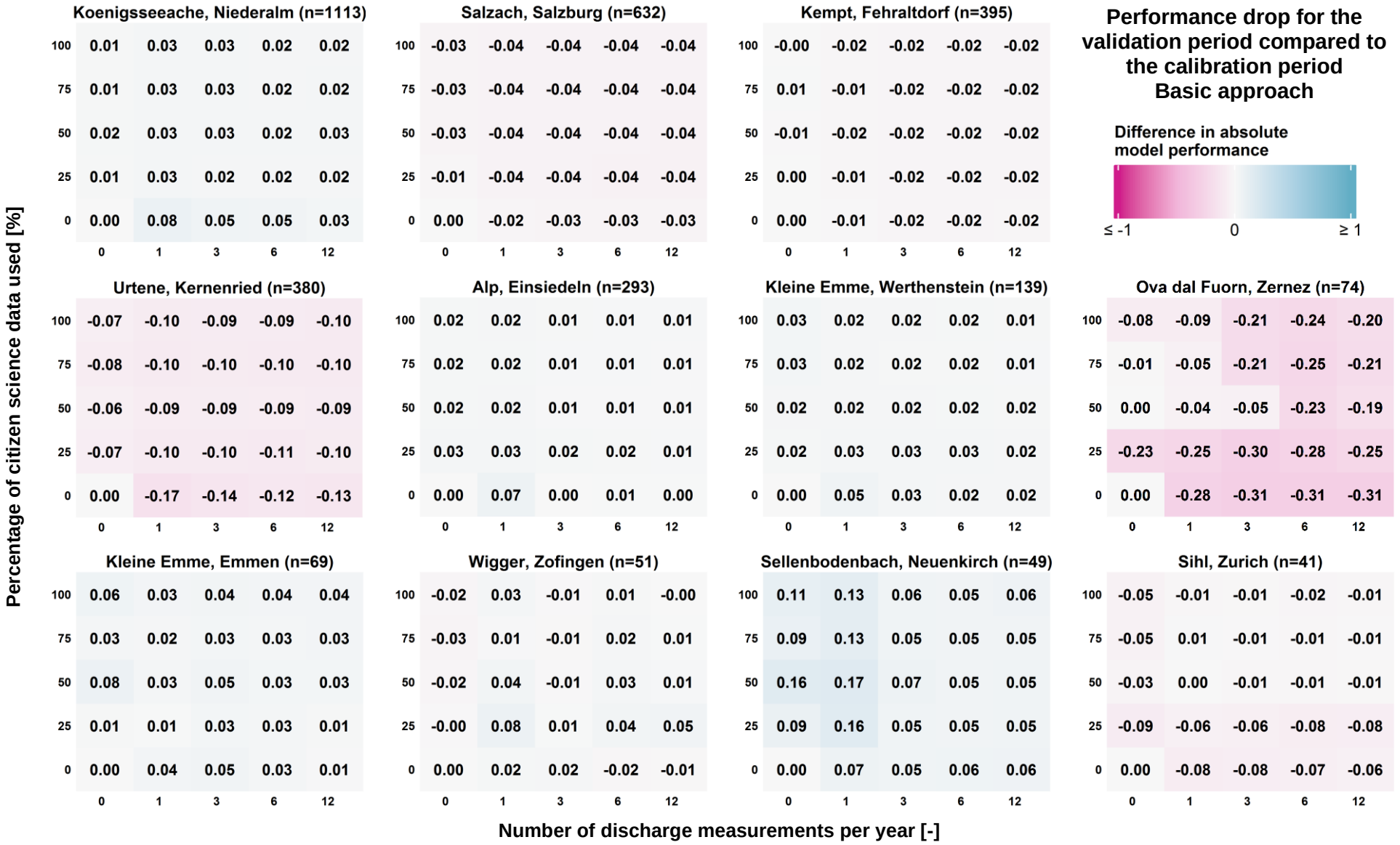


Figure 21: Difference in absolute NPE performance of the ensemble mean for all catchments and all scenarios between the validation period and the calibration period. A performance drop in the validation period is indicated with negative values, an increase of performance in the validation period is indicated with positive values.

For the calibration period, the relative model performances regarding the components of the NPE (see section 4.6.4.1) are additionally shown in Figure 22 (relative performance regarding α), Figure 23 (relative performance regarding β) and Figure 24 (relative performance regarding r_S). In the case of α and r_S, the relative value compared to the benchmarks is shown directly. In the case of β, the relative values show the relative deviation from 1, whereby the deviation of the lower benchmark corresponds to a value of 0 and the deviation of the upper benchmark corresponds to a value of 1.

There was no clear trend for the goodness of the simulation of the flow duration curve (Figure 22). For most catchments, there seemed to be a tendency for rather bad simulations of the flow duration curve if only citizen science data was used for the calibration of the model. However, this was not true for the Koenigsseeache, the Kleine Emme in Emmen as well as for the Sellenbodenbach. An interesting observation was the surprisingly bad performance regarding α at the Salzach and vice-versa the surprisingly good performance regarding α at the Sihl. Regarding α, a clear break can be observed in the Alp catchment, as scenarios in which at least 25% of the available citizen science data and at least 3 discharge measurements per year are used for calibration were able to simulate the flow duration curve much better than the remaining scenarios. This break could not be compensated by β or r_S and thus remained visible in the relative NPE performance, especially for the calibration period (Figure 17).

The performance patterns were less patchy regarding the performances of the resulting ensemble means of the different scenarios in simulating the mean discharge (Figure 23). In general, the missing volume information in the scenarios using citizen science data was reflected in a rather low performance of these scenarios regarding β. Other than that, the performances regarding the discharge volume did not differ too strongly among the different scenarios, however more discharge measurements per hydrological year (and thus more information about the discharge volume) tended to lead to better simulations of the mean discharge.

The influence of the quality of the water level class observations was clearly reflected in the relative performances of the resulting ensemble means regarding the Spearman rank correlation (Figure 24). At the Kleine Emme in Emmen, the Sellenbodenbach and the Sihl, the resulting r_S-values were poorer than those of the lower benchmark, at least if citizen science data (showing a low correlation with the discharge time series) was used to calibrate the model. On the other hand, the water level class data collected at the Koenigsseeache, Salzach, Kempt and Alp (showing a high correlation with the discharge time series) led to similarly high or higher r_S-values than this was the case for the upper benchmark. Thus, the use of good water level class observations to calibrate the model led to a good simulation of the discharge dynamics in a stream.

While the quality of the citizen science data seemed to have a large influence on the resulting model performance regarding the Spearman rank correlation, the amount of citizen science data seemed to have a smaller influence: The increase in performance when moving from 25% of citizen science data to 100% of citizen science data was not very large. A bit of a larger difference between the rows showing the results for 25% of citizen science data and 50% of citizen science data could be observed at the Urtene. However, the performance increase with more citizen science data could not be observed anymore when comparing the 50%-scenarios to the scenarios using 75% or 100% of the available citizen science data. This finding could be confirmed when comparing different catchments with each other: About the same number of citizen science data (though not necessarily distributed similarly over the calibration period) was used for the 25%-scenarios at the Alp as for the 100%-scenarios at the Kleine Emme in Emmen. Still, the corresponding performances regarding the Spearman rank correlation at the Alp were way better than those at the Kleine Emme in Emmen.

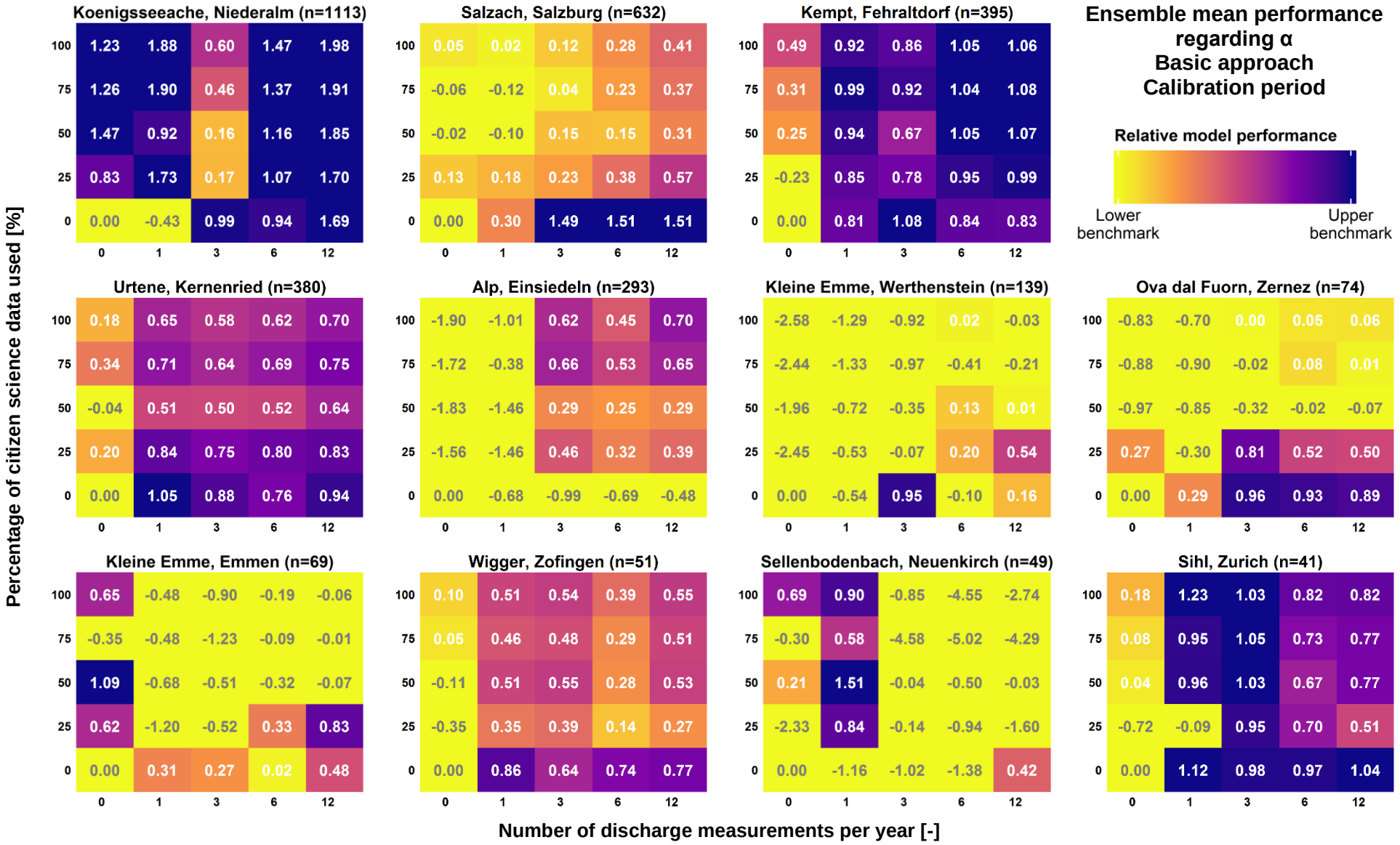


Figure 22: Relative performance of the ensemble mean for all catchments and all scenarios, considering only α (first component of the NPE). Use of basic approach, results for the calibration period. The number of citizen science observations corresponding to 100% is given as n after the name of the catchment.

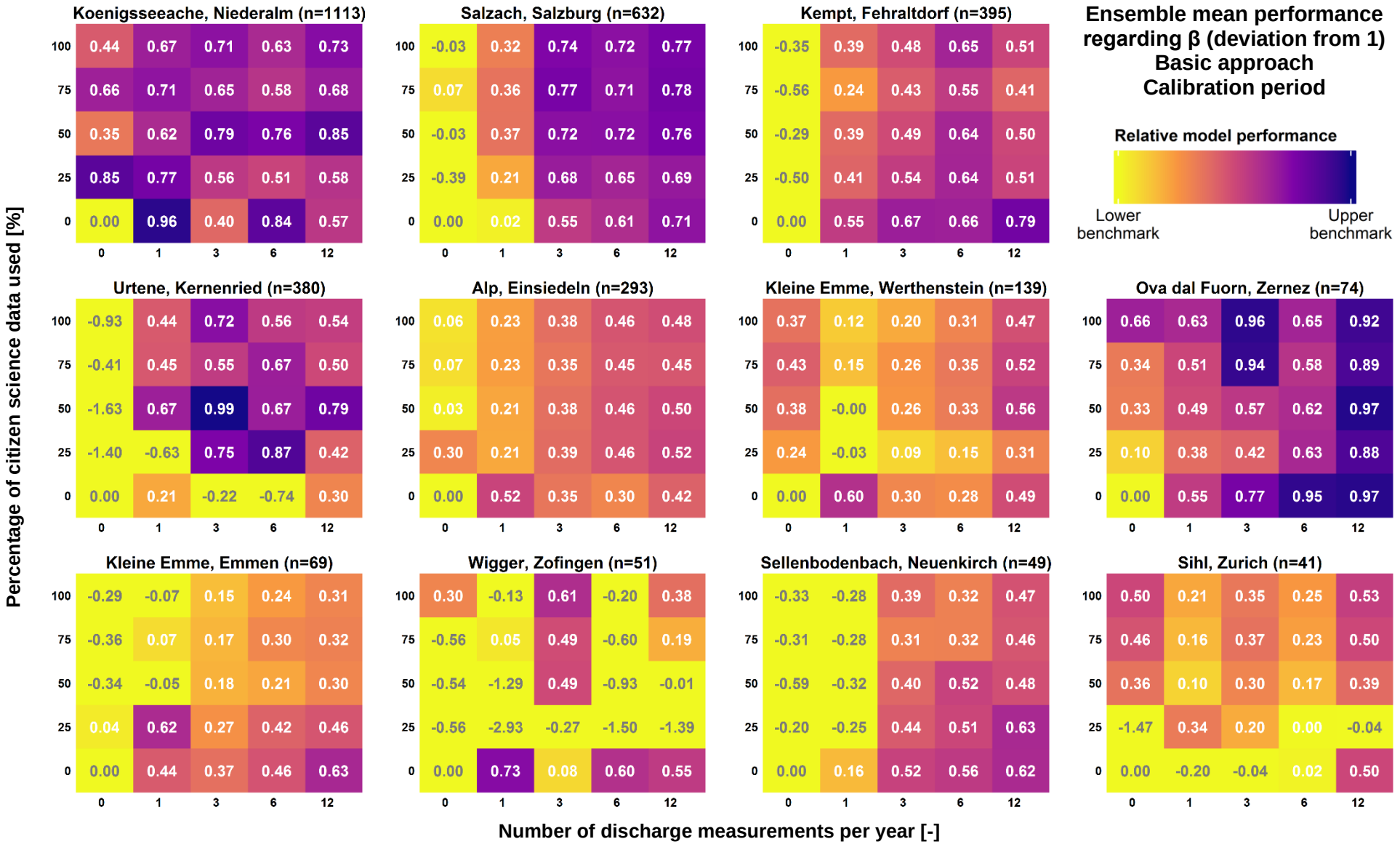


Figure 23: Relative performance of the ensemble mean for all catchments and all scenarios, considering only β (second component of the NPE). Use of basic approach, results for the calibration period. The number of citizen science observations corresponding to 100% is given as n after the name of the catchment.

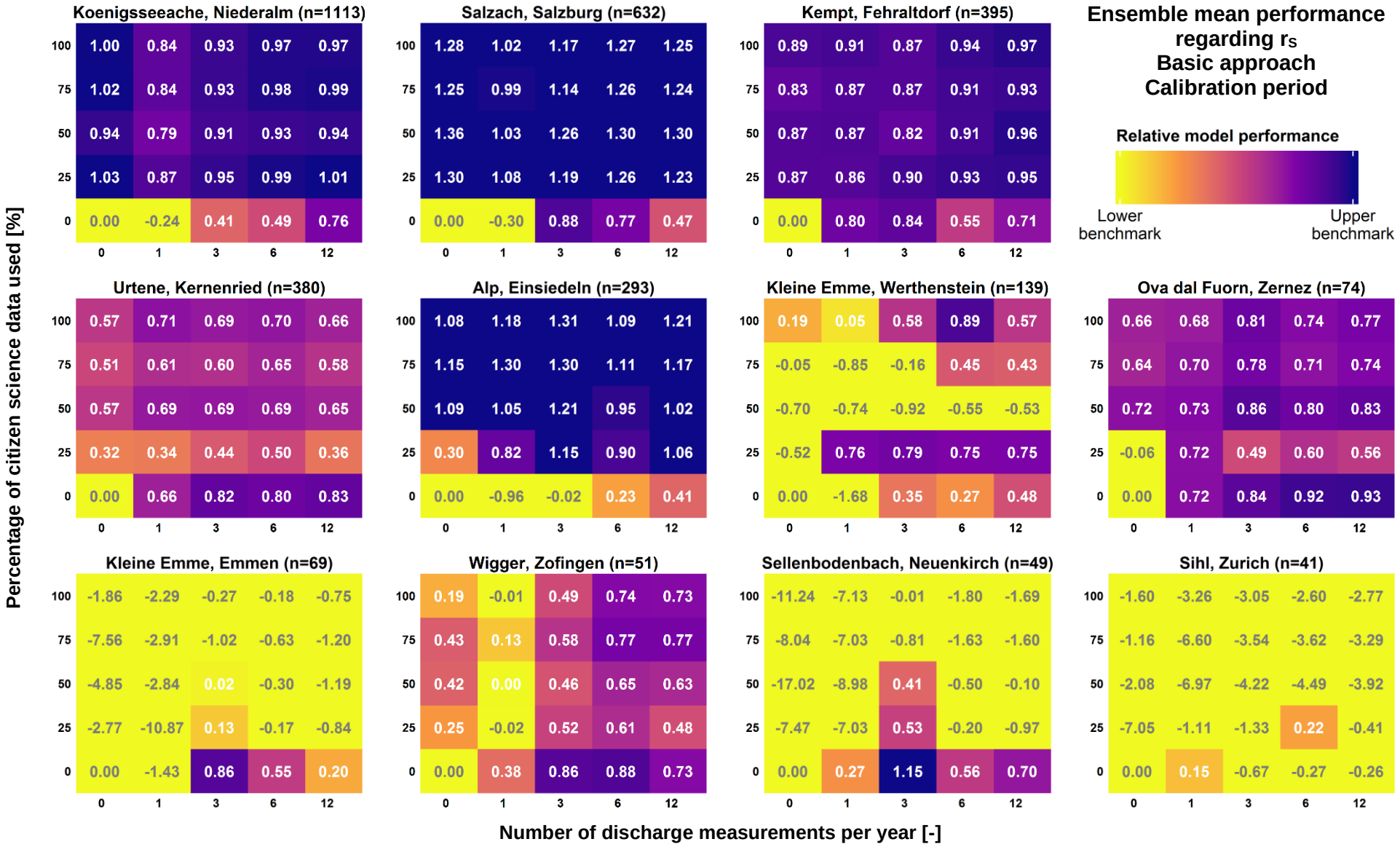


Figure 24: Relative performance of the ensemble mean for all catchments and all scenarios, considering only rs (third component of the NPE). Use of basic approach, results for the calibration period. The number of citizen science observations corresponding to 100% is given as n after the name of the catchment.

The one million parameter sets were ranked differently in each data availability scenario than when the full discharge time series was available to rank the parameter sets (Figure 25 and appendix 10.7). In the squares, each point represents one parameter set. In the x-direction, the parameter sets were plotted according to their logarithmic rank regarding the calibration against the full discharge time series (upper benchmark calibrated with the Monte Carlo approach). In the y-direction, the parameter sets were plotted according to their logarithmic rank regarding the calibration of the scenarios. The scenarios were sorted as in the preceding figures, whereby at the position of the lower benchmark (lower left corner) the upper benchmark was used to show the ideal case. The pink line is the linear regression line. The two red lines indicate rank 100, thus all points on the left and below the red line respectively were chosen as one of the top 100 parameter sets. The number of points in the lower left

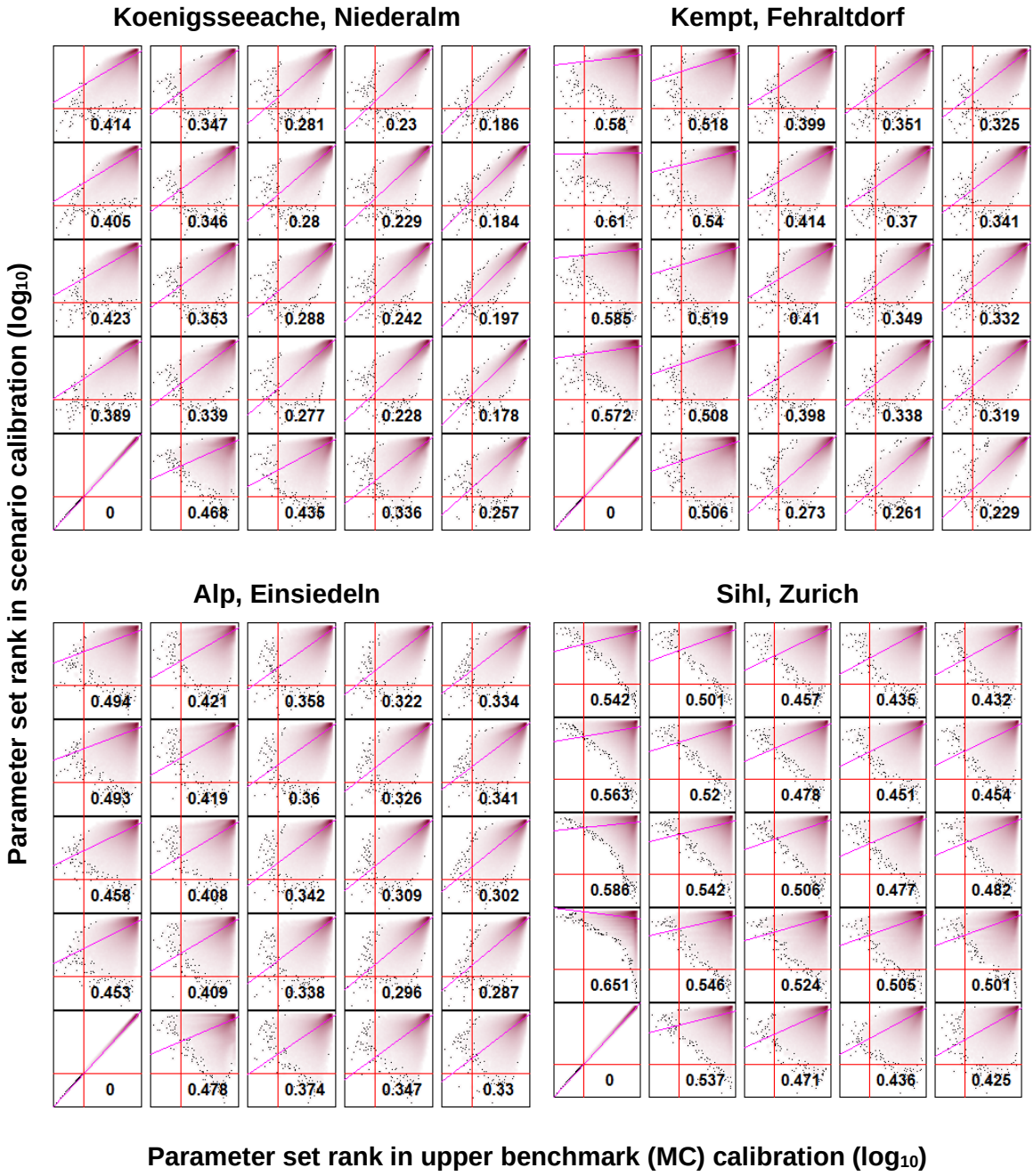


Figure 25: Plots showing the logarithmic rank of each of the one million parameter sets in each scenario against the logarithmic rank in the upper benchmark for four exemplary catchments. Each square represents one scenario. The pink line shows the linear regression, the number is the root mean squared error.

corner of each square indicates the number of parameter sets that were in the top 100 regarding the full discharge time series as well as regarding the corresponding scenario. The number given for each scenario is the root mean squared error of the position of the parameter sets to the 1:1-line.

Scenarios resulting in a good model performance ranked the parameter sets similarly to the ranking of the upper benchmark and thus showed a small root mean squared error and a linear regression line close to the 1:1-line. This could be observed well for the good scenarios of the Koenigsseeache, Kempt and Alp. Scenarios resulting in a bad model performance tended to rank the parameter sets opposite to the ranking of the upper benchmark. This was best visible for the Sihl. Such an opposite ranking could also be observed for other catchments such as the Sellenbodenbach (see appendix 10.7). In the scenarios using citizen science data only at the Kempt, the ranking also differed strongly from the ranking of the upper benchmark which resulted in a bad model performance for these scenarios. However, since the ranking was not opposite to the one of the upper benchmark, and the ranking of the scenarios using discharge measurements only was similar to the one of the upper benchmark, the mixed scenarios still ended up with a reasonable ranking and a good model performance at the Kempt.

The number of parameter sets that were shared in the top 100 parameter sets of two neighbouring scenarios differed between 0 and 90 (Figure 26 and appendix 10.8). The white fields contain the names of the scenarios and the coloured fields in between them contain the number of parameter sets that were in the top 100 for both scenarios.

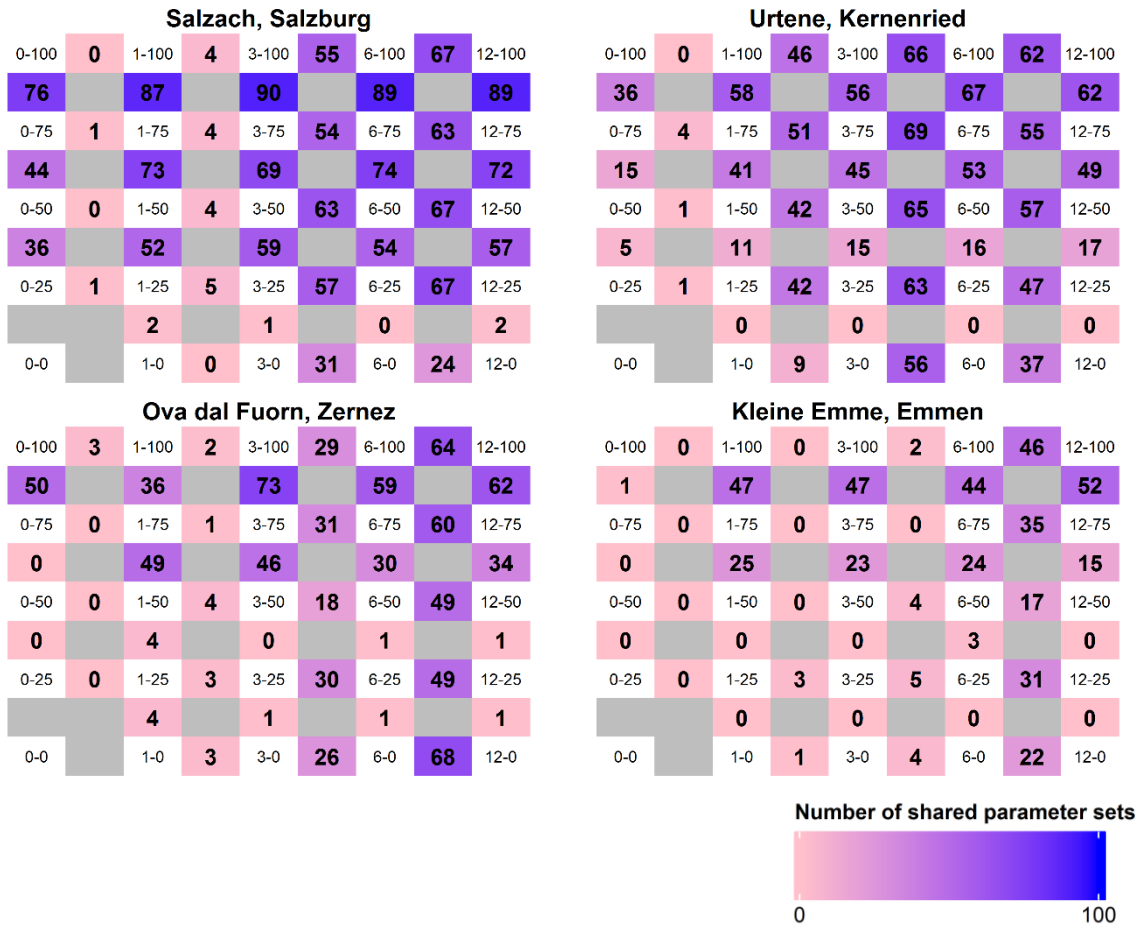


Figure 26: The number of the top 100 parameter sets that are shared among neighbouring scenarios, for four exemplary catchments. The names of the scenarios are given in the white fields, the number of shared parameter sets in the coloured fields, whereby a high number is indicated with a dark colour and a low number is indicated with a bright colour. The grey fields serve as placeholders.

In the examples shown here, but also in all other catchments, the top 100 parameter sets got replaced completely or almost completely when considering two data types instead of only one, i.e., the number of shared parameter sets between the scenarios 1-0, 3-0, 6-0, 12-0 and 1-25, 3-25, 6-25, 12-25 as well as between 0-25, 0-50, 0-75, 0-100 and 1-25, 1-50, 1-75, 1-100 was always very small. This implies that the parameter sets that got a low rank regarding the Spearman rank correlation with the water level class observations did usually get a high rank regarding the NPE with the discharge measurements, or in other words, a rank that was too high to be among the top 100 when the mean rank was considered. The same was valid in the other direction, too.

If only the mixed scenarios or only the scenarios using one type of data were considered, there was a trend for more shared parameter sets if more data was used, i.e., the ranking of the one million parameter sets stabilized if more data was used. This trend was visible for all catchments, no matter how well-performing the simulations with the different data availability scenarios were. In general, the number of shared parameter sets got highest when moving from left to right or from bottom to top in those catchments that showed good model performances. This was well visible for the Salzach (Figure 26). In catchments in which the approach did not work that well, as for example at the Kleine Emme in Emmen, this increase was less pronounced, respectively, there were fewer parameter sets shared in general among the scenarios. The constraints given by the data available in these catchments thus seemed to be too weak to find well-performing parameter sets.

Another indicator for a too weak constraint induced by the data available in catchments in which the approach did not work well could be found in the resulting parameter values compared to the resulting parameter values in the parameter sets used for the upper benchmark (Figure 27 and appendix 10.9).

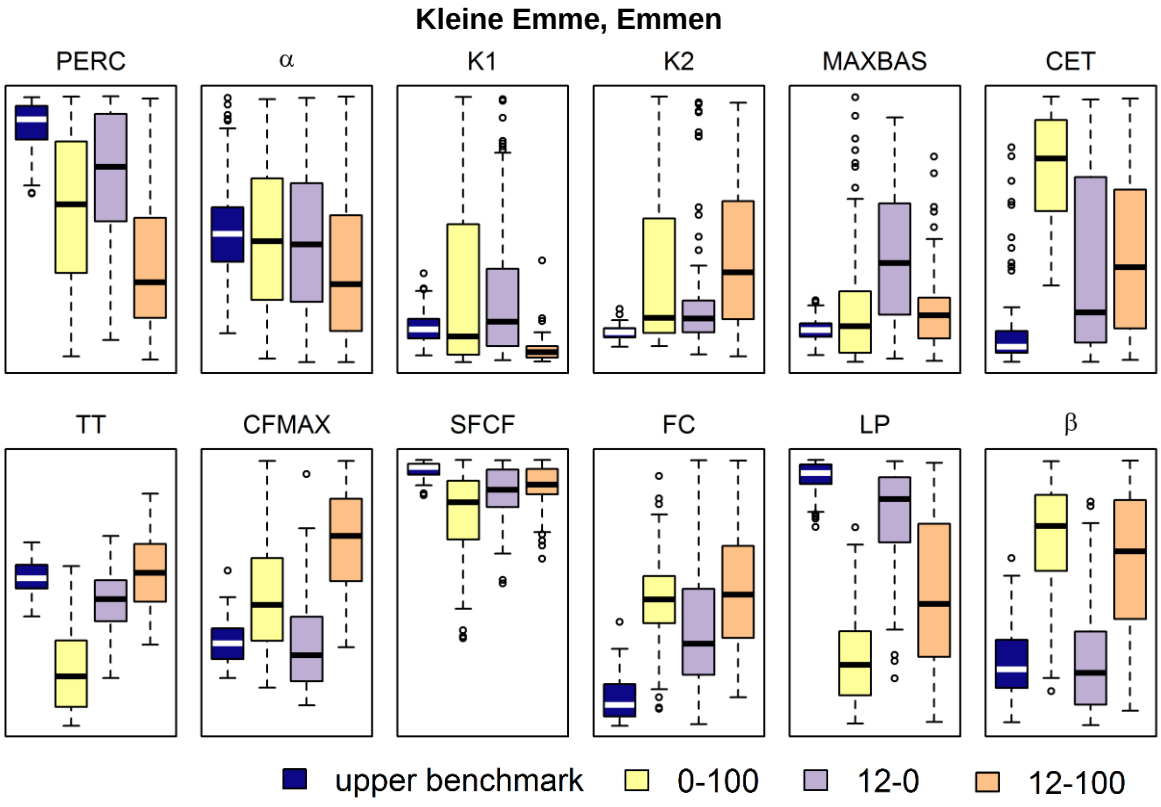


Figure 27: Distribution of the parameter values for the 100 parameter sets building the ensemble mean of the upper benchmark (calibration with the GAP approach) and the scenarios 0-100, 12-0 and 12-100 (i.e., the corners of Table 5), for the catchment of the Kleine Emme in Emmen. The y-axes cover the ranges that were allowed for each parameter (Table 4).

At the Kleine Emme in Emmen, the use of all water level class observations (0-100) acted disinformatively and a set of twelve discharge measurements per hydrological year (12-0) acted informatively. The scenario using both these input data sets for calibration (12-100) led to a medium relative performance, i.e., the valuable information contained in the discharge measurements did not get lost completely when combined with the citizen science data. The distribution of the parameter values obtained in scenario 0-100 differed quite strongly from the one of the parameter values obtained in the upper benchmark. This was especially strong in the soil routine. The difference was less strong for scenario 12-0 (and partially also for scenario 12-100) which indicates that in this catchment, the information contained in the twelve discharge measurements was helpful to find good parameter values and the information (or disinformation) contained in the water level class observations was not.

Also in most other catchments, the parameter values obtained when calibrating the model with twelve discharge measurements per year were more similar to those of the upper benchmark than the parameter values obtained when calibrating the model with all available citizen science observations (see appendix 10.9). At the Koenigsseeache, where scenario 0-100 reached a similar model performance as scenario 12-0, this difference was less pronounced. Even though parameter values can compensate for each other in the HBV model, more informative input data in general lead to a better constraint of the parameter values and thus a better model performance.

5.3 Additional mean discharge

Different accuracies of estimates of the mean discharge were tested as an additional constraint for the model. To do so, the one million parameter sets were filtered according to their resulting volume error. The number of parameter sets that fulfilled the conditions of each filter are given in Table 8.

Table 8: Number of parameter sets leading to a volume error smaller than or equal to a certain percentage.

Catchment	2.5%	5%	10%	20%	30%	50%
Koenigsseeache, Niederalm	64'052	127'101	250'286	467'386	640'028	880'505
Salzach, Salzburg	27'048	55'058	117'854	291'615	505'433	888'213
Kempt, Fehraltdorf	40'779	81'587	164'592	340'257	524'399	829'653
Urtene, Kernenried	64'435	128'067	251'787	479'091	666'424	906'490
Alp, Einsiedeln	4'871	9'899	22'649	86'968	287'940	961'711
Kleine Emme, Werthenstein	10'307	21'322	49'196	162'063	408'390	960'263
Ova dal Fuorn, Zernez	42'301	84'022	167'154	329'825	482'672	760'534
Kleine Emme, Emmen	8'903	18'430	43'742	146'999	375'778	930'507
Wigger, Zofingen	74'725	148'985	295'843	564'317	767'224	955'032
Sellenbodenbach, Neuenkirch	600	1291	3634	21'737	82'290	416'315
Sihl, Zurich	71'157	141'809	277'524	515'819	698'318	908'770

In general, the application of such a filter improved the model performance. If good results were already achieved without a filter, the increase in model performance was less pronounced than when the unfiltered results were rather bad. The impact of the different filters on the different scenarios was similar in the calibration period (Figure 28) as in the validation period (Figure 29). A better estimate of the mean annual discharge (represented by a narrower filter or in other words a stronger constraint on the parameter sets that were allowed to be part of the top 100 parameter sets) did not always lead to a better model performance: At the Alp and the Sellenbodenbach for example, the model performances got worse when a maximal volume error of 2.5% or 5% was allowed than when using a filter that allowed for a deviation of 10%. On the other hand, where a very narrow filter worked well (as for example at the Salzach and the Sihl), the 10% filter did not lead to a decrease in model performance either, i.e., it was not disadvantageous to use the 10% filter instead of the 2.5% or the 5% filter (Figure 28 and Figure 29).

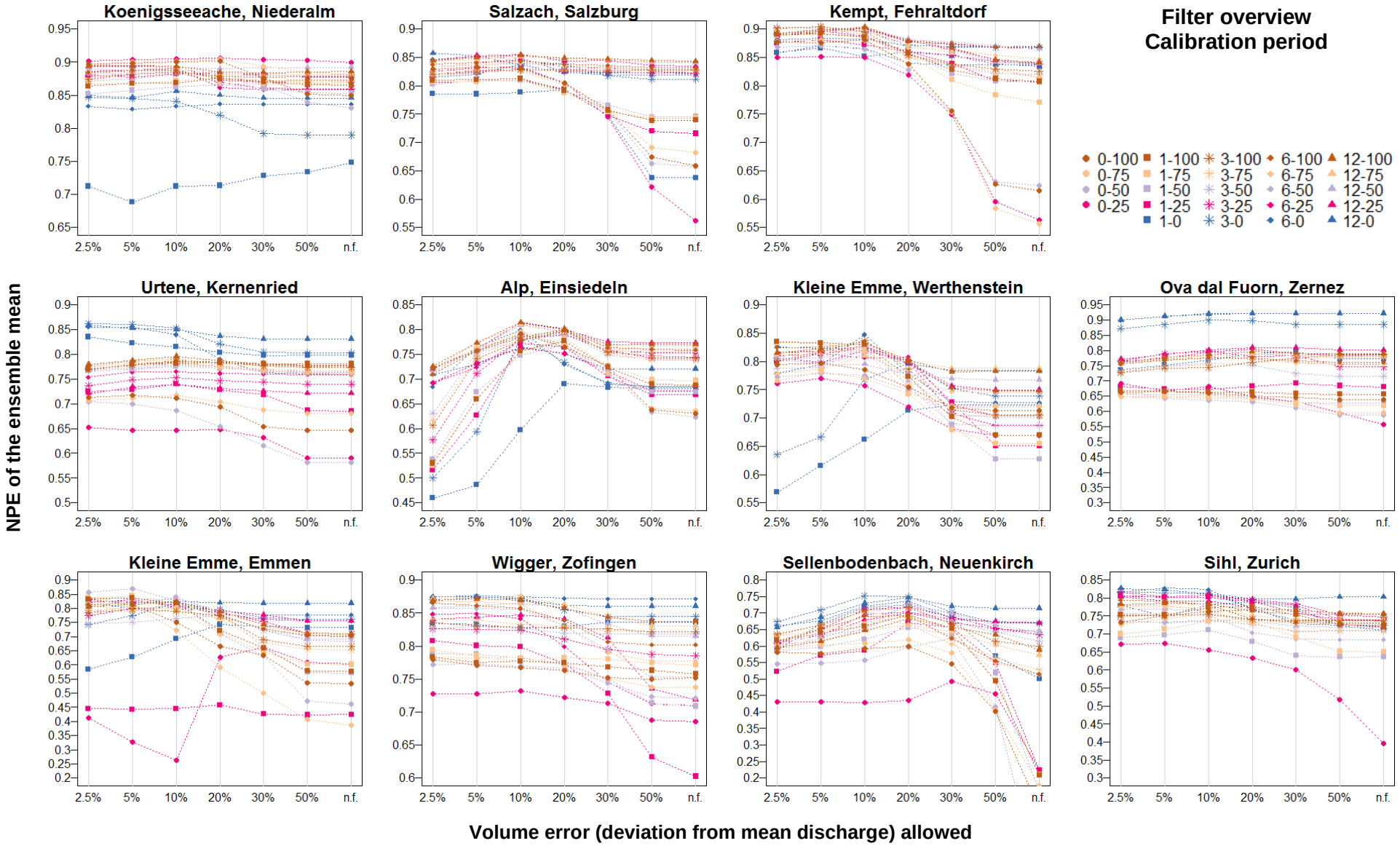


Figure 28: Impact of different volume error filters on the ensemble mean performances in all data availability scenarios and all catchments. The last column in each plot shows the model performances of all scenarios when no volume error filter is applied (corresponding to the basic approach). Results for the calibration period.

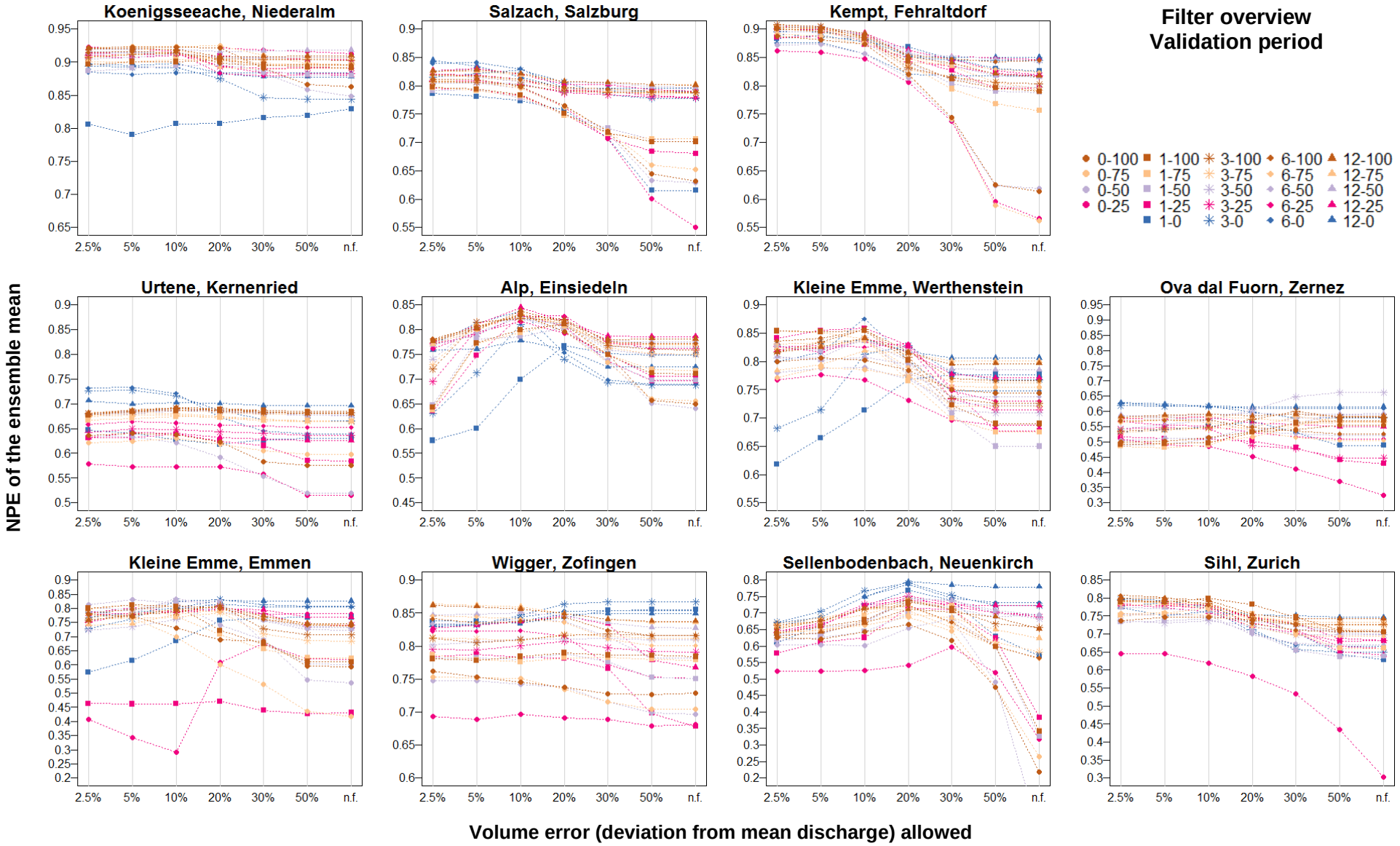


Figure 29: Impact of different volume error filters on the ensemble mean performances in all data availability scenarios and all catchments. The last column in each plot shows the model performances of all scenarios when no volume error filter is applied (corresponding to the basic approach). Results for the validation period

One reason why a narrower filter led to worse model performances than a wider filter can be found in the distribution of the parameter sets according to the NPE and the volume error that they produced (Figure 30). The same plots for the remaining seven catchments can be found in appendix 10.9. For catchments showing a distribution like the Salzach or the Sihl, the application of a volume error filter excluded most parameter sets that did not perform well regarding the NPE, thus the remaining parameter sets mostly led to a high NPE value. This was less pronounced at the Ova dal Fuorn, where the bandwidth of NPE performances remained high as the negative correlation between the NPE and the volume error was less strong. In catchments such as the Alp, the application of a too narrow filter even excluded the best-performing parameter sets regarding the NPE, while the parameter sets that surpassed the filter led to a low or medium performance in terms of the NPE.

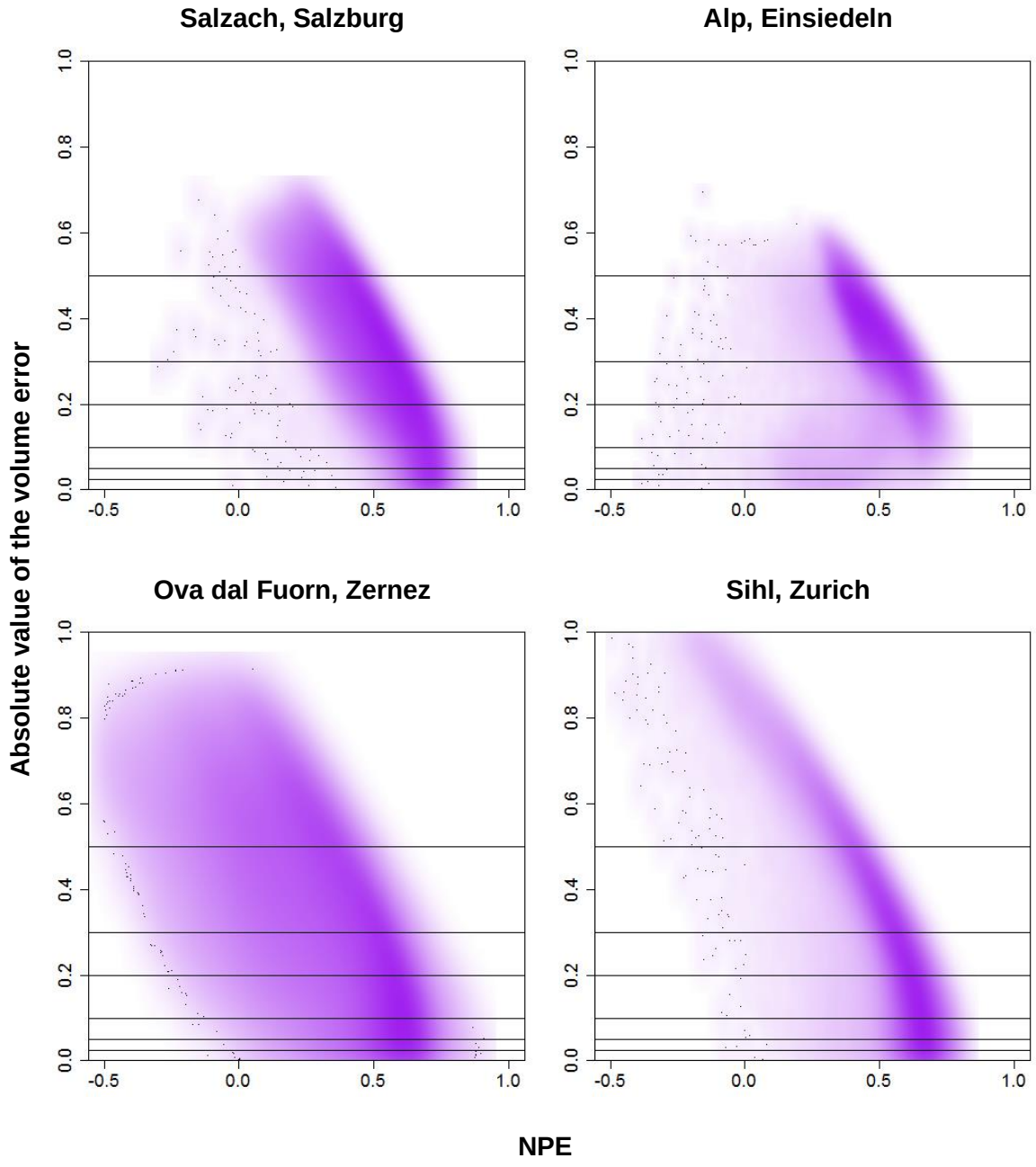


Figure 30: Density of the one million parameter sets when the NPE is plotted against the volume error. Results are shown for the calibration period and are very similar for the validation period.

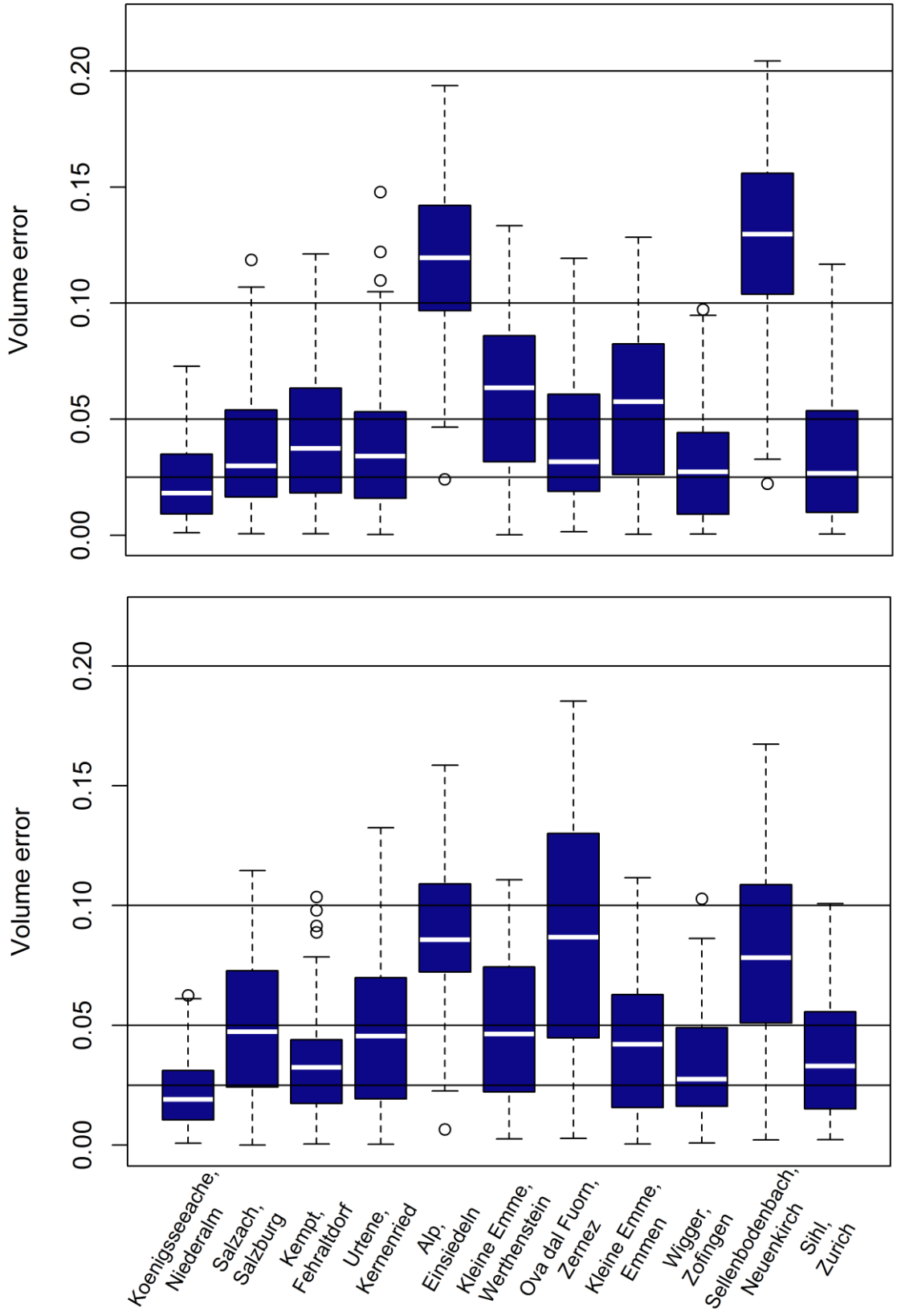


Figure 31: Volume error of the 100 parameter sets out of the one million that led to the highest NPE values (upper benchmark when the Monte Carlo approach was used for calibration) for each study catchment in the calibration period (top plot) and in the validation period (bottom plot). The horizontal lines indicate the 20%, 10%, 5% and 2.5% filter, i.e., parameter sets with a volume error below that line were excluded when applying this filter.

Consequently, the 100 parameter sets per catchment that led to the best performances regarding the NPE were not necessarily the catchments with the smallest volume errors (Figure 31). In all catchments except the Koenigsseeache, the median volume error of these parameter sets was above 2.5%. In most catchments, a considerable number of these top 100 parameter sets led to a volume error between 5% and 10%. A filter that allowed for a deviation larger than 10% often led to a slight decrease in model performance, i.e., the filters allowing for deviations of 20%, 30% or 50% did not constrain the model as well as the one allowing for a deviation of 10% (Figure 28 and Figure 29). However, even a filter that allowed for a deviation of 50% from the observed mean discharge was of value: The comparison with the situation without filter revealed a slight improvement or at least no decrease in model performance. This is surprising since a large majority of the one million parameter sets surpassed the 50%-filter in most catchments, thus the constraint was not very strong.

Based on the reasoning above, the filter allowing for a 10% deviation of the mean discharge was chosen to represent the situation when a mean discharge estimate was used additionally to the citizen science data and the discharge measurements. By applying this filter, almost all scenarios in almost all catchments outperformed the lower benchmark, i.e., the situation without any discharge information (Figure 33 for the calibration period and Figure 34 for the validation period). Compared to the situation in which no filter was applied (the basic approach, see section 5.2), the application of the 10% filter was advantageous: For almost all scenarios and almost all catchments, the model performance increased under the application of the filter (Figure 35 and Figure 36). The model performances resulting for the other filters that were considered can be found in appendix 10.11.

In most catchments, the increase in model performance was especially strong for those scenarios that had citizen science data only available for the calibration of the model. In scenario 0-50 at the Salzach for example, the parameter sets underestimating the summer discharge could be eliminated by applying the filter (Figure 32). The performances of the scenarios using citizen science data only without a filter often performed worse than other scenarios and thus left most room for improvement (see section 5.2). Additionally, the volume error was the only information about the amount of water in the stream when otherwise only using water level class data and thus was an important information in these scenarios. The improvement of the calibrations was so strong that the resulting model performances in these scenarios reached values similar to those reached in other scenarios that had discharge measurements that could help to constrain the model. Thus, the value of a mean discharge estimate is comparable to the value of several discharge measurements per hydrological year.

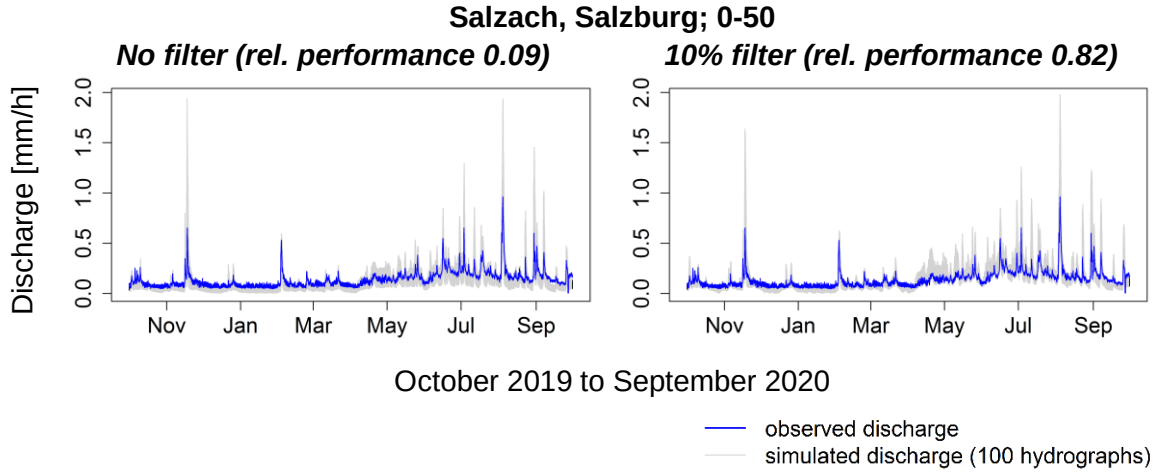


Figure 32: Observed and simulated hydrographs of the Salzach for the hydrological year 2020. Scenario 0-50 as in the basic approach and with an additional mean discharge filter allowing for a maximal deviation of 10% from the mean discharge. Observed hydrograph shown in blue, 100 simulated hydrographs building the ensemble mean shown in grey.

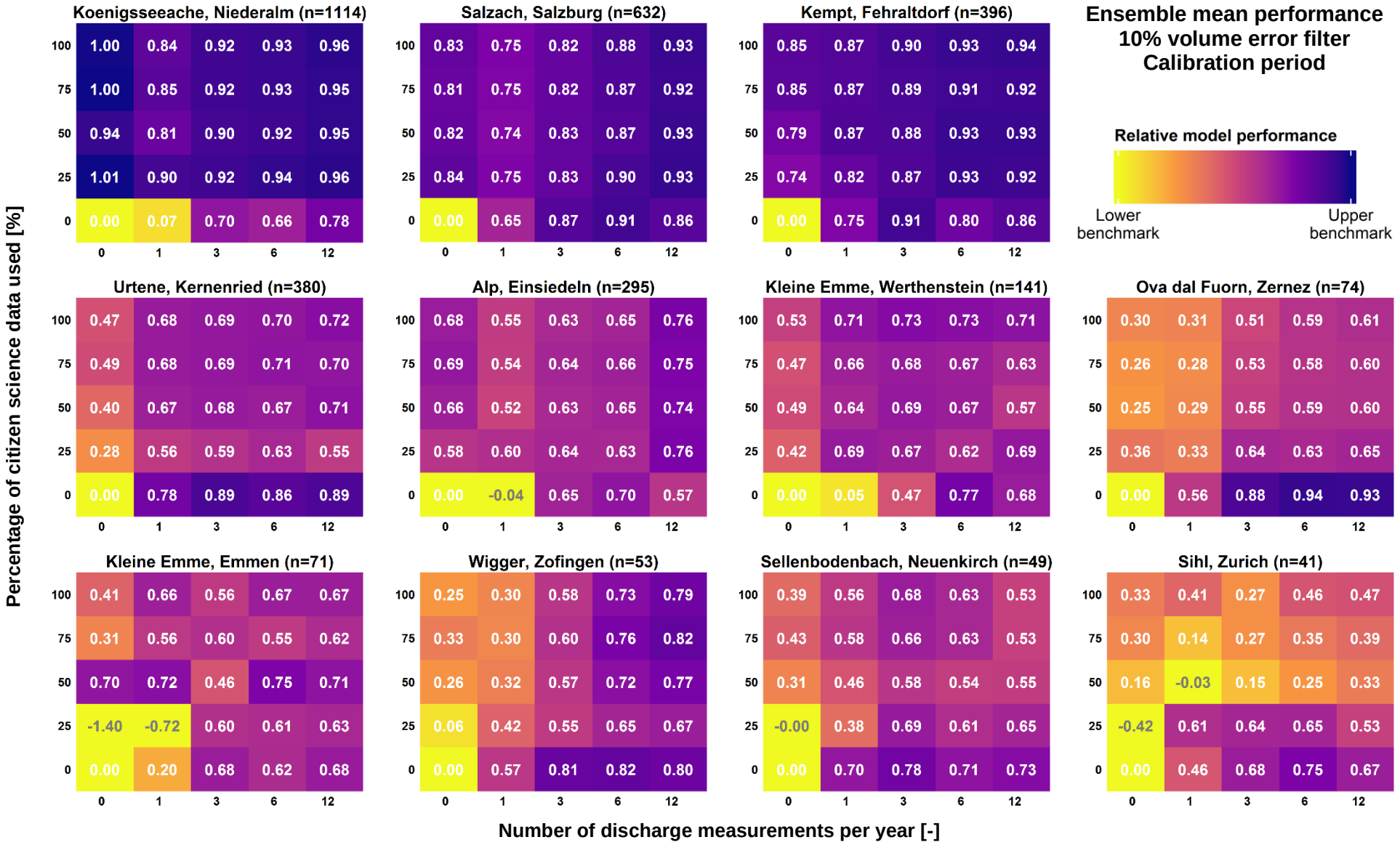


Figure 33: Relative performance of the ensemble mean for all catchments and all scenarios, when allowing a volume error of 10%. Results for the calibration period. The number of citizen science observations corresponding to 100% is given as n after the name of the catchment.

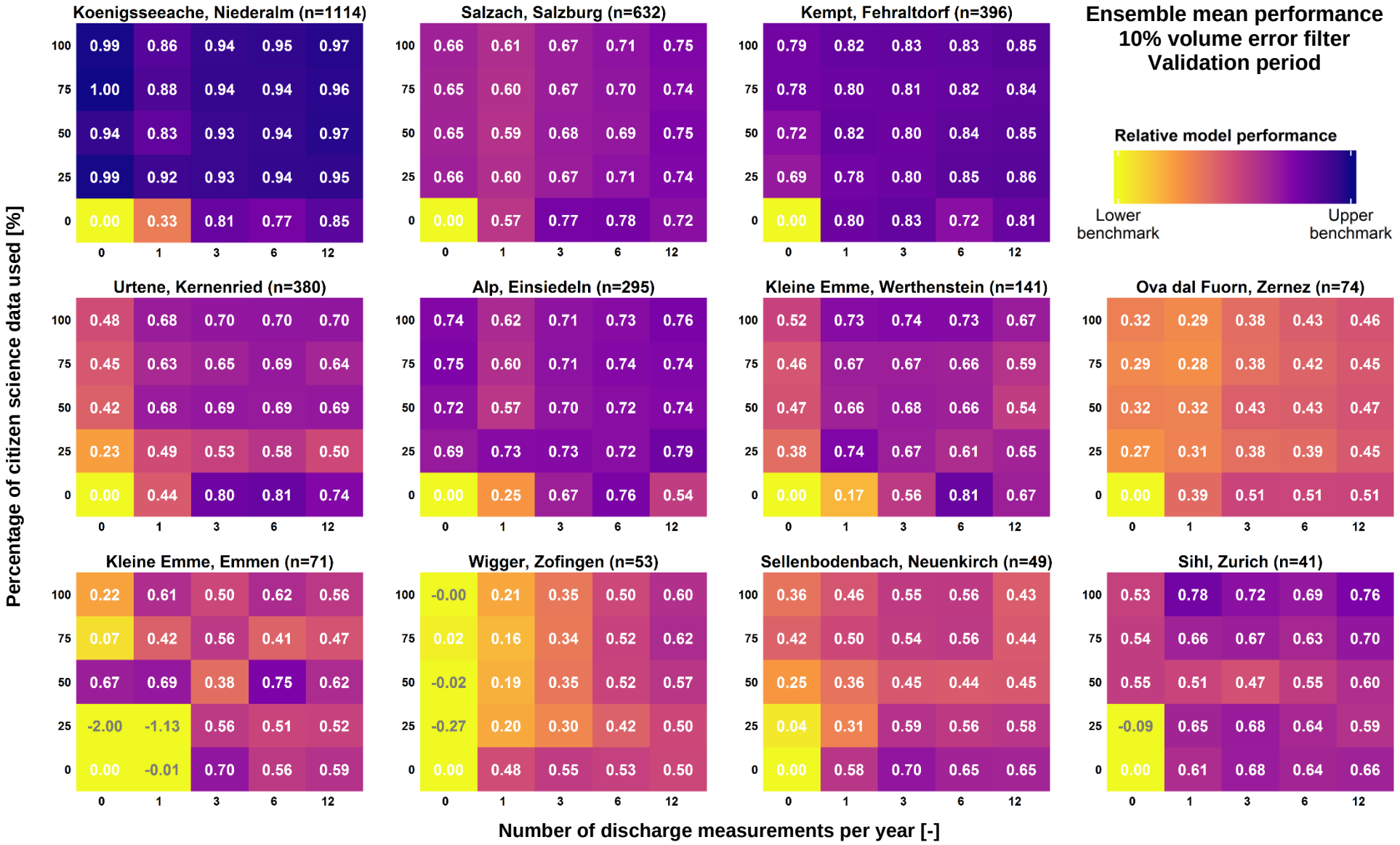


Figure 34: Relative performance of the ensemble mean for all catchments and all scenarios, when allowing a volume error of 10%. Results for the validation period. The number of citizen science observations corresponding to 100% is given as n after the name of the catchment.

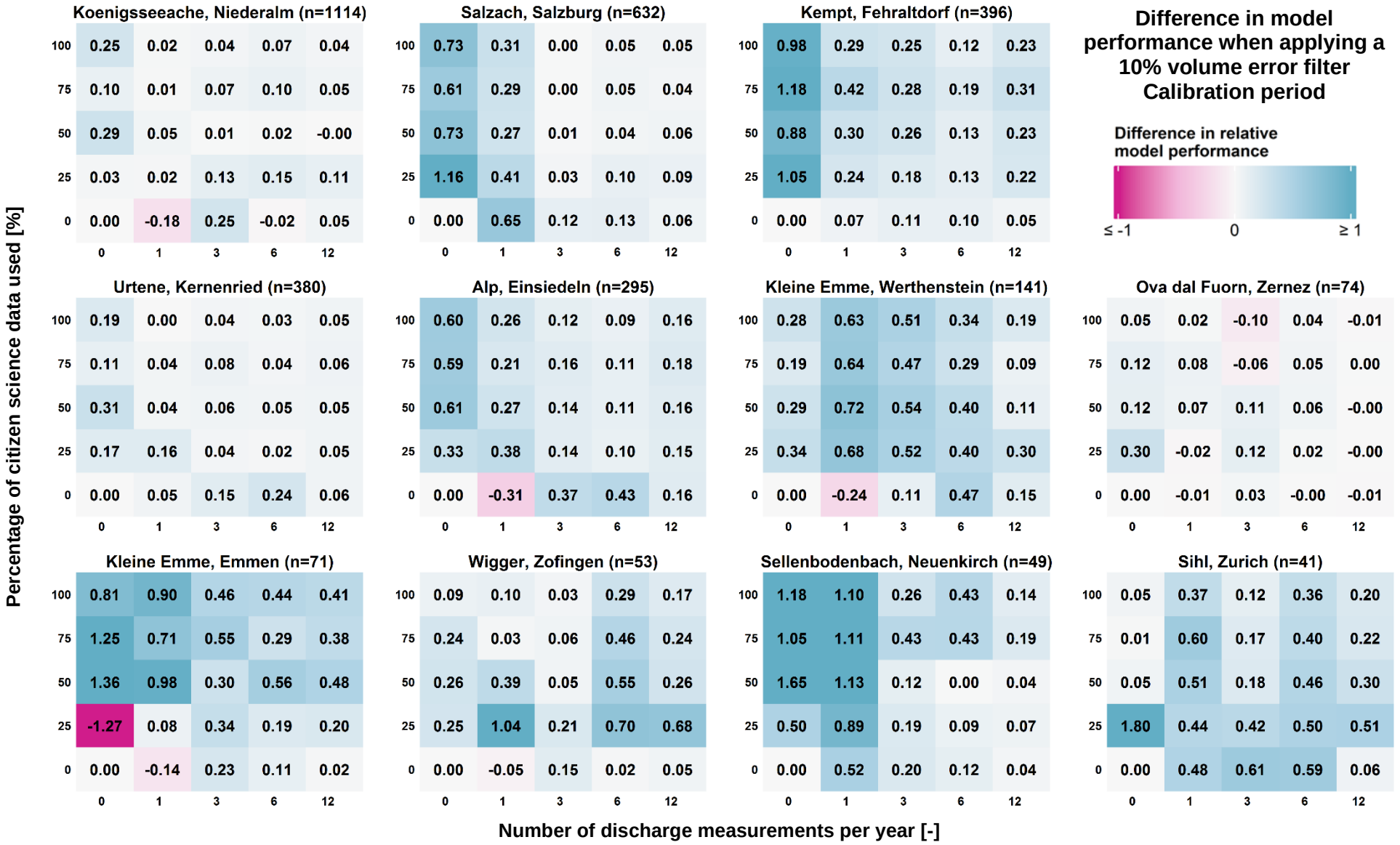


Figure 35: Change in the relative performance of the ensemble mean for all catchments and all scenarios when a filter allowing a volume error of 10% is applied. Results for the calibration period. The number of citizen science observations corresponding to 100% is given as n after the name of the catchment.

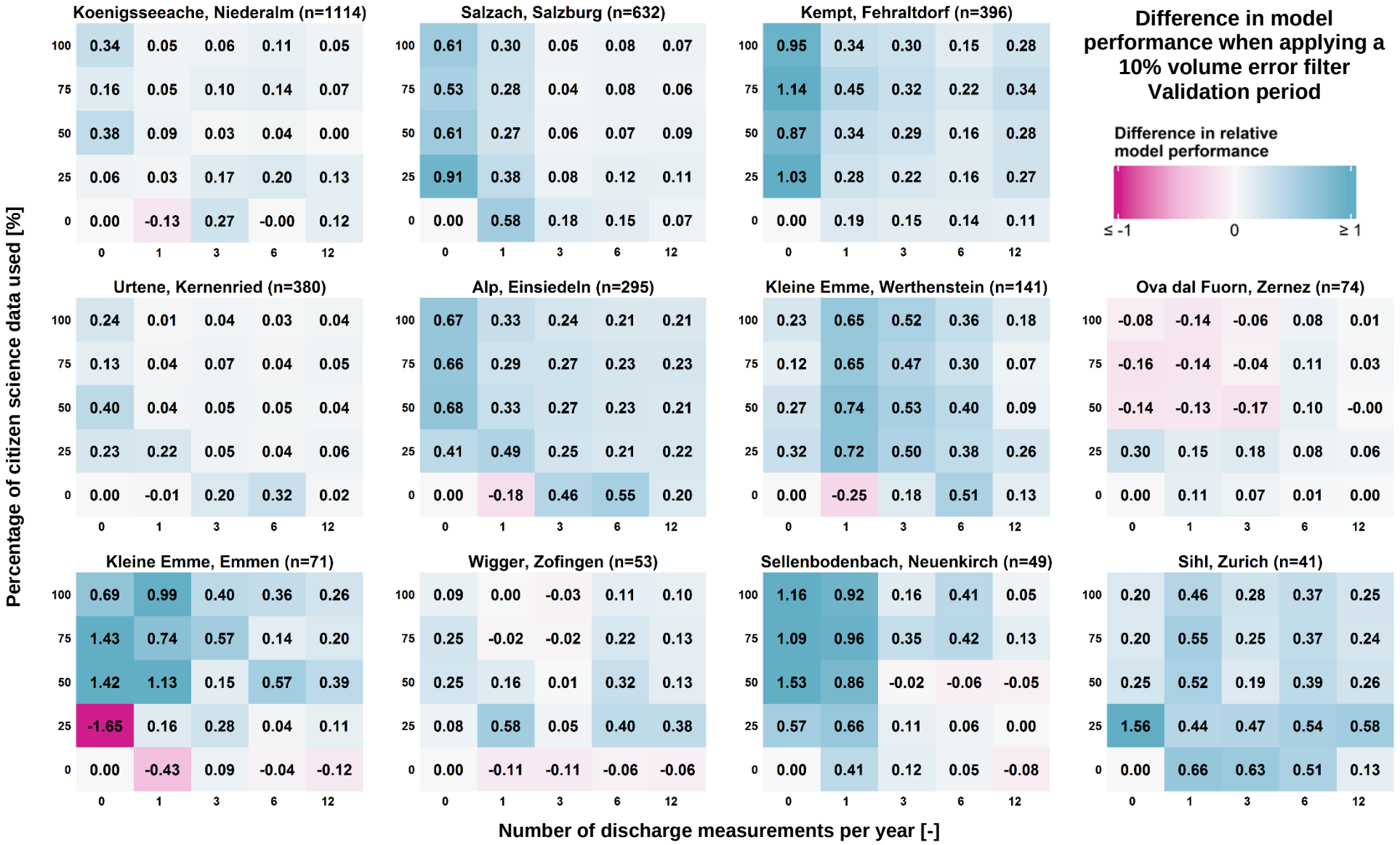


Figure 36: Change in the relative performance of the ensemble mean for all catchments and all scenarios when a filter allowing a volume error of 10% is applied. Results for the validation period. The number of citizen science observations corresponding to 100% is given as n after the name of the catchment.

5.4 Water levels instead of water level classes

In order to simulate the situation in which citizen scientists read the water level from a staff gauge with the accuracy and precision of the measurements done by the authorities, the water level class observations were replaced by the official water level measurements (to be precise, by discharge measurements that were treated as if they were water level measurements, see section 4.8.2). Other than in the preceding approach, the resulting relative model performances differed strongly among the different scenarios in most catchments (Figure 38 for the calibration period and Figure 39 for the validation period). Still, for most catchments and most scenarios there was an improvement in the model performances compared to the basic approach (Figure 40 and Figure 41).

Catchments with a lot of citizen science data that was highly correlated with the discharge time series did not profit much from the even higher correlated water level data. In general, catchments tended to profit more from the replacement if the correlation between the water level observations and the discharge time series was low. However, scenarios using citizen science data only and not performing well in the basic approach did still not reach performances better than the lower benchmark. This was especially the case for the Kempt, the Kleine Emme in Emmen, the Wigger and the Sellenbodenbach. Meanwhile, the performances of the scenarios using both, discharge measurements and water level data, were strongly improved in catchments with little citizen science data. Oftentimes, the lower benchmark could not be outperformed with the water level class data and got clearly outperformed with the water level measurements. The improvement was most pronounced in scenarios that did not perform well in the basic approach, i.e., where there was most room for improvement. Thus, the replacement brought all scenarios closer together. Especially at the Kleine Emme in Werthenstein, a slight performance drop could be observed in scenarios that were among the better performing scenarios in the basic approach. This further strengthened the effect of a smaller range among all relative model performances in each catchment. As nothing was changed in the scenarios that did not make use of any citizen science data at all, also the performances of these scenarios remained unchanged.

In the validation period, there was a performance drop when using water levels at the Ova dal Fuorn. As the validation period in this catchment already showed some unexpected behaviour in the basic approach, this drop was not considered any further. The most exceptional catchment in this approach was the Sihl with strong performance drops compared to the basic approach for most scenarios. The higher resolved water level measurements seemed to act disinformatively in this case and did not allow to simulate the hydrograph better than when using water level class observations (Figure 37).

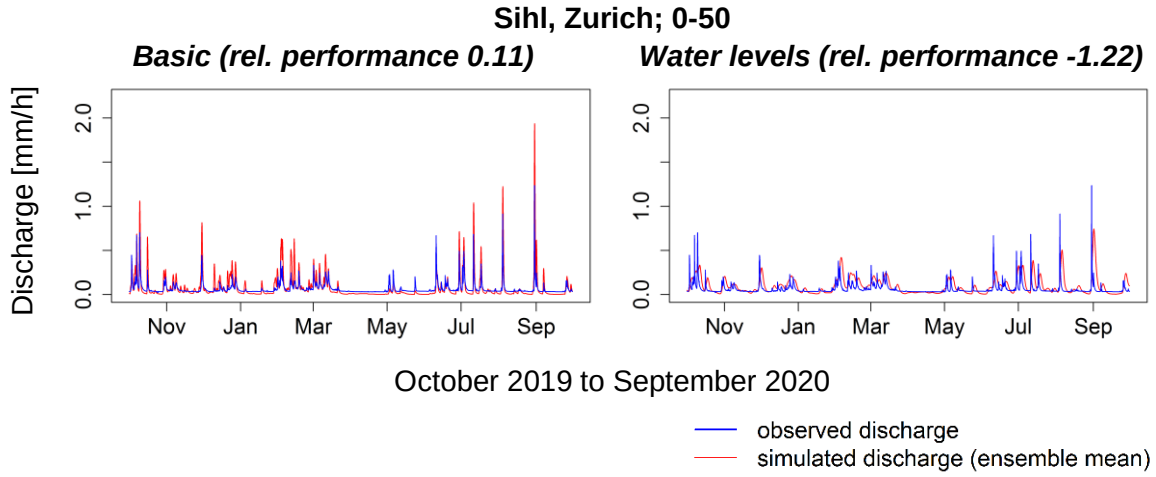


Figure 37: Example hydrographs of the Sihl for the hydrological year 2020. Scenario 0-50 using water level classes and water levels. Observed hydrograph shown in blue, ensemble mean of the top 100 simulated hydrographs shown in red.

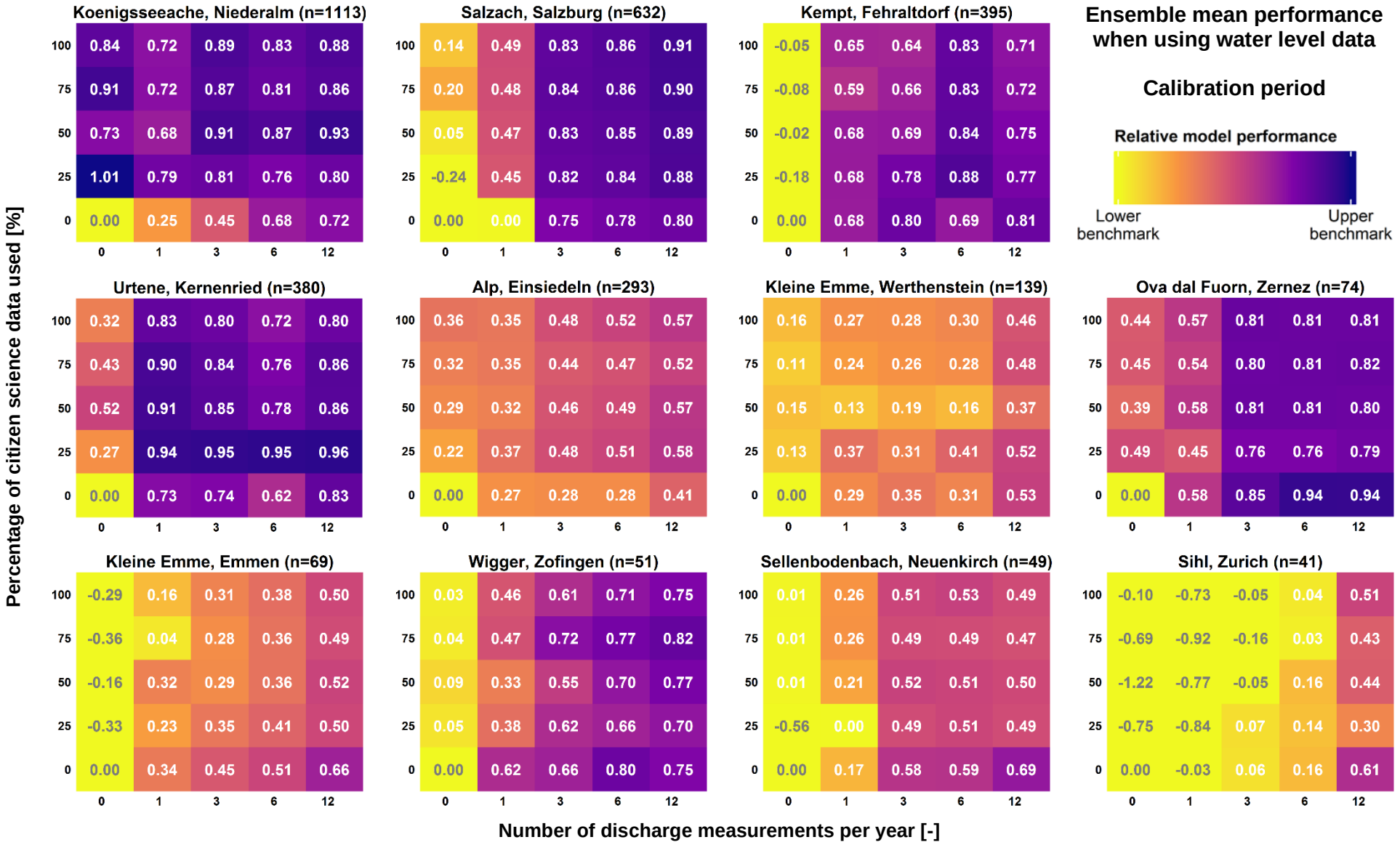


Figure 38: Relative performance of the ensemble mean for all catchments and all scenarios, when using water level data instead of water level class observations for the calibration of the model. Results for the calibration period. The number of citizen science observations corresponding to 100% is given as n after the name of the catchment.

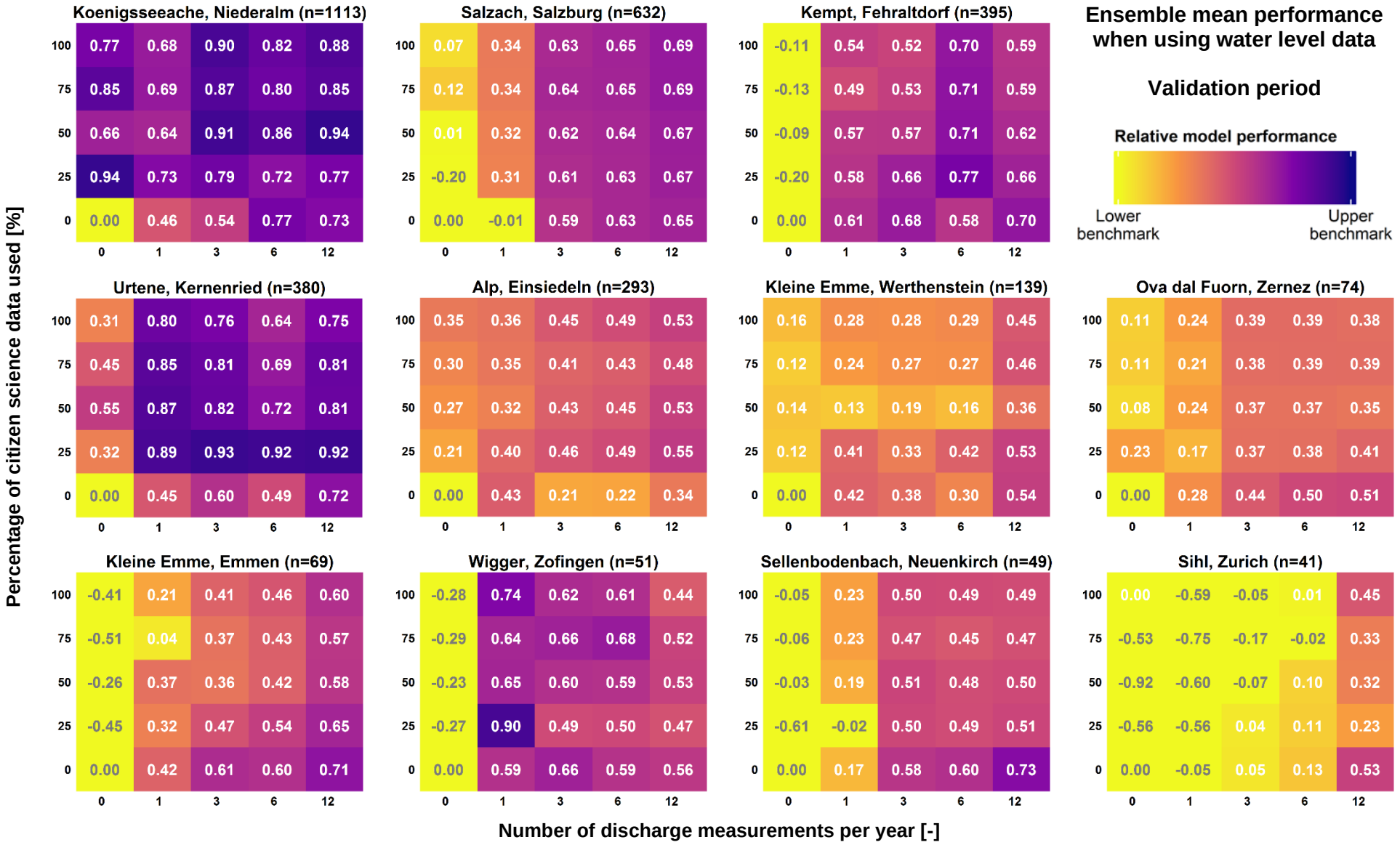


Figure 39: Relative performance of the ensemble mean for all catchments and all scenarios, when using water level data instead of water level class observations for the calibration of the model. Results for the validation period. The number of citizen science observations corresponding to 100% is given as n after the name of the catchment.

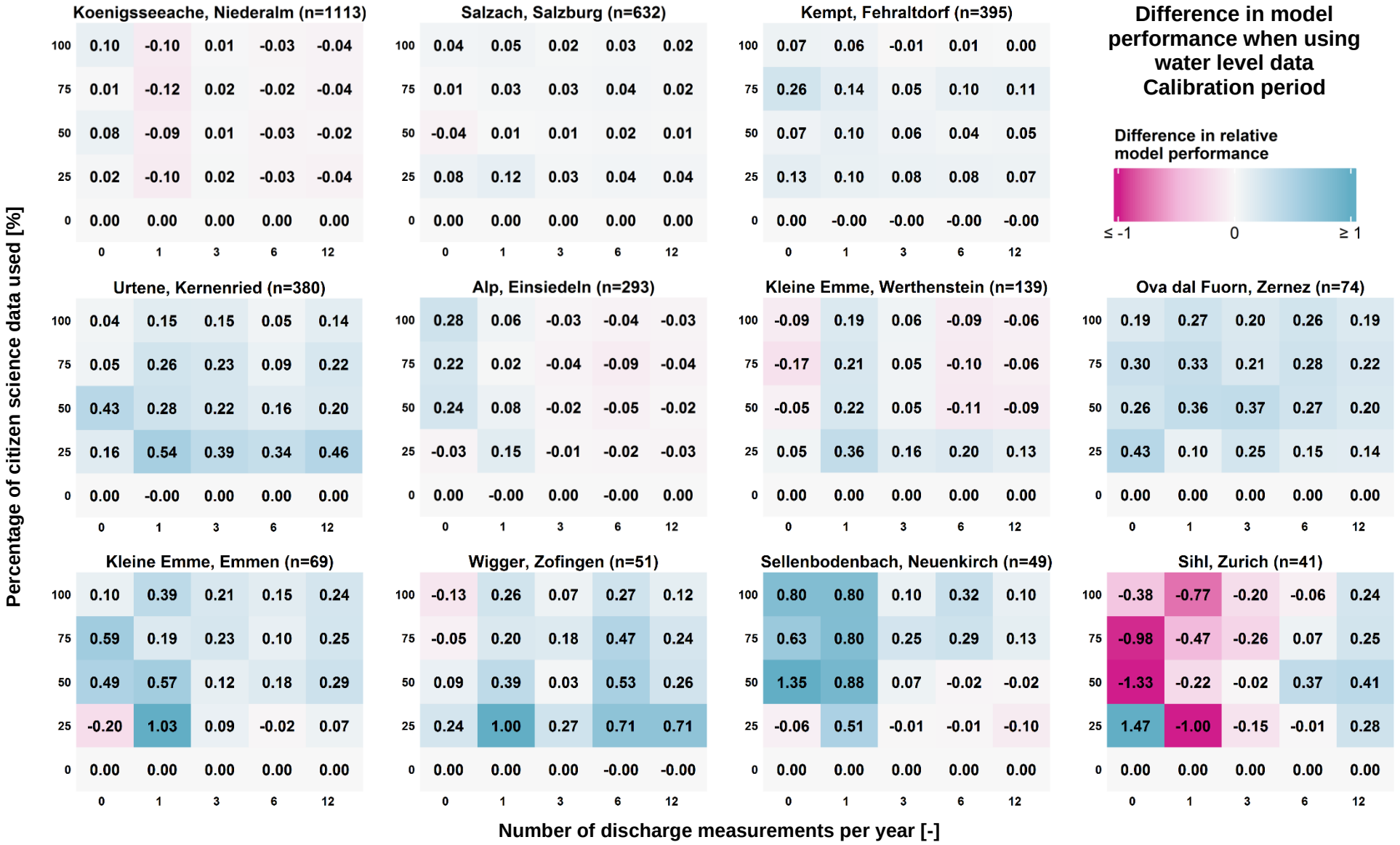


Figure 40: Change in the relative performance of the ensemble mean for all catchments and all scenarios when using water level data instead of water level class observations for the calibration of the model. Results for the calibration period. The number of citizen science observations corresponding to 100% is given as n after the name of the catchment.

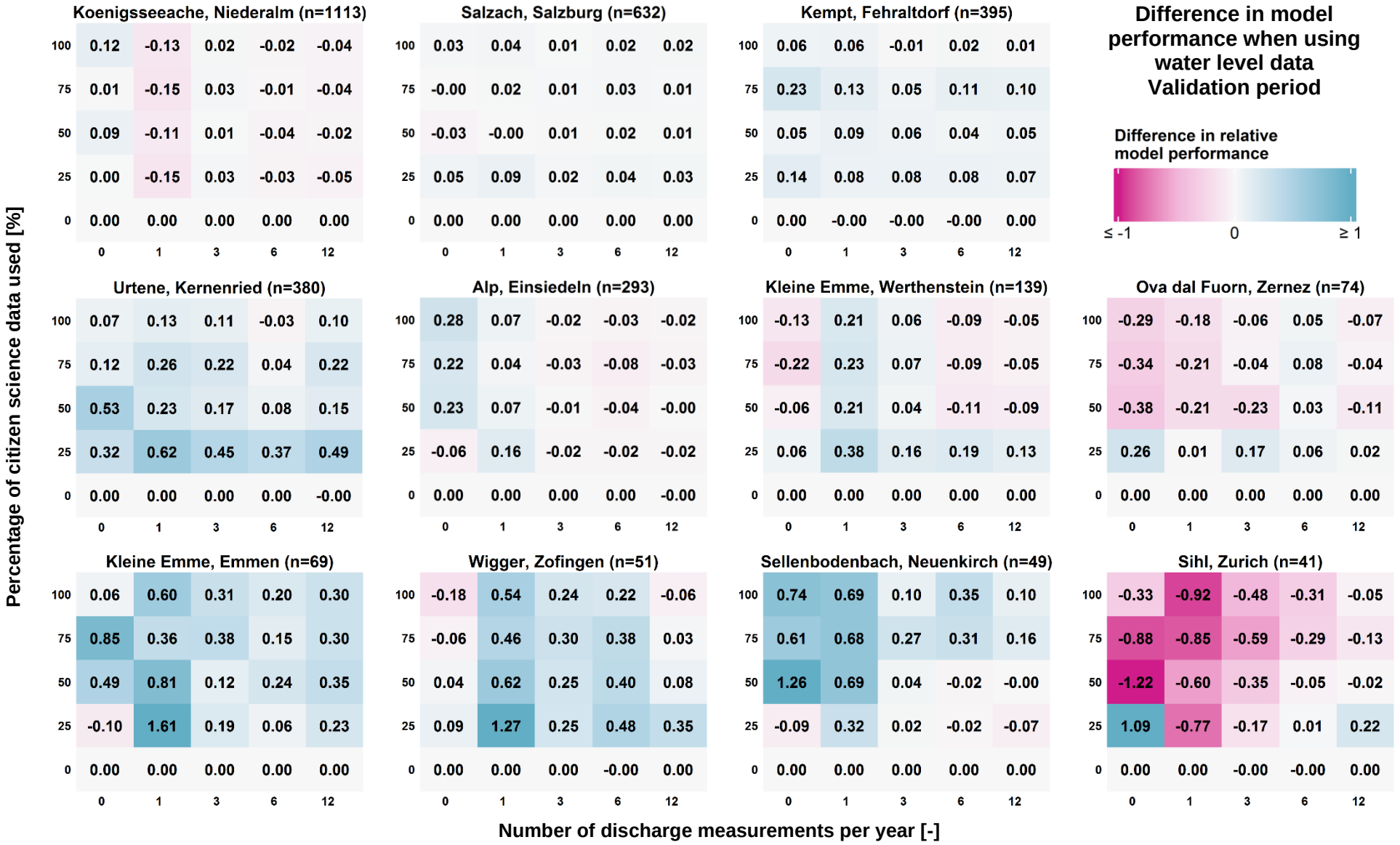


Figure 41: Change in the relative performance of the ensemble mean for all catchments and all scenarios when using water level data instead of water level class observations for the calibration of the model. Results for the validation period. The number of citizen science observations corresponding to 100% is given as n after the name of the catchment.

5.5 Water level class data checked by citizen scientists

5.5.1 Adjusted benchmarks

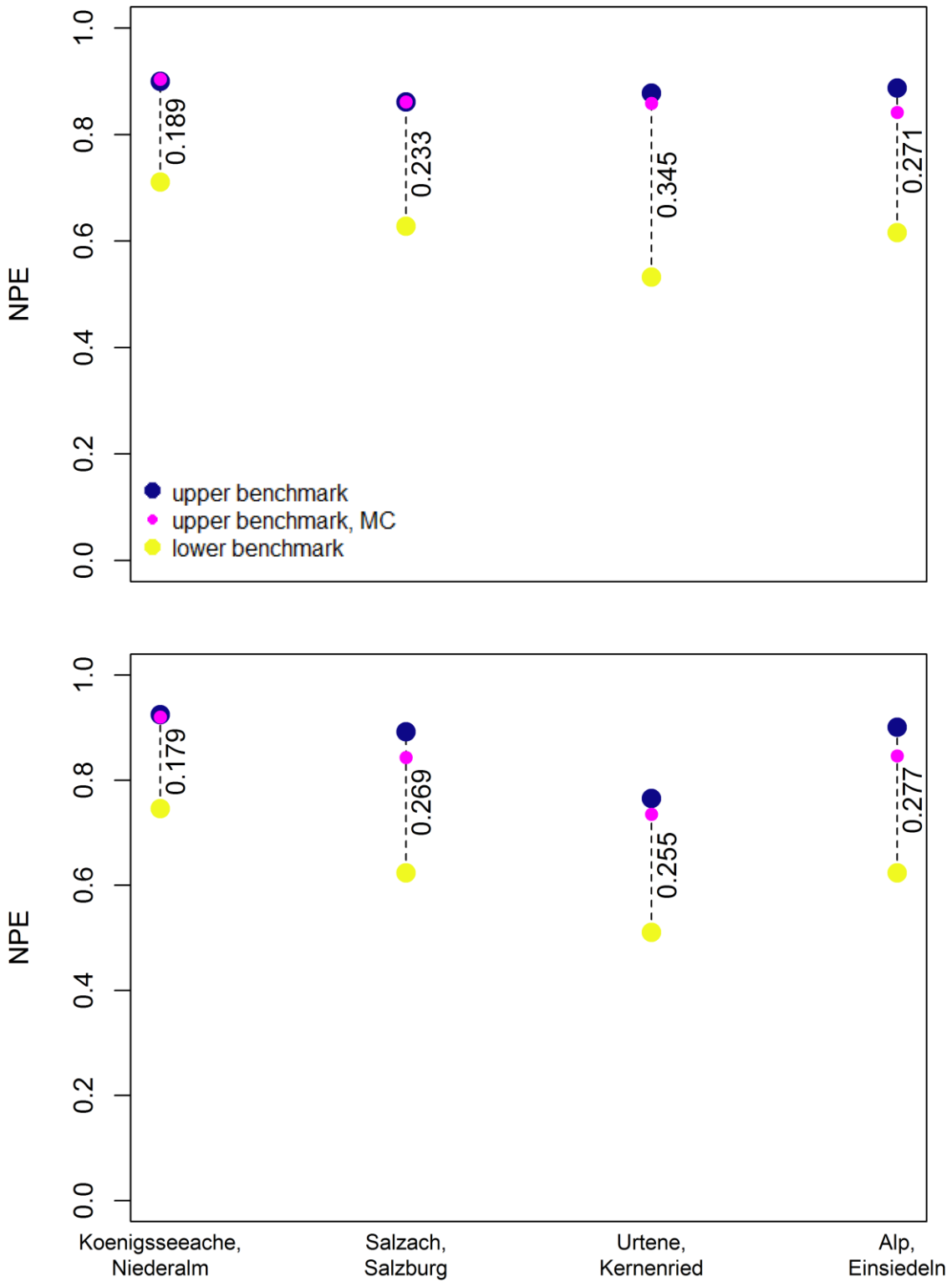


Figure 42: Benchmarks for the calibration period (upper plot) and the validation period (lower plot) of the catchments in which quality-controlled data was available. The number describes the difference between the upper and lower benchmark. The upper benchmarks for the Monte Carlo approach are plotted for comparison but were not used for the calculation of the relative performance. Note that the upper benchmark for the validation period is rather low for the Urtene.

The benchmarks for the calibration as well as for the validation period were calculated separately for this part of the thesis (Figure 42). Note that there was not much of a difference between the calibration and the validation period for all catchments except the Urtene: While the lower benchmark at the Urtene showed a rather bad performance for both periods, the upper benchmark resulted in a clearly worse performance during the validation period compared to the resulting model performance during the calibration period. This led to a larger span between the two benchmarks in the calibration period compared to the span in the validation period. Furthermore, note that the span between the upper and the lower benchmark at the Koenigsseeache is rather small for both periods, so the possible improvement that could be achieved when using citizen science data and discharge measurements was limited.

Again, the upper benchmark calibrated using the Monte Carlo approach is shown as a comparison but was not used for the calculations of the relative model performances.

5.5.2 Resulting model performances

To investigate on the value of the quality-control conducted by citizen scientists in the CrowdWater game, a modified basic approach (shorter calibration period, data points selected such that an amount as large as possible could be replaced with classified data from the game, see section 4.8.3) was conducted. These results were then compared to the results of the approach in which all water level class observations that had already been classified by at least 15 votes were replaced with the trimmed mean of all game votes.

Older data points get classified earlier than newer data points. Thus, there was a tendency that for the scenarios with a small amount of citizen science data (where all or almost all data can be replaced with data from the game), more citizen science observations from the beginning than from the end of the observation period were used (Figure 43). At the Koenigsseeache and at the Salzach, a wide range of observed water level classes was already covered by the observations used for the scenarios with 25% and 50% of the citizen science data. As already seen earlier (Figure 13), the range of water level classes resulting from the CrowdWater game was way smaller than the original range of water level classes observed with the CrowdWater app at the Urtene. Thus, in this catchment scenarios using only a small part of the citizen science data available had a smaller range of water level classes covered than scenarios using more citizen science data (and therefore more original values collected in the app). Aside the very highest and very lowest values observed at the Alp, a wide range of water level classes was also covered in the scenarios with only 25% or 50% of the citizen science data at the Alp. Note that in the full data set, there are observations in the water level classes between -3 to 6 available at the Koenigsseeache and the Salzach, but only observation in the water level classes between -1 and 3 at the Urtene and the Alp.

In the four catchments used here, three different cases could be observed when comparing the resulting model performances if app data only was used with the resulting model performances if the data was replaced with the values from the game where possible (Figure 45 for the calibration period and Figure 46 for the validation period). The top row shows the relative performances when app data only was used, the second row shows the relative performances when some of the data was replaced with data from the game and the third row shows the difference between the two approaches, whereby a negative value indicates a better model performance when app data only was used, and a positive value indicates an improvement if some of the app data was replaced with game data. The bottom row of each subplot does not change among these two approaches. Note that the percentage of citizen science data used (y-axis) still represents how much of the data that was available in total has been used and not how much of the data has been replaced with classified data from the game.

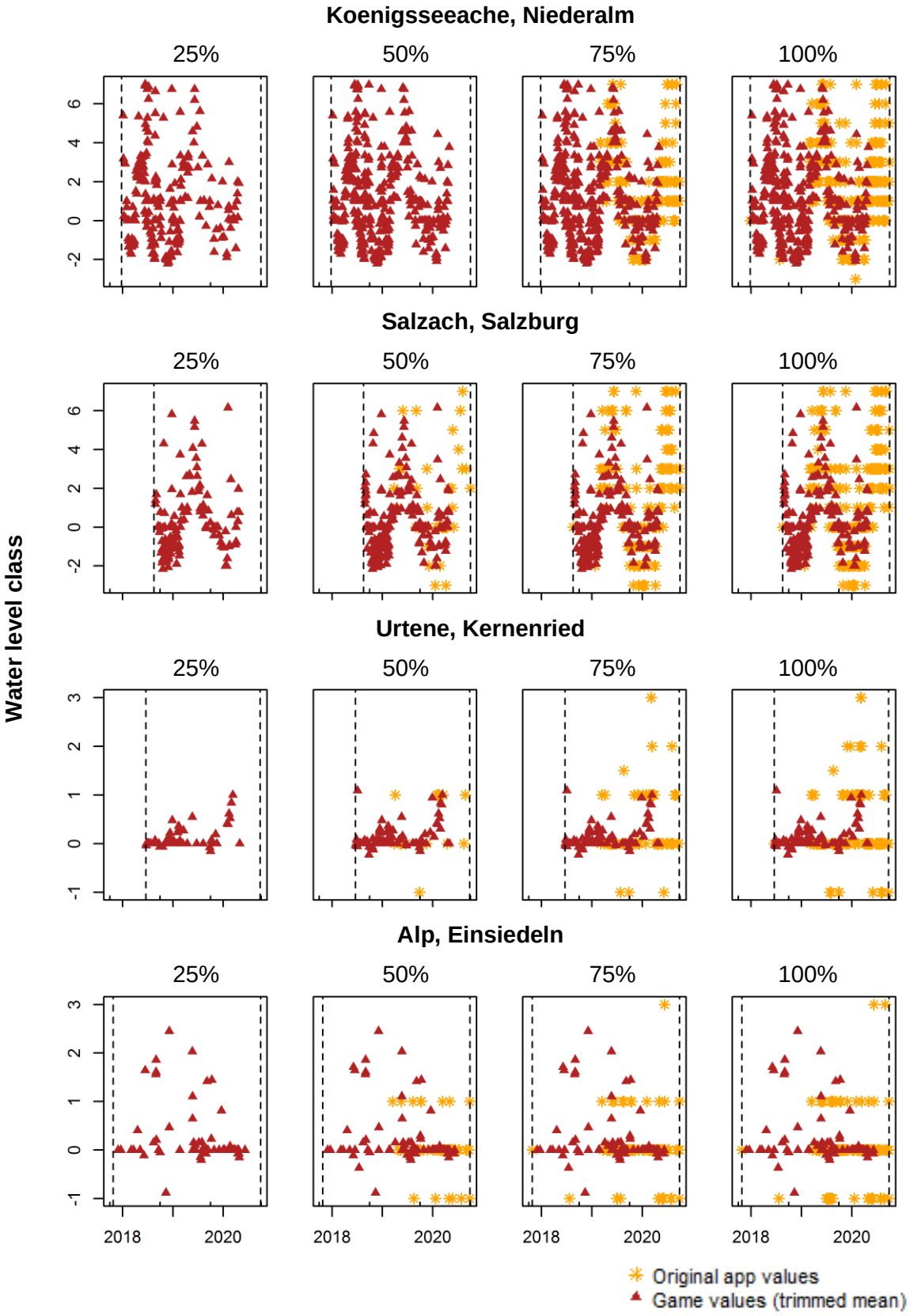


Figure 43: Citizen science data from the CrowdWater app and the CrowdWater game used for the different scenarios. The dashed lines mark the date of the first and the last citizen science observation within the calibration period (October 2017 to September 2020). The regular tick marks indicate the beginning of a calendar year and the small tick marks indicate the beginning of a hydrological year.

Three cases could be observed regarding the impact of the CrowdWater game on the correlation between the citizen science data and the official discharge measurements (Figure 14). These three cases were also represented in the resulting model performances (Figure 45 and Figure 46). At the Koenigsseeache and the Salzach hardly any difference between the use of app data and game data resulted in the correlation as well as in the model performances. At the Koenigsseeache, all relative performances showed high values, even though the room for improvement was rather small in this catchment due to a small span width between the upper and the lower benchmark (see section 5.5). At the Salzach, the relative performances resulted in rather low values when using less than three discharge measurements per year. This could not be improved by using the quality-controlled citizen science data from the CrowdWater game.

At the Urtene, the correlation as well as the model performance dropped if game data instead of app data was used. The Urtene seems to be a CrowdWater spot at which it is easier to determine the water level class directly onsite than by comparing two pictures in the CrowdWater game. Thus, the value of the citizen science data at the Urtene decreased due to the classification of the data in the game. This performance drop was largest when all the citizen science data used originated from the CrowdWater game, i.e., when 25% of all available citizen science observations were used, as well as when citizen science data only and no discharge measurements were considered for calibration.

In the Alp catchment, the contrary was the case. The replacement of the app data with the quality-controlled game data led to an improvement in model performance. The improvement was pronounced strongest in scenarios in which all the citizen science data used could be replaced with game data (scenarios using 25% of the citizen science data points available) and in which not too many discharge measurements (three or less) were available, as well as in scenarios that used citizen science data only for the calibration of the model. At the Alp, the correlation between the water level classes and the discharge measurement time series was much higher when replacing the app data by the game data (Figure 14). Thus, by replacing the app data by the game data, the quality of the data could be improved. However, note that the scenarios showing the strongest improvement were the scenarios leading to very low model performances when using the data from the app and thus offered most room for improvement when using different data. An example is scenario 0-25, where not even the performance of the lower benchmark could be reached when data from the app were used, and the lower benchmark could be outperformed by using classified data from the game (Figure 44).

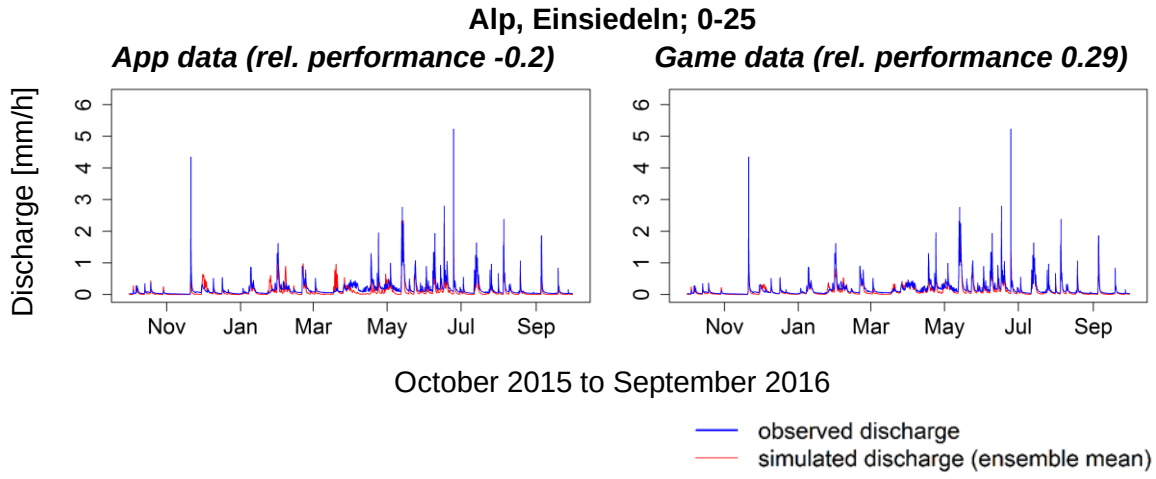


Figure 44: Observed and simulated hydrographs of the Alp for the hydrological year 2016 for scenario 0-25 calibrated using data originating from the CrowdWater app (left) and the higher resolution data from the CrowdWater game. Observed hydrograph shown in blue, simulated ensemble mean hydrograph shown in red.

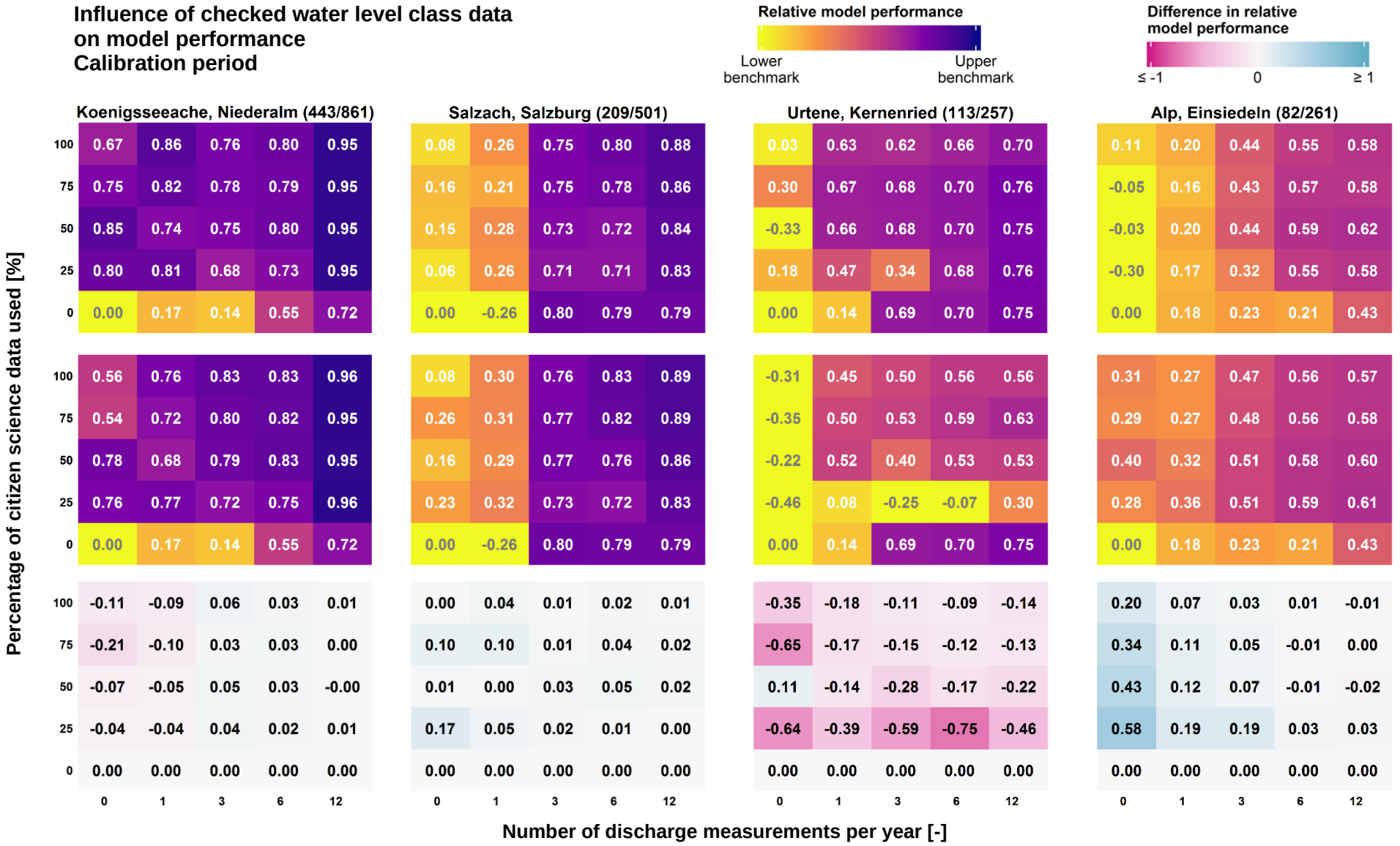


Figure 45: Ensemble mean model performance for four catchments (columns) and all scenarios when using observations obtained by one citizen scientist (first row) and observations checked by at least 15 citizen scientists (second row), and the difference between them (third row). Results for the calibration period. The first number in brackets after the catchment name refers to the number of checked citizen science observations, the second number refers to the total number of citizen science observations.

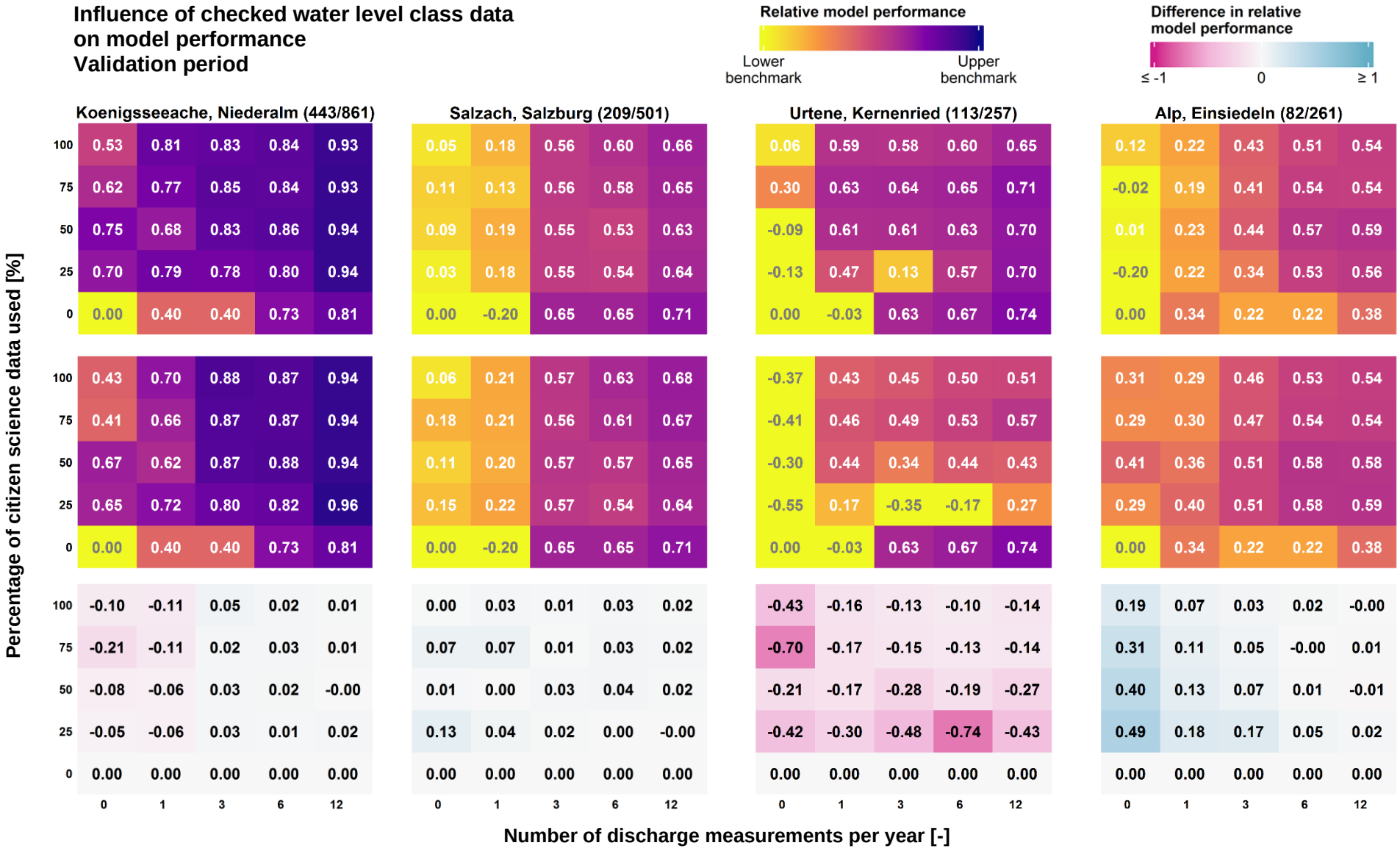


Figure 46: Ensemble mean model performance for four catchments (columns) and all scenarios when using observations obtained by one citizen scientist (first row) and observations checked by at least 15 citizen scientists (second row), and the difference between them (third row). Results for the validation period. The first number in brackets after the catchment name refers to the number of checked citizen science observations, the second number refers to the total number of citizen science observations.

6 Discussion

6.1 Value of only water level class observations

The main research question posed in the beginning of this thesis asked if the calibration of a hydrological model based on citizen science data and a limited number of discharge measurements leads to an accurate simulation of discharge. This can be affirmed if accuracy is seen as an improvement compared to the situation without any information about the amount of water in a stream, i.e., the situation simulated by the lower benchmark in this thesis. However, discharge measurements do in general have more value for the overall model performance than water level observations if both data types are available and used. If the water level class observations are of a high-quality, they contribute to a good model calibration, especially for the simulation of the discharge dynamics.

High-quality observations are required for the calibration with water level class data alone (i.e., all scenarios of the basic approach that do not use any discharge measurements) to be valuable for hydrological model calibration (such as at the Koenigsseeache in Niederalm). But even if the observations are of high quality (such as at the Salzach in Salzburg and the Kempt in Fehraltdorf), water level class data alone may not be sufficient to constrain the model well enough and to reach a good model calibration. The shape of the hydrograph can be simulated well using good water level class data (cf. Figure 24), but the missing information on discharge volume limit the calibration success in other aspects of the simulation (cf. Figure 22 and Figure 23).

A low rank correlation between the water level class observations and the measured discharge resulted in a poor model calibration. Etter et al. (2020a) compared the quality of water level class observations collected with the CrowdWater app to those collected with the forms (see section 4.4.1). Their finding that observations made using the app are of a higher quality than those made using the forms is reflected in the model performance when these data are used for calibrating the model: Data collected with the app helps the model to get the shape of the hydrograph right, even if other components of the hydrograph are not simulated well. Data collected on the forms tended to be disinformative, even for the shape of the hydrograph, i.e., the Spearman rank correlation was lower in these cases than for the lower benchmark.

In the Crowdhydrology project, clearly wrong observations, i.e., observations that are not in a realistic range of the water level, could easily be filtered out (Fienen & Lowry, 2012). Such an approach may also be valuable for the water level classes collected in the CrowdWater project. Thanks to the photos that are uploaded with the observations, experts could possibly check if the data point is indeed an outlier or if it is an extreme event that was captured by a citizen scientist and should therefore be used to calibrate the model. In a real case application, some form of control on the quality of the data collected by citizen scientists should definitely be applied (see also section 6.5 on the discussion of the value of the data from the CrowdWater game).

Previous studies found that the value of citizen science observations is lower for flashier catchments (Davids et al., 2017; Luffman & Connors, 2022). This could not be confirmed here. Possible other factors that influence the value of water level class observations by citizen scientists for model calibration could be hydrological conditions of the catchments which were not investigated here, the length of the time series, or the amount of citizen science data. Additional observations were shown to be valuable in some cases (e.g., for the Urtene in Kernenried and the Wigger in Zofingen), but do not necessarily lead to an increase in model performance (e.g., for the Alp in Einsiedeln and the Kempt in Fehraltdorf). However, it is not possible to say if many more data points would have had an impact because for each catchment only a certain number of water level class observations was

available and thus the amount could not be increased arbitrarily to find out about the value of more data points.

Collectively the findings of the calibration based on only the water level class data imply that water level class observations are a good starting point for the collection of hydrological data in an otherwise ungauged catchment. If one citizen scientist regularly observes the water level of the stream based on a well-set virtual staff gauge, and she or he gets trained, her or his observations will be very valuable to correctly simulate the shape of the hydrograph. Since no clear statement about the required length of the time series of water level class observations can be drawn from this study, it is best to collect time series that are as long as possible, i.e., to start as early as possible with observing the water level class. A similar recommendation can be made regarding the temporal resolution of the observations. In general, more observations seem to be better, but no clear conclusion on the required temporal resolution of the observations can be drawn. However, in their study with synthetic water level class data, Etter et al. (2020b) found that on average one water level class observation per week for a duration of one year is already informative (Etter et al., 2020b). This provides some guidance on the temporal resolution one should strive for, even though it may be difficult to quantify the exact amount of data that is required at a specific site beforehand, as already mentioned by Mazzoleni et al. (2017).

6.2 Value of additional discharge measurements

The findings of this study highlight the value of adding some information on the volume of discharge to the water level class observations in model calibration by doing a few discharge measurements spread over the calibration period. The resulting model performances show that a few discharge measurements have a high value on their own (i.e., without the citizen science observations, cf. the corresponding cells in Figure 17 and Figure 18). This result is in line with the findings of Pool et al. (2017) and Seibert & Beven (2009). However, these previous studies also showed that too few discharge measurements can be disinformative. Aside from the scenario 1-0 (i.e., one discharge measurement and no citizen science observations) for the Sihl, this was not observed in this study. If the quality and number of citizen science observations was rather low, a few discharge measurements per hydrological year always led to better results than using a combination of discharge measurements and water level classes. The low-quality citizen science data acted disinformatively in these cases. Still, the recommendation to use more than one source of data if only limited information is available stated by Avellaneda et al. (2020); Seibert & McDonnell (2015) and Starkey et al. (2017) can be supported by the results of this study: If the citizen science data contain information about the discharge dynamics, or in other words if the citizen science data are of a sufficiently high quality, a combination of the two data types is a promising method for model calibration. If the discharge measurements alone are not able to reach very good model performances, accurate water level observations by citizen scientists can improve the calibration and lead to a better simulation of the observed hydrograph (see the results for the Koenigsseeache in Niederalm and the Alp in Einsiedeln, where the performances with discharge measurements only are lower than those with citizen science data and discharge measurements). Mazzoleni et al. (2017) and Starkey et al. (2017) reached a similar conclusion.

Because a limited number of discharge measurements taken at a regular time step are informative for the calibration of the model, it is recommended to take these measurements where possible. The measurements can be done in regular time intervals but for easily accessible sites, it may be valuable to make an intelligent choice of the sampling days (cf. Pool et al., 2017; Pool & Seibert, 2021; Seibert & Beven, 2009). To avoid the data set being dependent on experts doing the discharge measurements, it may be valuable to train citizen scientists in measuring discharge. The salt dilution method may be a suitable method to do so because citizen scientists can measure discharge quite accurately using this method (Davids et al., 2019) and because it is not required to go into the water to do a measurement

with this method (and the method is thus less dangerous than other methods). Furthermore, if citizen scientists are trained to do discharge measurements, the accessibility of the site is not an issue anymore and the number of discharge measurements per hydrological year may be increased. However, it is not recommended that citizens just estimate the discharge instead of water level classes because Strobl et al. (2020a) showed that these estimates contain large errors and Etter et al. (2018) showed that they cannot be used to reliably calibrate a hydrological model.

6.3 Value of estimating the mean discharge

The different filters investigated in this study to determine the value of an estimate of the mean discharge for model calibration revealed that a more accurate estimate does not necessarily mean better results. Allowing for a deviation of only 2.5% from the mean discharge led to worse (or not better) results than allowing for a deviation of 10% because the negative correlation between a high overall performance (here, the NPE) and the volume error is not perfect: There are parameter sets leading to a very high overall performance but a comparably large volume error (cf. Figure 30 and appendix 10.10). These parameter sets cannot surpass a too narrow filter. The application of a narrow filter thus excludes these parameter sets from being part of the top 100 parameter sets building the ensemble mean, even though they would lead to a good simulation of the hydrograph. A wider filter avoids this, while still filtering out parameter sets with a very large volume error (that do usually also have a rather small overall performance). This is a positive outcome because detailed information on the mean discharge in a stream (as used by Seibert & Vis (2016)) is not available for a data-scarce catchment. Furthermore, a measurement error of only 2.5% is unlikely to be achieved for either precipitation or the discharge measurements (Davids et al., 2019; Horner et al., 2018). Thus, estimating the mean discharge with such a high accuracy is impossible and it does not make sense to apply a filter that only allows for such a small deviation.

The value of a rough estimate of the mean discharge for model calibration was already highlighted by Weeser et al. (2019) for a catchment in Kenya. They applied a simple water balance filter based on precipitation measurements and evapotranspiration estimates based on remote sensing to further constrain a model that was calibrated using water levels observed by participants of their citizen science project. To compensate for errors in their precipitation measurements and in the evapotranspiration estimates derived from the *MODIS* data set as well as for neglected storage changes and other uncertainties, they allowed a deviation of 30% from their calculated water balance for a parameter set to be considered well enough in terms of discharge volume simulation. Their results indicated that better model calibrations can be achieved by applying this filter compared to calibrating the model based on water level data only, i.e., without any volume information (Weeser et al., 2019). The filter allowing for a deviation of 30% tested here (cf. Figure 28 and Figure 29 as well as appendix 10.11) supports this finding: The resulting model performance when using this filter is higher for a majority of the catchments and scenarios than when not using a filter constraining the simulated discharge volume.

In summary, the sub-question if an estimate of the mean annual discharge improves the model performance can clearly be answered with yes. Thereby, it can be stressed that already a rough estimation leads to an improvement. Thus, even if highly uncertain, a mean discharge estimate is indispensable to improve model calibrations based on citizen science data and a limited number of discharge measurements. Such an estimation of the mean discharge could be based on regionalization (Seibert & Vis, 2016) or some simple water balance calculations (Weeser et al., 2019). Thereby, it is recommended to define a comparably wide range of simulated mean discharges that are accepted and not to constrain the model too strongly to an uncertain mean discharge estimate.

6.4 Value of water levels instead of water level classes

The use of water level data instead of water level class data had on average a smaller impact on model performance than the constraint on the mean discharge (see section 6.3). For catchments for which the water level class observations were highly correlated with the measured discharge (i.e., the quality of the water level class observations was very high, such as at the Koenigsseeache in Niederalm and the Salzach in Salzburg), the replacement of the water level class data with water levels only had a very limited effect on the model performance. This result is in line with the findings of van Meerveld et al. (2017), who found that the HBV model can be constrained similarly well when using continuous water level class data (five classes) and higher resolved water level data.

However, for catchments for which the Spearman rank correlation between the citizen science data and the measured discharge time series was rather low (e.g., at the Kleine Emme in Emmen or the Sellenbodenbach in Neuenkirch), model calibrations improved by using water level data instead of water level class data. However, the model performance still did not reach the same level as for catchments with a lot of high-quality citizen science data. Thus, it can be assumed that a larger amount of data (and a higher temporal resolution) is more important than very precise observations when it comes to water level classes. This conclusion can be drawn since the catchments with high-quality citizen science data are also the catchments with a lot of citizen science data. This confirms the finding by Etter et al. (2020b), who showed that a higher temporal resolution of water level class estimates has a larger impact on the model performance than a smaller error in the data (Etter et al., 2020b).

These findings imply that if applied correctly, water level class observations based on a virtual staff gauge are good enough and do not need to be replaced with water level data. This is of advantage since the installation of physical staff gauges may be an obstacle due to cost for material and expertise needed for installation. The sub-question asking if model performances can be improved by using water levels instead of water level classes can thus be negated: Water level class data of a high accuracy are sufficient to constrain the model. Hence, there is no reason to assume that a physical staff gauge that provides the possibility to directly read the water level would have an advantage over the use of the virtual staff gauge, if the virtual staff gauge is set correctly (i.e., at a good location, and in a suitable size). In a real application, the water level class data with water levels could be realized by letting citizen scientists read the water level from a physical staff gauge installed in a stream. The Crowdhydrology project (Lowry & Fienen, 2013) as well as a citizen science project in Kenya (Weeser et al., 2018, 2019) have shown that data collected by citizen scientists this way are quite accurate but cannot assumed to be error-free either.

The results for the Sihl catchment show that there may be catchments for which the use of water level class data for calibration leads to a better overall model performance than the use of water level data. A possible explanation could be small-scale variations in the baseflow caused by human influences, such as the water release from the Sihlsee. These variations are levelled out when water level class data are used (since the water levels still belong into the same class). The model may try to simulate these variations in the water levels but fails to do so because such human influences are not included in the model. As a result, the calibration based on water level data may result in a worse model performance than the calibration based on water level class data. However, this is just one hypothesis for the drop in performance when using the water level data for calibration for the Sihl and requires more investigation.

6.5 Value of water level class data checked by citizen scientists

By playing the CrowdWater game, citizen scientists can increase the value of the water level data collected with the CrowdWater app (Strobl et al., 2019). The influence of the use of these checked water level class data was only investigated for a small number of catchments in this thesis. For the catchment for which there was a gain in data-quality after data quality checks in the game (i.e., the Alp in Einsiedeln), the model performance improved. This was especially the case for the scenarios in which only citizen science data were used to calibrate the model, i.e., the value of the citizen science data for the calibration of the model was increased in the game.

There are also cases in which the use of data from the CrowdWater game leads to a worse model performance than the original data from the CrowdWater app. This was the case for the Urtene and, to a lesser extent and only for some of the scenarios, for the Koenigsseeache. At the Koenigsseeache, the drop in performance may possibly be explained by the higher resolution of the data, which could cause the model to try to simulate small-scale variations that it cannot simulate (cf. the drop in performance for the Sihl when using water level data instead of water level class data (see section 6.4)). At the Urtene, the drop in model performance had to be expected because of the decrease in the correlation between the citizen science data and the discharge (cf. Figure 14). This drop revealed a loss in data-quality caused by the CrowdWater game. A possible explanation could be that the citizen scientist collecting the data had the better overview of the water level variations than it was visible on the uploaded pictures. Thus, the game players were not able to determine the water level classes equally well as the citizen scientist who made the observations on site.

Based on the findings of Strobl et al. (2019), one can assume that the data-quality of the CrowdWater data is increased when the data is running through the CrowdWater game but that this is not the case for about 10% of the data. The Urtene in Kernenried seems to be a location from which the data belongs to these 10%. However, the situation at the Alp with an increase of the data-quality may be considered the normal case as for 75% of the data points, the water level class value resulting from the game was found to be better than the water level class value uploaded to the app (Strobl et al., 2019). These findings imply that the realistic increase in the quality of the citizen science data that can be reached with the CrowdWater game improves the model performance. The last sub-question asking for the impact of checked citizen science data can thus be affirmed; however, some additional checking is still recommended to avoid that data of a lower quality are used due to a drop in quality caused by the CrowdWater game or some other (community- or computer-) based data-quality control mechanism.

6.6 Limitations of this study

6.6.1 Study catchments and model

This thesis focused on eleven catchments that are all located relatively close together in Central Europe. The climate is humid for all catchments. The catchments mainly differ in their area (Figure 5 and Table 1) and discharge regimes (Figure 7). The choice of study catchments was strongly limited by the availability of water level class data collected in the CrowdWater project and official discharge data from a nearby measurement station. Therefore, the results of this thesis are limited to the value of data for similar catchments and cannot directly be transferred to very different catchments, e.g., in arid and semi-arid climates.

All simulations were done using the HBV model. As described in section 4.5, the model fits the purposes of this thesis well. However, statements about the value of the different types of data for hydrological models must be limited to other lumped models.

The calibration process (described in section 0) contained a huge limitation, as it was not possible to calibrate against the water level class and the discharge measurements at the same time. Thus, the two sources of information could not be used in the calibration process simultaneously but only one after the other. As a result, it was also not possible to calibrate the model using the more efficient GAP algorithm for the different data availability scenarios and Monte Carlo simulations had to be used instead. By using Monte Carlo simulations, the parameter space was limited to a random choice of one million parameter sets that may not contain parameter sets that would lead to better results. Furthermore, also the selected parameter ranges (see section 4.6.3) constrain the parameter space.

6.6.2 Data used for calibration

Even though the same questions were asked for all catchments, the different amount and quality of the citizen science data available for each catchment made it difficult to find general answers on the value of water level class data. The water level class data used in this thesis originated from two different approaches: The data collected with the app and the data collected on the forms at the pen and paper stations. These data types were treated equally, even though they differ in quality (Etter et al., 2020a; Figure 12). In this study, it was possible to determine the quality of the citizen science data. However, if only citizen science data are available and need to be used for hydrological model calibration, this would not be possible.

The limited number of discharge measurements were simulated by extracting a certain number of data points from the full discharge time series for each catchment. It was thus assumed that a point measurement of the discharge would match the discharge measured based on the stage-discharge relationship. However, point measurements of the discharge, as well as the calculation of the discharge based on a stage-discharge relationship are subject to uncertainty. This uncertainty was neglected here.

Thanks to the discharge data available for the study catchments, the mean discharge could be calculated precisely. Starting with this mean discharge, the intervals for the filters could be set in both directions, e.g., the 30%-filter could allow for an over- and underestimation of 30% from the exact mean discharge. In a real-world application, the range of the simulated mean discharge that would still be accepted has to be defined without any knowledge about the actual mean discharge. Thus, the mean discharge filter applied may be subject to a larger uncertainty than it was the case here.

The water level data used to simulate citizen science data that are perfectly correlated with the discharge were extracted from the available discharge data. It was thus assumed that the stage-discharge relationship did not change during the four calibration years and that the water level measurements are connected to the discharge in the stream by a bijective function. However, the stage-discharge-relationship can change with time and this effect was neglected here.

Water level class data that has gone through the CrowdWater game was only available for four of the eleven study catchments. The sample size was thus very small and a general conclusion on the impact of the changes in the data when going through the CrowdWater game on the performance of the model can hardly be drawn. Even though data-quality is improved in the CrowdWater game in most cases (Strobl et al., 2019), the impact on the model performance when these data instead of data originating from the app is used to calibrate the model requires more research.

Finally, yet importantly, the meteorological data are subject to uncertainties and may contain errors (e.g., due to the high spatial variation in precipitation). This is also the case for the discharge data obtained by the authorities. The uncertainties are even higher for the periods for which the data has

not been quality-controlled by the authorities yet (see appendix 10.3). However, these data were used as a reference because they represent the best possible estimate for these variables. While it was possible to use high resolution meteorological data for all catchments here, it is unlikely that data of such a high quality are available in all remote and data-scarce areas. The meteorological data that would be used in a real-world application may thus contain larger errors than the meteorological data used here, which would lead to a higher uncertainty in the simulated hydrographs.

6.7 Outlook

To obtain more insight in the value of the data used in this thesis, the limitations in the calibration process must be addressed. This requires an objective function that takes the discharge measurements and the citizen science data simultaneously into account. Aside this objective function, the HBV model needs to be extended with a functionality that allows the use two different kinds of discharge information at the same time, even if these two information types cannot directly be linked to each other. With this extension, the aforementioned objective function could be designed easily for example by combining the NPE and the Spearman rank correlation used here. Other ways of combining the two different data types to calibrate a hydrological model based on them should be investigated too.

To be able to make more general statements on the value of the citizen science data, the analyses done here should be expanded to more catchments covering a large range in catchment characteristics and especially to catchments in different climate zones than the humid climate of Central Europe. As it is difficult to find enough motivated citizen scientists that collect data of a high quality at suitable locations, and as it takes a long time until enough data is gathered, synthetic data could be considered to further investigate on the value of the data (cf. Etter et al., 2020b).

The treatment of the water level class observations in the CrowdWater game can lead to a decrease in data-quality. Thus, some expert checking may be required to see if the game data should be used for model calibration. Since manual checking of the data is laborious and may not be feasible in large applications, image recognition algorithms may be used for this task in the future. By doing so, the collaboration of citizen scientists and machine learning may provide even more valuable information on the amount of water in a stream (cf. Wang et al., in review).

Initially, it was planned to also investigate on the value of the water level class data collected in the CrowdWater project for hydrological regionalization. However, this went beyond the scope of this thesis. Still, the question if water level class data collected by citizen scientists can improve regionalization approaches would be worth another investigation. To conduct such an "informed regionalization approach" (cf. Pool et al., 2019) for a catchment with no discharge information but water level class observations, the parametrizations originating from the donor catchments (i.e., those parameter sets that would be used to build an ensemble mean hydrograph in each donor catchment) could be used to obtain simulated hydrographs for the almost ungauged catchment. Instead of an equal weighting of these hydrographs to obtain an ensemble mean for the almost ungauged catchment, the weighting could be done according to the Spearman rank correlation between the simulated discharge time series and the water level class observations available. Based on the value of high-quality water level class data for hydrological model calibration demonstrated here, it can be assumed that also regionalization approaches can profit from the information content of these observations.

7 Conclusions

This study examined the value of water level class observations by citizen scientists and limited other information for hydrological model calibration. The study found that accurate citizen science observations help to simulate the discharge dynamics in a catchment and that the simulations obtained like this can be improved with a few discharge measurements or an estimate of the mean discharge. Furthermore, the study found that water level class data based on a well-set virtual staff gauge can be about as valuable as exact water level data and that a high temporal resolution of water level observations is more valuable than a high accuracy. It also found that if an improvement of the data-quality is achieved in a quality-control approach by citizen scientists, water level class observations can get even more valuable for the calibration of a hydrological model.

Observations and measurements that do not meet the typical standards of high-quality data in terms of temporal resolution and precision are valuable for the calibration of a hydrological model. The comparably easily achievable combination of water level class observations by citizen scientists and a rough estimate of the mean discharge allows for a calibration that far outperforms the situation without any information about the amount of water in a stream. This finding highlights the value of citizen science in hydrology and is especially encouraging for regions where neither hydrological data nor the means to build and maintain a network of measurement stations is available, but discharge simulations are urgently needed.

This thesis closes with a recommendation for the collection of hydrological data if only limited resources are available. First, it is valuable to collect as much water level class data as possible with the CrowdWater app. To get high quality data, the staff gauge should be placed correctly and in an appropriate size. It is furthermore beneficial if the same person adds the repeated observations to the app. The volume information included in a few discharge measurements can increase the value of the citizen science data and thus discharge measurements should be taken if possible. Furthermore, the mean discharge of the stream should be estimated. This can be a very rough estimate. By excluding simulations that do not simulate the mean discharge within a realistic range, the reliability of the model performance can be increased significantly.

Water level class observations collected in the CrowdWater app by an experienced citizen scientist using a well-set staff gauge have a very similar value for model calibration as water level measurements with the same temporal resolution. Thus, it is not necessary to install a physical staff gauge everywhere where water level information needs to be collected. A well-set virtual staff gauge does the job. However, the installation of a physical staff gauge may increase the quality of citizen science data collected by different people, as citizen scientists contributing for the first time may have more issues with a virtual than with a physical staff gauge.

In general, data-quality of water level classes can successfully be controlled by citizen scientists, as it is done in the CrowdWater game. Before using the water level class data collected in the CrowdWater app, it is recommended to improve the quality of these data by letting at least 15 citizen scientists vote on the water level class of each observation in the CrowdWater game. However, the resulting data should be carefully double-checked by experts to avoid a loss in data-quality. A decrease instead of an increase in data-quality can happen if game players have difficulties to determine the water level class shown on the picture.

The combination of all these data is informative for the calibration of a hydrological model and thus provides a remedy to the lack of data in hydrology.

8 Acknowledgements

I was fortunate to be mentored in this thesis by two brilliant scientists and wonderful personalities whom I admire very much. Thank you, Ilja and Jan, for all the inspiring meetings, for your appreciation and your support. Thank you for always making time for me and letting me ask all my questions. Thank you for your valuable feedback on earlier versions of this thesis. Thank you for sharing your knowledge about and excitement for water and modelling with me. It was a real pleasure to discuss my plans, challenges, and outcomes with you. I am very excited to work with you in the future.

I thank Marc Vis for listening to my minor and major problems, for helping me solve them and for all his explanations that made me understand so many details regarding hydrological modelling. I thank him for speeding up my loops, showing me the advantages of Notepad++ that saved me so much time, for writing me a script to download data much faster than it would be possible manually, for providing me with private GIS lectures, for giving me hints on how to solve coding puzzles and for the many times asking me how me and my thesis are doing.

I thank Simon Meili-Etter and Barbara Strobl for enlightening the CrowdWater flame in me and for their support when I had questions while working on my thesis. I thank Maria Staudinger and Daniel Viviroli for helping me with data issues and Sandra Pool for providing me with her latest research. I thank everyone else of the H2K group for offering me a nice working environment.

The research done in this thesis would never have been possible without the citizen scientists of CrowdWater. Especially the incredible number of contributions by Auria Buchs, Monika Dietschi Hanselmann, Elisabeth Strobl, Karin Ebermann and Martin Ringer made this thesis possible. I am very grateful that I was able to work with data collected by motivated people. Many thanks to all of them. I furthermore thank Hanspeter Hodel for sending me the pen and paper data collected at several locations in Switzerland and for maintaining the stations over all these years.

I thank Michèle Oberhänsli, Hélène Monnard, Noël Leber and Nik Jauer from the Federal Office for the Environment, Michael Solomir from the Canton of Zurich, Simon Jaun and Martin Liniger from the Canton of Berne, Udo Ebner from the Hydrological Service Salzburg, Klaus Häfliger from the ZAMG and the people behind IDAWEB of MeteoSwiss for their help with getting the meteorological and hydrological data required to write this thesis.

With all my heart, I thank my parents Natalie and Thomas for supporting me during all these years of education, for enabling me to go my way and for always being there for me. I thank Noël for his love, support and understanding, for being an oasis when numbers were turning in my head and for reminding me to take breaks. I thank my sisters Flavia and Olivia and my grandfather Karl for being so proud of me and for their interest in what I am doing. And I thank Sophia for our wonderful friendship and for listening to me when I was wondering about the discharge of the Danube in Budapest.

9 Bibliography

Aceves-Bueno, E., Adeleye, A. S., Feraud, M., Huang, Y., Tao, M., Yang, Y., & Anderson, S. E. (2017). The Accuracy of Citizen Science Data: A Quantitative Review. *Bulletin of the Ecological Society of America*, *98*(4), 278–290. https://doi.org/10.1002/bes2.1336

Addor, N., Jaun, S., Fundel, F., & Zappa, M. (2011). An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): skill, case studies and scenarios. *Hydrology and Earth System Sciences*, *15*(7), 2327–2347. https://doi.org/10.5194/hess-15-2327-2011

Assumpção, T. H., Popescu, I., Jonoski, A., & Solomatine, D. P. (2018). Citizen observations contributing to flood modelling: opportunities and challenges. *Hydrology and Earth System Sciences*, *22*(2), 1473–1489. https://doi.org/10.5194/hess-22-1473-2018

Avellaneda, P. M., Ficklin, D. L., Lowry, C. S., Knouft, J. H., & Hall, D. M. (2020). Improving Hydrological Models With the Assimilation of Crowdsourced Data. *Water Resources Research*, *56*(5), e2019WR026325. https://doi.org/10.1029/2019WR026325

Baker, D. B., Richards, R. P., Loftus, T. T., & Kramer, J. W. (2004). A new flashiness index: Characteristics and applications to Midwestern rivers and streams. *Journal of the American Water Resources Association*, *40*(2), 503–522. https://doi.org/10.1111/j.1752-1688.2004.tb01046.x

Bergström, S. (1991). Principles and confidence in hydrological modelling. *Nordic Hydrology*, *22*(2), 123–136. https://doi.org/10.2166/nh.1991.0009

Bergström, S. (1992). The HBV Model – its structure and applications. *SMHI Reports Hydrology* (No. 4). 32 pages. Norrköping, Sweden.

Bergström, S. (1995). The HBV model. In V. P. Singh (Ed.), *Computer Models of Watershed Hydrology* (pp. 443–476). Highlands Ranch, Colorado, USA: Water Resources Publications.

Beven, K. J. (2012). *Rainfall-Runoff Modelling: The Primer, Second Edition*. 488 pages. Chichester, UK: Wiley-Blackwell. https://doi.org/10.1002/9781119951001

Brath, A., Montanari, A., & Toth, E. (2004). Analysis of the effects of different scenarios of historical data availability on the calibration of a spatially-distributed hydrological model. *Journal of Hydrology*, *291*(3–4), 232–253. https://doi.org/10.1016/j.jhydrol.2003.12.044

Brunner, M. I., Slater, L., Tallaksen, L. M., & Clark, M. (2021). Challenges in modeling and predicting floods and droughts: A review. *Wiley Interdisciplinary Reviews: Water*, *8*(3), e1520. https://doi.org/10.1002/wat2.1520

Buytaert, W., Zulkafli, Z., Grainger, S., Acosta, L., Alemie, T. C., Bastiansen, J., et al. (2014). Citizen science in hydrology and water resources: opportunities for knowledge generation, ecosystem service management, and sustainable development. *Frontiers in Earth Science*, *2*(26), 1–21. https://doi.org/10.3389/feart.2014.00026

Correa, A., Windhorst, D., Crespo, P., Célleri, R., Feyen, J., & Breuer, L. (2016). Continuous versus event-based sampling: how many samples are required for deriving general hydrological understanding on Ecuador's páramo region? *Hydrological Processes*, *30*(22), 4059–4073. https://doi.org/10.1002/hyp.10975

Le Coz, J., Renard, B., Bonnifait, L., Branger, F., & Le Boursicaud, R. (2014). Combining hydraulic knowledge and uncertain gaugings in the estimation of hydrometric rating curves: A Bayesian approach. *Journal of Hydrology*, *509*, 573–587. https://doi.org/10.1016/j.jhydrol.2013.11.016

Le Coz, J., Patalano, A., Collins, D., Guillén, N. F., García, C. M., Smart, G. M., et al. (2016). Crowdsourced data for flood hydrology: Feedback from recent citizen science projects in Argentina, France and New Zealand. *Journal of Hydrology*, *541*, *Part*, 766–777. https://doi.org/10.1016/j.jhydrol.2016.07.036

Davids, J. C., van de Giesen, N., & Rutten, M. (2017). Continuity vs. the Crowd—Tradeoffs Between Continuous and Intermittent Citizen Hydrology Streamflow Observations. *Environmental Management*, *60*(1), 12–29. https://doi.org/10.1007/s00267-017-0872-x

Davids, J. C., Rutten, M. M., Pandey, A., Devkota, N., David Van Oyen, W., Prajapati, R., & Van De Giesen, N. (2019). Citizen science flow - an assessment of simple streamflow measurement methods. *Hydrology and Earth System Sciences*, *23*(2), 1045–1065. https://doi.org/10.5194/hess-23-1045-2019

Drogue, G. P., & Plasse, J. (2014). How can a few streamflow measurements help to predict daily hydrographs at almost ungauged sites? *Hydrological Sciences Journal*, *59*(12), 2126–2142. https://doi.org/10.1080/02626667.2013.865031

Dwarakish, G. S., & Ganasri, B. P. (2015). Impact of land use change on hydrological systems: A review of current modeling approaches. *Cogent Geoscience*, *1*(1), 1115691. https://doi.org/10.1080/23312041.2015.1115691

Eitzel, M. V, Cappadonna, J. L., Santos-Lang, C., Duerr, R. E., Virapongse, A., West, S. E., et al. (2017). Citizen Science Terminology Matters: Exploring Key Terms. *Citizen Science: Theory and Practice*, *2*(1), 1–20. https://doi.org/10.5334/cstp.96

Elmi, O., Tourian, M. J., & Sneeuw, N. (2015). River discharge estimation using channel width from satellite imagery. In *International Geoscience and Remote Sensing Symposium (IGARSS)* (pp. 727–730). IEEE. https://doi.org/10.1109/IGARSS.2015.7325867

Eng, K., & Milly, P. C. D. (2007). Relating low-flow characteristics to the base flow recession time constant at partial record stream gauges. *Water Resources Research*, *43*(1), W01201. https://doi.org/10.1029/2006WR005293

Engeland, K., Hisdal, H., & Frigessi, A. (2004). Practical Extreme Value Modelling of Hydrological Floods and Droughts: A Case Study. *Extremes*, *7*(1), 5–30. https://doi.org/10.1007/s10687-004-4727-5

Etter, S. (2020). *CrowdWater: Motivations of Citizen Scientists, the Accuracy and the Potential of Crowd-Based Data for Hydrological Model Calibration*. 267 pages. Doctoral dissertation. University of Zurich, Faculty of Science. https://doi.org/10.5167/uzh-188314

Etter, S., Strobl, B., Seibert, J., & van Meerveld, H. J. I. (2018). Value of uncertain streamflow observations for hydrological modelling. *Hydrology and Earth System Sciences*, *22*(10), 5243–5257. https://doi.org/10.5194/hess-22-5243-2018

Etter, S., Strobl, B., van Meerveld, H. J. I., & Seibert, J. (2020a). Quality and timing of crowd-based water level class observations. *Hydrological Processes*, *34*(22), 4365–4378. https://doi.org/10.1002/hyp.13864

Etter, S., Strobl, B., Seibert, J., & van Meerveld, H. J. I. (2020b). Value of Crowd-Based Water Level Class Observations for Hydrological Model Calibration. *Water Resources Research*, *56*(2), e2019WR026108. https://doi.org/10.1029/2019WR026108

Fasipe, O. A., Izinyon, O. C., & Ehiorobo, J. O. (2021). Hydropower potential assessment using spatial technology and hydrological modelling in Nigeria river basin. *Renewable Energy*, *178*, 960–976. https://doi.org/10.1016/j.renene.2021.06.133

Fienen, M. N., & Lowry, C. S. (2012). Social.Water-A crowdsourcing tool for environmental data acquisition. *Computers and Geosciences*, *49*, 164–169. https://doi.org/10.1016/j.cageo.2012.06.015

Fung, K. F., Huang, Y. F., Koo, C. H., & Soh, Y. W. (2020). Drought forecasting: A review of modelling approaches 2007–2017. *Journal of Water and Climate Change*, *11*(3), 771–799. https://doi.org/10.2166/wcc.2019.236

Garnier, S., Ross, N., Rudis, R., Camargo, P. A., Sciaini, M., & Scherer, C. (2021). viridis - Colorblind-Friendly Color Maps for R. R package version 0.6.2. https://doi.org/10.5281/zenodo.4679424

Getirana, A. C. V., Bonnet, M. P., Calmant, S., Roux, E., Rotunno Filho, O. C., & Mansur, W. J. (2009). Hydrological monitoring of poorly gauged basins based on rainfall-runoff modeling and spatial altimetry. *Journal of Hydrology*, *379*(3–4), 205–219. https://doi.org/10.1016/j.jhydrol.2009.09.049

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, *32*(18), 2847–2849. https://doi.org/10.1093/bioinformatics/btw313

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*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Gustard, A., Bullock, A., & Dixon, J. M. (1992). Low flow estimation in the United Kingdom. *Report - UK Institute of Hydrology* (Vol. 108). 292 pages. Oxfordshire, UK.

Haberlandt, U. (2010). From hydrological modelling to decision support. *Advances in Geosciences*, *27*, 11–19. https://doi.org/10.5194/adgeo-27-11-2010

Haiden, T., Kann, A., Wittmann, C., Pistotnik, G., Bica, B., & Gruber, C. (2011). The integrated nowcasting through comprehensive analysis (INCA) system and its validation over the Eastern Alpine region. *Weather and Forecasting*, *26*(2), 166–183. https://doi.org/10.1175/2010WAF2222451.1

Hannah, D. M., Demuth, S., van Lanen, H. A. J., Looser, U., Prudhomme, C., Rees, G., et al. (2011). Large-scale river flow archives: Importance, current status and future needs. *Hydrological Processes*, *25*(7), 1191–1200. https://doi.org/10.1002/hyp.7794

Harlin, J. (1991). Development of a process oriented calibration scheme for the HBV hydrological model. *Nordic Hydrology*, *22*(1), 15–36. https://doi.org/10.2166/nh.1991.0002

Harrison, R. L. (2009). Introduction to Monte Carlo simulation. *AIP Conference Proceedings*, *1204*, 17–21. https://doi.org/10.1063/1.3295638

Haslinger, K., & Bartsch, A. (2016). Creating long-term gridded fields of reference evapotranspiration in Alpine terrain based on a recalibrated Hargreaves method. *Hydrology and Earth System Sciences*, *20*(3), 1211–1223. https://doi.org/10.5194/hess-20-1211-2016

Hiebl, J., & Frei, C. (2016). Daily temperature grids for Austria since 1961—concept, creation and applicability. *Theoretical and Applied Climatology*, *124*(1), 161–178. https://doi.org/10.1007/s00704-015-1411-4

Horner, I., Renard, B., Le Coz, J., Branger, F., McMillan, H. K., & Pierrefeu, G. (2018). Impact of Stage Measurement Errors on Streamflow Uncertainty. *Water Resources Research*, *54*(3), 1952–1976. https://doi.org/10.1002/2017WR022039

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

Jian, J., Ryu, D., Costelloe, J. F., & Su, C. H. (2017). Towards hydrological model calibration using river level measurements. *Journal of Hydrology: Regional Studies*, *10*, 95–109. https://doi.org/10.1016/j.ejrh.2016.12.085

Juston, J., Seibert, J., & Johansson, P. (2009). Temporal sampling strategies and uncertainty in calibrating a conceptual hydrological model for a small boreal catchment. *Hydrological Processes*, *23*(21), 3093–3109. https://doi.org/10.1002/hyp.7421

Kauzlaric, M., Nicolet, G., & Viviroli, D. (2021). EXAR: Entwicklung hydrometeorologischer Grundlagen. In N. Andres, N. Steeb, A. Badoux, & C. Hegg (Eds.), *Extremhochwasser an der Aare, Hauptbericht Projekt EXAR. Methodik und Resultate.* (pp. 29–37). WSL Berichte.

Kim, U., & Kaluarachchi, J. J. (2009). Hydrologic model calibration using discontinuous data: an example from the upper Blue Nile River Basin of Ethiopia. *Hydrological Processes*, *23*(26), 3705–3717. https://doi.org/10.1002/hyp.7465

Kirkwood, C. W. (1982). A Case History of Nuclear Power Plant Site Selection. *Journal of the Operational Research Society*, *33*(4), 353–363. https://doi.org/10.1057/jors.1982.77

Kundzewicz, Z. W. (1997). Water resources for sustainable development. *Hydrological Sciences Journal*, *42*(4), 467–480. https://doi.org/10.1080/02626669709492047

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. https://doi.org/10.1016/S0022-1694(97)00041-3

Lowry, C. S., & Fienen, M. N. (2013). CrowdHydrology: Crowdsourcing hydrologic data and engaging citizen scientists. *Groundwater*, *51*(1), 151–156. https://doi.org/10.1111/j.1745-6584.2012.00956.x

Lowry, C. S., Fienen, M. N., Hall, D. M., & Stepenuck, K. F. (2019). Growing Pains of Crowdsourced Stream Stage Monitoring Using Mobile Phones: The Development of CrowdHydrology. *Frontiers in Earth Science*, *7*(128), 1–10. https://doi.org/10.3389/feart.2019.00128

Luffman, I., & Connors, D. (2022). Stream Stage Monitoring with Community Science-Contributed Stage Data. *Hydrology*, *9*(1), 11. https://doi.org/10.3390/hydrology9010011

Mazzoleni, M., Verlaan, M., Alfonso, L., Monego, M., Norbiato, D., Ferri, M., & Solomatine, D. P. (2017). Can assimilation of crowdsourced data in hydrological modelling improve flood prediction? *Hydrology and Earth System Sciences*, *21*(2), 839–861. https://doi.org/10.5194/hess-21-839-2017

McMillan, H., Freer, J., Pappenberger, F., Krueger, T., & Clark, M. (2010). Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions. *Hydrological Processes*, *24*(10), 1270–1284. https://doi.org/10.1002/hyp.7587

van Meerveld, H. J. I., Vis, M. J. P., & Seibert, J. (2017). Information content of stream level class data for hydrological model calibration. *Hydrology and Earth System Sciences*, *21*(9), 4895–4905. https://doi.org/10.5194/hess-21-4895-2017

Melsen, L. A., Teuling, A. J., van Berkum, S. W., Torfs, P. J. J. F., & Uijlenhoet, R. (2014). Catchments as simple dynamical systems: A case study on methods and data requirements for parameter identification. *Water Resources Research*, *50*(7), 5577–5596. https://doi.org/10.1002/2013WR014720

Menzel, L., Lang, H., & Rohmann, M. (1999). Mean Annual Actual Evaporation 1973-1992. *Hydrological Atlas of Switzerland, Plate 4.1*.

Merz, R., & Blöschl, G. (2005). Flood frequency regionalisation—spatial proximity vs . catchment attributes. *Journal of Hydrology*, *302*(1–4), 283–306. https://doi.org/10.1016/j.jhydrol.2004.07.018

Mishra, A. K., & Coulibaly, P. (2009). Developments in hydrometric network design: A review. *Reviews of Geophysics*, *47*(2), RG2001. https://doi.org/doi:10.1029/ 2007RG000243

Neuwirth, E. (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1-2.

OED (Oxford English Dictionary) (2021). 'citizen science'. Retrieved 17 November 2021, from https://www.oed.com/view/Entry/33513?redirectedFrom=citizen+science%23eid316619123

de Oliveira Serrão, E. A., Silva, M. T., Ferreira, T. R., de Ataide, L. C. P., Wanzeler, R. T. S., da Silva, V. de P. R., et al. (2021). Large-Scale hydrological modelling of flow and hydropower production, in a Brazilian watershed. *Ecohydrology and Hydrobiology*, *21*(1), 23–35. https://doi.org/10.1016/j.ecohyd.2020.09.002

Perrin, C., Oudin, L., Andreassian, V., Michel, C., & Mathevet, T. (2007). Impact of limited streamflow data on the efficiency and the parameters of rainfall—runoff models. *Hydrological Sciences Journal*, *52*(1), 131–151. https://doi.org/10.1623/hysj.52.1.131

Pool, S., & Seibert, J. (2021). Gauging ungauged catchments – Active learning for the timing of point discharge observations in combination with continuous water level measurements. *Journal of Hydrology*, *598*, 126448. https://doi.org/10.1016/j.jhydrol.2021.126448

Pool, S., Viviroli, D., & Seibert, J. (2017). Prediction of hydrographs and flow-duration curves in almost ungauged catchments: Which runoff measurements are most informative for model calibration? *Journal of Hydrology*, *554*, 613–622. https://doi.org/10.1016/j.jhydrol.2017.09.037

Pool, S., Vis, M., & Seibert, J. (2018). Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency. *Hydrological Sciences Journal*, *63*(13–14), 1941–1953. https://doi.org/10.1080/02626667.2018.1552002

Pool, S., Viviroli, D., & Seibert, J. (2019). Value of a Limited Number of Discharge Observations for Improving Regionalization: A Large-Sample Study Across the United States. *Water Resources Research*, *55*(1), 363–377. https://doi.org/10.1029/2018WR023855

Rojas-Serna, C., Michel, C., Perrin, C., & Andreassian, V. (2006). Ungauged catchments: How to make the most of a few streamflow measurements? In *Large Sample Basin Experiments for Hydrological Model Parametrization: Results of the Model Parameter Experiment MOPEX* (pp. 230–236). IAHS-AISH Publication 307.

Rojas-Serna, C., Lebecherel, L., Perrin, C., Andréassian, V., & Oudin, L. (2016). How should a rainfall-runoff model be parameterized in an almost ungauged catchment? A methodology tested on 609 catchments. *Water Resources Research*, *52*(6), 4765–4784. https://doi.org/10.1111/j.1752-1688.1969.tb04897.x

Ruhi, A., Messager, M. L., & Olden, J. D. (2018). Tracking the pulse of the Earth's fresh waters. *Nature Sustainability*, *1*(4), 198–203. https://doi.org/10.1038/s41893-018-0047-7

Savic, D. A., Kapelan, Z. S., & Jonkergouw, P. M. R. (2009). Quo vadis water distribution model calibration? *Urban Water Journal*, *6*(1), 3–22. https://doi.org/10.1080/15730620802613380

Schaefli, B., & Gupta, H. V. (2007). Do Nash values have value? *Hydrological Processes*, *21*(15), 2075–2080. https://doi.org/10.1002/hyp.6825

Seibert, J. (1999). Regionalisation of parameters for a conceptual rainfall-runoff model. *Agricultural and Forest Meteorology*, *98–99*, 279–293. https://doi.org/10.1016/S0168-1923(99)00105-7

Seibert, J. (2000). Multi-criteria calibration of a conceptual runoff model using a genetic algorithm. *Hydrology and Earth System Sciences*, *4*(2), 215–224. https://doi.org/10.5194/hess-4-215-2000

Seibert, J. (2001). On the need for benchmarks in hydrological modelling. *Hydrological Processes*, *15*(6), 1063–1064. https://doi.org/10.1002/hyp.446

Seibert, J., & Bergström, S. (2022). A retrospective on hydrological modelling based on half a century with the HBV model. *Hydrology and Earth System Sciences*, *26*(5), 1371–1388. https://doi.org/10.5194/hess-26-1371-2022

Seibert, J., & Beven, K. J. (2009). Gauging the ungauged basin: how many discharge measurements are needed? *Hydrology and Earth System Sciences*, *13*(6), 883–892. https://doi.org/10.5194/hess-13-883-2009

Seibert, J., & McDonnell, J. J. (2015). Gauging the Ungauged Basin: Relative Value of Soft and Hard Data. *Journal of Hydrologic Engineering*, *20*(1), A4014004-1–6. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000861

Seibert, J., & Vis, M. J. P. (2012). Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. *Hydrology and Earth System Sciences*, *16*(9), 3315–3325. https://doi.org/10.5194/hess-16-3315-2012

Seibert, J., & Vis, M. J. P. (2016). How informative are stream level observations in different geographic regions? *Hydrological Processes*, *30*(14), 2498–2508. https://doi.org/10.1002/hyp.10887

Seibert, J., Vis, M. J. P., Lewis, E., & van Meerveld, H. J. I. (2018). Upper and lower benchmarks in hydrological modelling. *Hydrological Processes*, *32*(8), 1120–1125. https://doi.org/10.1002/hyp.11476

Seibert, J., Strobl, B., Etter, S., Hummer, P., & van Meerveld, H. J. I. (2019). Virtual staff gauges for crowd-based stream level observations. *Frontiers in Earth Science*, *7*(70), 1–10. https://doi.org/10.3389/feart.2019.00070

Sideris, I. V., Gabella, M., Erdin, R., & Germann, U. (2014). Real-time radar-rain-gauge merging using spatio-temporal co-kriging with external drift in the alpine terrain of Switzerland. *Quarterly Journal of the Royal Meteorological Society*, *140*(680), 1097–1111. https://doi.org/10.1002/qj.2188

Silvertown, J. (2009). A new dawn for citizen science. *Trends in Ecology and Evolution*, *24*(9), 467–471. https://doi.org/10.1016/j.tree.2009.03.017

Singh, S. K., & Bárdossy, A. (2012). Calibration of hydrological models on hydrologically unusual events. *Advances in Water Resources*, *38*, 81–91. https://doi.org/10.1016/j.advwatres.2011.12.006

Sivapalan, M. (2003). Prediction in ungauged basins: a grand challenge for theoretical hydrology. *Hydrological Processes*, *17*(15), 3163–3170. https://doi.org/10.1002/hyp.5155

Solomatine, D. P., & Wagener, T. (2011). Hydrological Modeling. In P. Wilderer (Ed.), *Treatise on Water Science* (Vol. 2, pp. 435–457). https://doi.org/10.1016/B978-0-444-53199-5.00044-0

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, *15*(1), 72–101. https://doi.org/doi:10.2307/1412159

Starkey, E., Parkin, G., Birkinshaw, S., Large, A., Quinn, P., & Gibson, C. (2017). Demonstrating the value of community-based ('citizen science') observations for catchment modelling and characterisation. *Journal of Hydrology*, *548*, 801–817. https://doi.org/10.1016/j.jhydrol.2017.03.019

Strobl, B. (2020). *Quality of crowdsourced water level observations*. 150 pages. Doctoral dissertation. University of Zurich, Faculty of Science. https://doi.org/10.5167/uzh-190608

Strobl, B., Etter, S., van Meerveld, H. J. I., & Seibert, J. (2019). The CrowdWater game: A playful way to improve the accuracy of crowdsourced water level class data. *PLoS ONE*, *14*(9), e0222579. https://doi.org/10.1371/journal.pone.0222579

Strobl, B., Etter, S., van Meerveld, H. J. I., & Seibert, J. (2020a). Accuracy of crowdsourced streamflow and stream level class estimates. *Hydrological Sciences Journal*, *65*(5), 823–841. https://doi.org/10.1080/02626667.2019.1578966

Strobl, B., Etter, S., van Meerveld, H. J. I., & Seibert, J. (2020b). Training citizen scientists through an online game developed for data quality control. *Geoscience Communication*, *3*(1), 109–126. https://doi.org/10.5194/gc-3-109-2020

Sun, W., Wang, Y., Wang, G., Cui, X., Yu, J., Zuo, D., & Xu, Z. (2017). Physically based distributed hydrological model calibration based on a short period of streamflow data: Case studies in four Chinese basins. *Hydrology and Earth System Sciences*, *21*(1), 251–265. https://doi.org/10.5194/hess-21-251-2017

Surowiecki, J. (2004). *The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. 336 pages. London, UK: Little Brown.

Tada, T., & Beven, K. J. (2012). Hydrological model calibration using a short period of observations. *Hydrological Processes*, *26*(6), 883–892. https://doi.org/10.1002/hyp.8302

Viviroli, D., & Seibert, J. (2015). Can a regionalized model parameterisation be improved with a limited number of runoff measurements? *Journal of Hydrology*, *529*, *Part*, 49–61. https://doi.org/10.1016/j.jhydrol.2015.07.009

Vogel, R. M., & Fennessey, N. M. (1995). Flow Duration Curves II: A Review of Applications in Water Resources Planning. *Water Resources Bulletin*, *31*(6), 1029–1039. https://doi.org/10.1111/j.1752-1688.1995.tb03419.x

Walker, D., Forsythe, N., Parkin, G., & Gowing, J. (2016). Filling the observational void: Scientific value and quantitative validation of hydrometeorological data from a community-based monitoring programme. *Journal of Hydrology*, *538*, 713–725. https://doi.org/10.1016/j.jhydrol.2016.04.062

Wang, Z., Seibert, J., van Meerveld, H. J. I., Lyu, H., & Zhang, C. (in review). Automatic water level estimation from repeated crowd-based images of streams. *Water Resources Research*, 2022WR032108.

Weeser, B., Stenfert Kroese, J., Jacobs, S. R., Njue, N., Kemboi, Z., Ran, A., et al. (2018). Citizen science pioneers in Kenya – A crowdsourced approach for hydrological monitoring. *Science of the Total Environment*, *631–632*, 1590–1599. https://doi.org/10.1016/j.scitotenv.2018.03.130

Weeser, B., Jacobs, S., Kraft, P., Rufino, M. C., & Breuer, L. (2019). Rainfall-Runoff Modeling Using Crowdsourced Water Level Data. *Water Resources Research*, *55*(12), 10856–10871. https://doi.org/10.1029/2019WR025248

Weeser, B., Gräf, J., Njue, N. K., Cerutti, P., Rufino, M. C., Breuer, L., & Jacobs, S. R. (2021). Crowdsourced Water Level Monitoring in Kenya's Sondu-Miriu Basin—Who Is 'The Crowd'? *Frontiers in Earth Science*, *8*(602422), 1–13. https://doi.org/10.3389/feart.2020.602422

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., et al. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Z'graggen, L., & Ohmura, A. (2000). Spatio-Temporal Variations in Net Radiation 1984-1993. *Hydrological Atlas of Switzerland, Plate 4.2*.

Zambrano-Bigiarini, M. (2020). hydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series. R package version 0.4-0. https://doi.org/10.5281/zenodo.839854

10 Appendix

10.1 Flashiness and baseflow indices

Table 9: Richards-Baker flashiness index and baseflow index for all study catchments, calculated on daily discharge values from the years 2013-2021.

Catchment	R-B index	BFI
Koenigsseeache, Niederalm	0.406	0.487
Salzach, Salzburg	0.191	0.734
Kempt, Fehraltdorf	0.474	0.478
Urtene, Kernenried	0.253	0.734
Alp, Einsiedeln	0.654	0.354
Kleine Emme, Werthenstein	0.485	0.443
Ova dal Fuorn, Zernez	0.092	0.851
Kleine Emme, Emmen	0.475	0.457
Wigger, Zofingen	0.310	0.626
Sellenbodenbach, Neuenkirch	0.744	0.280
Sihl, Zurich	0.448	0.540

10.2 Temperature measurement stations

Table 10: Full names and locations of the temperature measurement stations in Switzerland of which the data was used.

Short name	Place name, canton	Elevation (m a.s.l.)	Coordinates
BAN	Bantiger, BE	1097	46.978°N / 7.529°E
BER	Bern / Zollikofen, BE	555	46.991°N / 7.464°E
BUF	Buffalora, GR	1973	46.648°N / 10.267°E
EGO	Egolzwil, LU	523	47.179°N / 8.005°E
EIN	Einsiedeln, SZ	912	47.133°N / 8.757°E
FLU	Fluehli, LU	942	46.889°N / 8.020°E
HOE	Hoernli, ZH	1134	47.371°N / 8.942°E
KOP	Koppigen, BE	486	47.119°N / 7.605°E
LUZ	Luzern, LU	456	47.036°N / 8.301°E
NAP	Napf, BE	1406	47.005°N / 7.940°E
PIL	Pilatus, OW	2107	46.979°N / 8.252°E
SAG	Sattel, SZ	792	47.081°N / 8.637°E
SMA	Zurich / Fluntern, ZH	558	47.378°N / 8.566°E
SPF	Schuepfheim, LU	746	46.947°N / 8.012°E
UEB	Uetliberg, ZH	1016	47.351°N / 8.490°E
WAE	Waedenswil, ZH	488	47.221°N / 8.678°E

Table 11: List of temperature stations used per study catchment in Switzerland, with weight according to percentage of area in Thiessen polygon.

Catchment	Temperature stations and corresponding weights
Kempt, Fehraltdorf	HOE: 0.929 / SMA: 0.071
Urtene, Kernenried	BER: 0.64 / BAN: 0.287 / KOP: 0.073
Alp, Einsiedeln	EIN: 0.506 / SAG: 0.494
Kleine Emme, Werthenstein	SPF: 0.435 / FLU: 0.399 / NAP: 0.165
Ova dal Fuorn, Zernez	BUF: 1
Kleine Emme, Emmen	SPF: 0.31 / FLU: 0.254 / PIL: 0.209 / NAP: 0.149 / LUZ: 0.079
Wigger, Zofingen	EGO: 0.749 / NAP: 0.251
Sellenbodenbach, Neuenkirch	LUZ: 1
Sihl, Zurich	EIN: 0.592 / WAE: 0.163 / SAG: 0.153 / UEB: 0.092

10.3 Details on discharge data

Table 12: Detailed information on discharge data time series obtained by the Swiss and Austrian authorities.

Catchment	Official station number	More information	Data provider	Validated until
Koenigsseeache, Niederalm	204230	not available	Hydrographic Service Salzburg	31.12.2019
Salzach, Salzburg	204180	salzburg.gv.at/wasser/hydro/#/Fliessgewässer?station=204180	Hydrographic Service Salzburg	31.12.2019
Kempt, Fehraltdorf	580	zh.ch/de/umwelt-tiere/wasser-gewaesser/messdaten/abfluss-wasserstand.html	Canton of Zurich	31.10.2019
Urtene, Kernenried	A042	wada.sites.be.ch/geoportal/hydromn/oberflaechengewaesser/stationsblatt/A042.pdf	Canton of Berne	31.12.2020
Alp, Einsiedeln	2609	hydrodaten.admin.ch/de/2609.html	Swiss Federal Office for the Environment	31.12.2018
Kleine Emme, Werthenstein	2487	hydrodaten.admin.ch/de/2487.html	Swiss Federal Office for the Environment	31.12.2018
Ova dal Fuorn, Zernez	2304	hydrodaten.admin.ch/de/2304.html	Swiss Federal Office for the Environment	31.12.2018
Kleine Emme, Emmen	2634	hydrodaten.admin.ch/de/2634.html	Swiss Federal Office for the Environment	31.12.2018
Wigger, Zofingen	2450	hydrodaten.admin.ch/de/2450.html	Swiss Federal Office for the Environment	31.12.2018
Sellenbodenbach, Neuenkirch	2608	hydrodaten.admin.ch/de/2608.html	Swiss Federal Office for the Environment	31.12.2018
Sihl, Zurich	2176	hydrodaten.admin.ch/de/2176.html	Swiss Federal Office for the Environment	31.12.2018

10.4 Links to CrowdWater spots

Table 13: Links to CrowdWater spots on interactive webapp by Spotteron. The number in the end of the URL refers to the spot number of the first observation at this site.

Catchment	Link to CrowdWater spot
Koenigsseeache, Niederalm	spotteron.com/crowdwater/spots/23445
Salzach, Salzburg	spotteron.com/crowdwater/spots/42809
Kempt, Fehraltdorf	spotteron.com/crowdwater/spots/221750
Urtene, Kernenried	spotteron.com/crowdwater/spots/35919
Alp, Einsiedeln	spotteron.com/crowdwater/spots/22659
Kleine Emme, Werthenstein	spotteron.com/crowdwater/spots/24494
Ova dal Fuorn, Zernez	spotteron.com/crowdwater/spots/20353
Kleine Emme, Emmen	spotteron.com/crowdwater/spots/24491
Wigger, Zofingen	spotteron.com/crowdwater/spots/24496
Sellenbodenbach, Neuenkirch	spotteron.com/crowdwater/spots/24493
Sihl, Zurich	spotteron.com/crowdwater/spots/17768

10.5 Form used at the pen and paper stations

crowd water

Formular: Wasserstand & Abfluss

Die mit einem * markierten Angaben müssen ausgefüllt werden. Das ausgefüllte Formular kann in den Briefkasten gelegt werden.

Datum*:_____ **Uhrzeit***:_____

Alter: _____

Geschlecht:
- o weiblich
- o männlich
- o anderes

Bildung:
- o Sekundarschule
- o Berufslehre
- o Matura
- o Uni/ Fachhochschule

Muttersprache: _____

Haben Sie dieses Formular bereits einmal ausgefüllt? o ja o nein
Wenn ja, wie oft? _____

1. Wasserstand

Betrachten Sie das untenstehende Bild. In welcher Kategorie der fiktiven Messlatte auf dem Bild würde sich der tatsächliche Wasserstand in der Kleinen Emme zurzeit befinden?

Kategorie: _____

Ein zusätzliches Foto mit Datum und Uhrzeit an info@crowdwater.ch würde uns sehr helfen!

BITTE WENDEN!

Figure 47: First page of the form used at all pen and paper stations. Example from the Kleine Emme in Werthenstein.

2. Abfluss – Wie viel Wasser fliesst den Fluss runter

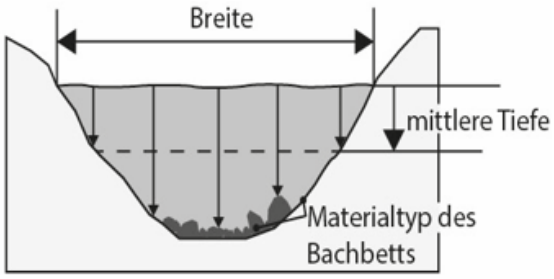
Um den Abfluss gut schätzen zu können, benötigt man Werte für die Breite, die durchschnittliche Tiefe und die Geschwindigkeit vom Fluss. Diese Schätzung kann noch zusätzlich verbessert werden, wenn auch der Materialtyp angegeben wird.

Breite [m]: _____

Mittlere Tiefe [m]: _____

Materialtyp im Bachbett:

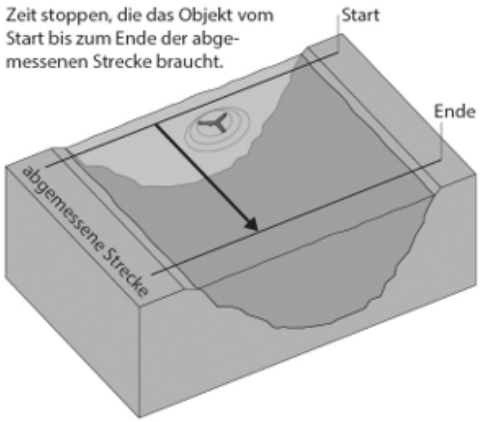
- Schlamm
- Sand
- Kies (kleiner als ein Hühnerei)
- Geröll (grösser als ein Hühnerei, aber unter 20cm Durchmesser)
- Felsblöcke (über 20cm Durchmesser)
- Felsuntergrund
- Beton



Bei einem Grenzfall können Sie auch bis zu **maximal zwei** Materialtypen notieren.

Fliessgeschwindigkeit [m/s]: _____

Abgemessene Strecke [m]: _____ Zeit [s]: _____



TIPP: Werfen Sie einen Stöckchen oder ein Blatt in den Fluss und stoppen sie die Zeit, die der Stecken für ca. 3 Meter benötigt. Diese Werte können umgerechnet werden, indem die Meteranzahl durch die Sekunden dividiert wird. Wenn Sie nicht rechnen möchten, können Sie auch die einzelnen Distanz- und Zeitangaben aufschreiben.

Was hat Sie motiviert hier mitzumachen? _____

_____

Was würde Sie motivieren, wiederholt teilzunehmen? _____

_____

Vielen herzlichen Dank
für Ihren Beitrag zu unserer Forschung!

Figure 48: Second page of the form used at all pen and paper stations. Example from the Kleine Emme in Werthenstein.

10.6 Elevation zones

Table 14: Elevation zones Koenigsseeache, Niederalm

Height of the elevation zone [m a. s. l.]	Proportion of catchment area in elevation zone
500	0.038
700	0.164
900	0.15
1100	0.128
1300	0.106
1500	0.094
1700	0.09
1900	0.082
2100	0.094
2300	0.043
2500	0.01
2700	0.001

Table 15: Elevation zones Salzach, Salzburg

Height of the elevation zone [m a. s. l.]	Proportion of catchment area in elevation zone
500	0.035
700	0.086
900	0.121
1100	0.122
1300	0.122
1500	0.111
1700	0.102
1900	0.093
2100	0.08
2300	0.054
2500	0.035
2700	0.022
2900	0.011
3100	0.004
3300	0.001
3500	0.001

Table 16: Elevation zones Kempt, Fehraltdorf

Height of the elevation zone [m a. s. l.]	Proportion of catchment area in elevation zone
560	0.393
650	0.362
750	0.173
860	0.072

Table 17: Elevation zones Urtene, Kernenried

Height of the elevation zone [m a. s. l.]	Proportion of catchment area in elevation zone
550	0.828
650	0.153
750	0.019

Table 18: Elevation zones Alp, Einsiedeln

Height of the elevation zone [m a. s. l.]	Proportion of catchment area in elevation zone
875	0.09
950	0.202
1050	0.143
1150	0.138
1250	0.132
1350	0.149
1450	0.115
1650	0.031

Table 19: Elevation zones Kleine Emme, Werthenstein

Height of the elevation zone [m a. s. l.]	Proportion of catchment area in elevation zone
650	0.021
750	0.088
850	0.122
950	0.141
1050	0.127
1150	0.094
1250	0.082
1350	0.078
1450	0.067
1550	0.058
1650	0.05
1750	0.033
1850	0.021
1950	0.011
2140	0.007

Table 20: Elevation zones Ova dal Fuorn, Zernez

Height of the elevation zone [m a. s. l.]	Proportion of catchment area in elevation zone
1770	0.009
1850	0.037
1950	0.088
2050	0.105
2150	0.116
2250	0.139
2350	0.117
2450	0.092
2550	0.095
2650	0.094
2750	0.055
2850	0.039
3000	0.014

Table 21: Elevation zones Kleine Emme, Emmen

Height of the elevation zone [m a. s. l.]	Proportion of catchment area in elevation zone
470	0.022
550	0.039
650	0.073
750	0.116
850	0.134
950	0.127
1050	0.105
1150	0.078
1250	0.068
1350	0.059
1450	0.051
1550	0.042
1650	0.036
1750	0.024
1850	0.014
2050	0.012

Table 22: Elevation zones Wigger, Zofingen

Height of the elevation zone [m a. s. l.]	Proportion of catchment area in elevation zone
465	0.086
550	0.278
650	0.33
750	0.16
850	0.076
950	0.043
1050	0.017
1250	0.01

Table 23: Elevation zones Sellenbodenbach, Neuenkirch

Height of the elevation zone [m a. s. l.]	Proportion of catchment area in elevation zone
550	0.624
650	0.204
765	0.172

Table 24: Elevation zones Sihl, Zurich

Height of the elevation zone [m a. s. l.]	Proportion of catchment area in elevation zone
470	0.034
550	0.049
650	0.082
750	0.058
850	0.109
950	0.161
1050	0.106
1150	0.098
1250	0.079
1350	0.078
1450	0.061
1550	0.032
1650	0.02
1750	0.016
1850	0.01
2050	0.007

10.7 Ranking of parameter sets

Figure 49: Plots showing the logarithmic rank of each parameter set in each scenario against the logarithmic rank in the upper benchmark for the catchments that were not shown in the results. The pink line shows the linear regression, the number states the root mean squared error. The scenarios are sorted as in the heatmaps in the results section.

10.8 Shared parameter sets

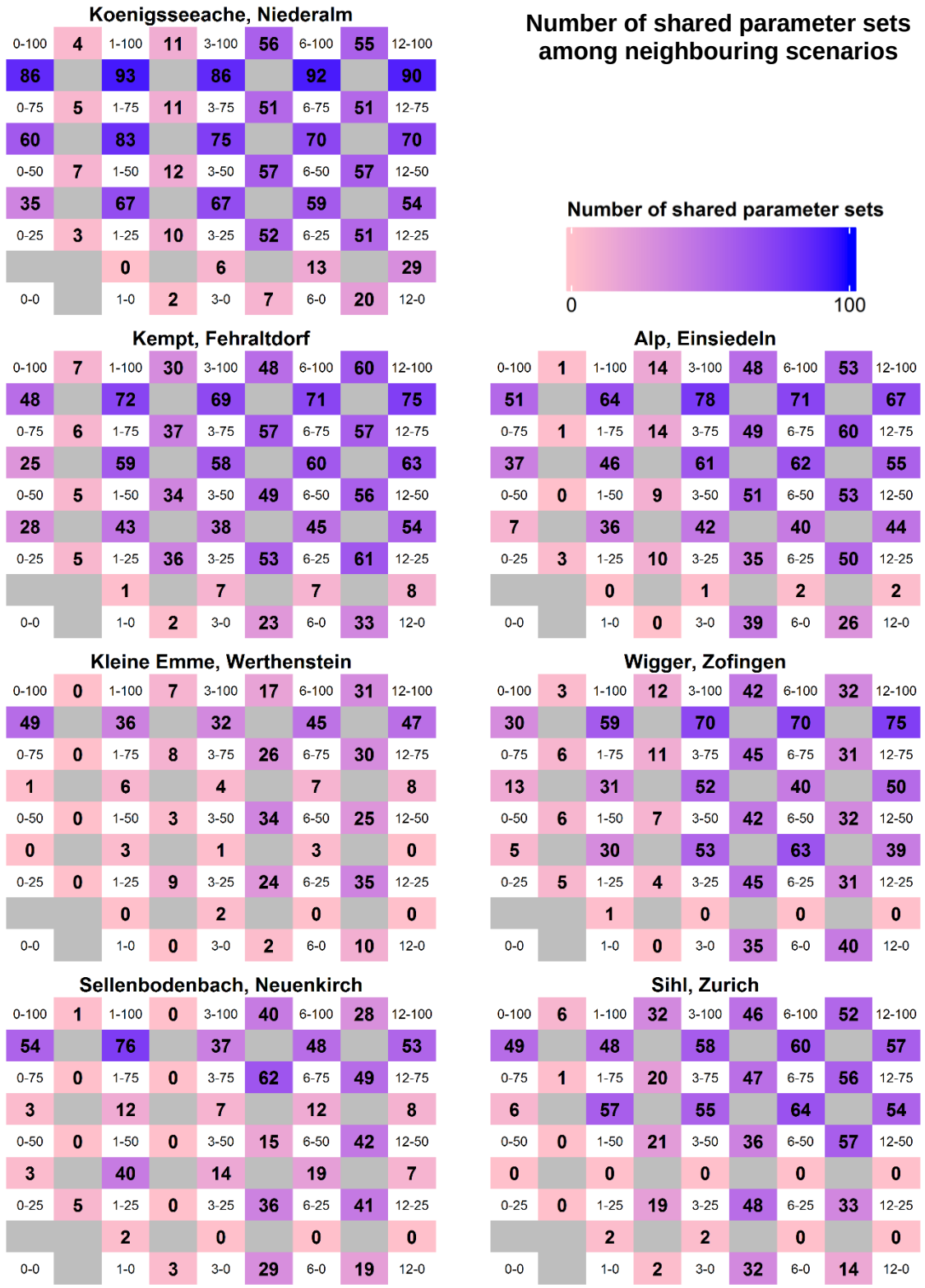
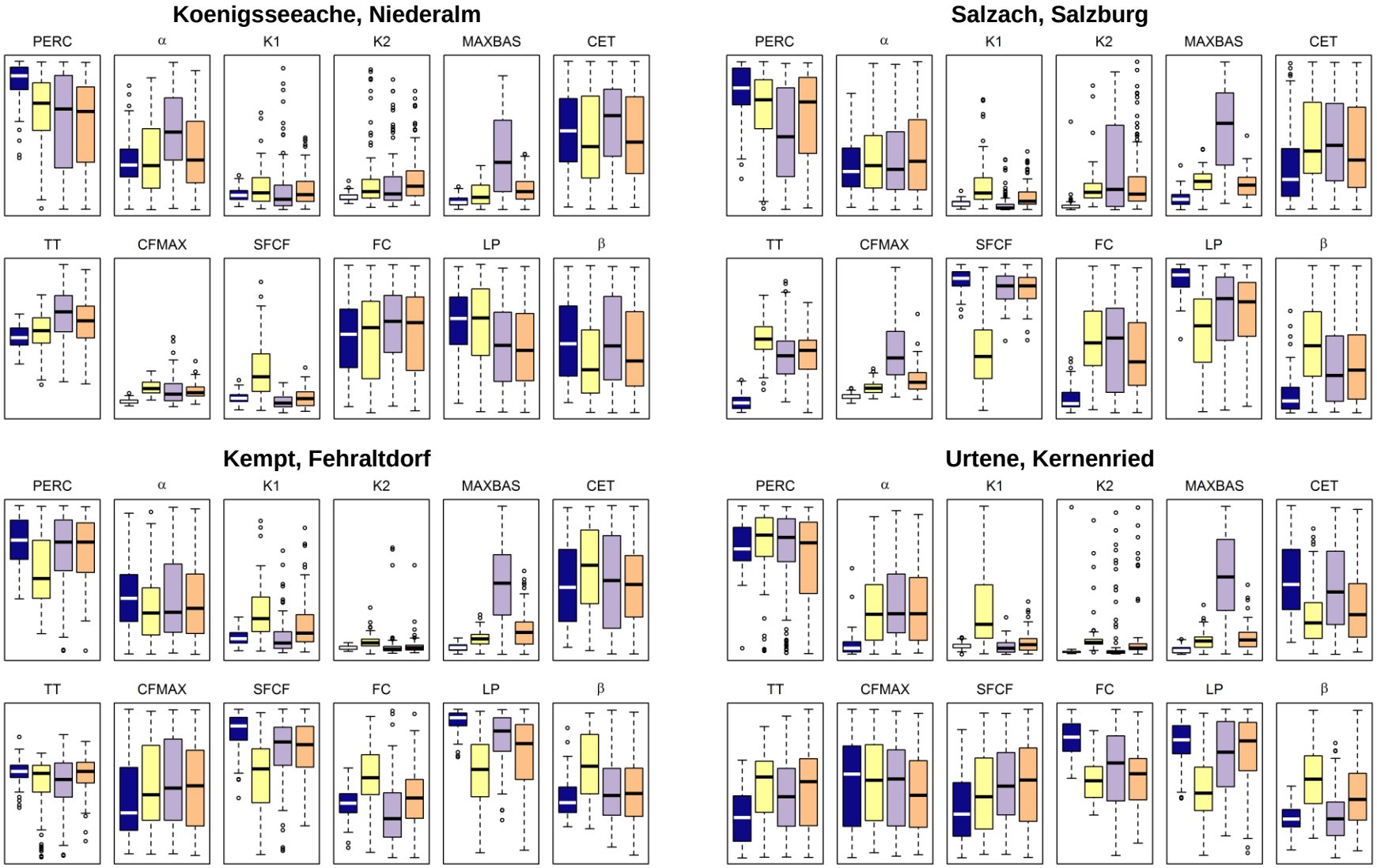
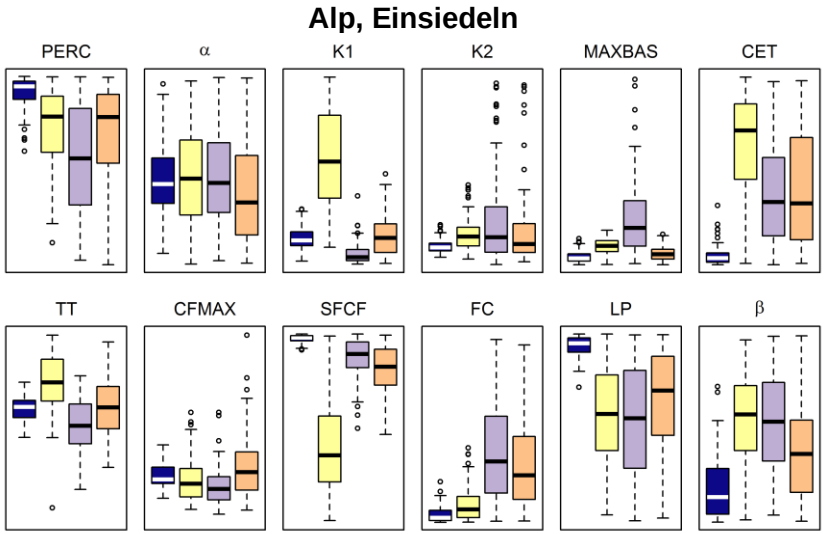
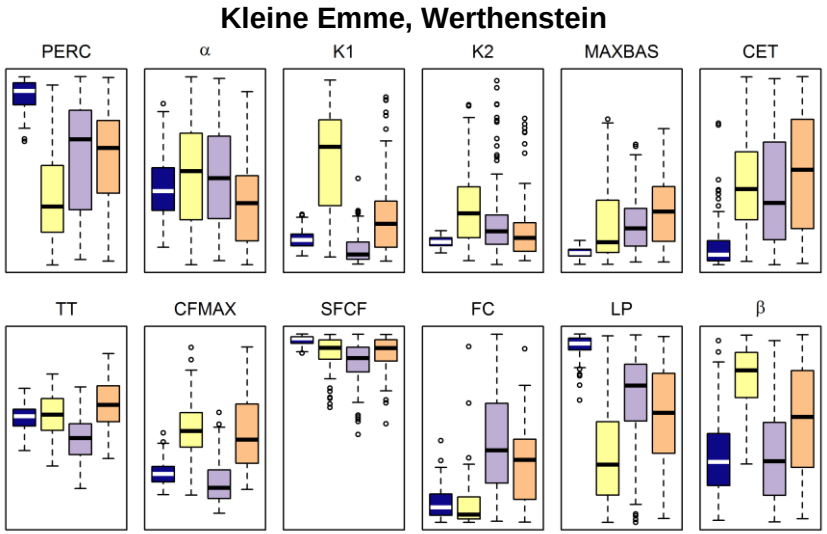


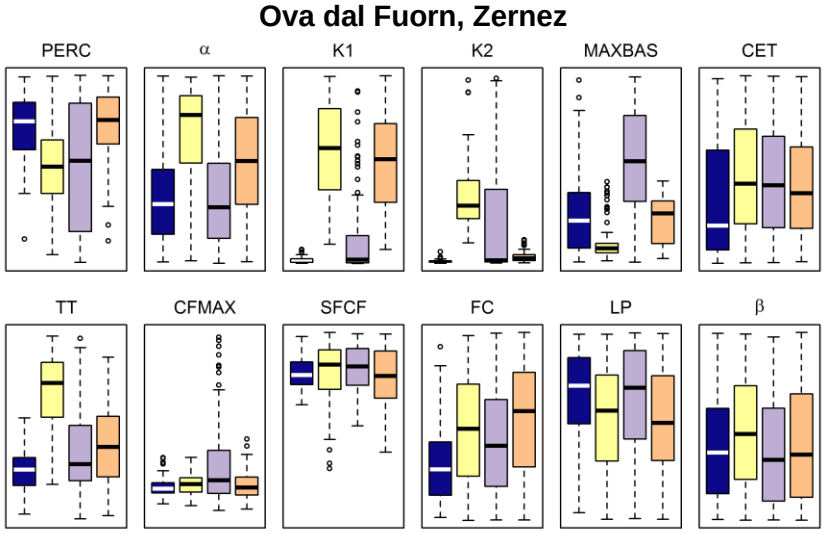
Figure 50: Plots showing how many of the top 100 parameter sets are shared among neighbouring scenarios, for all catchments that were not shown in the results. The names of the scenarios are given in the white fields, the number of shared parameter sets are given in the coloured fields, whereby a high number is indicated with a dark colour and a low number is indicated with a bright colour. The grey fields serve as placeholders.

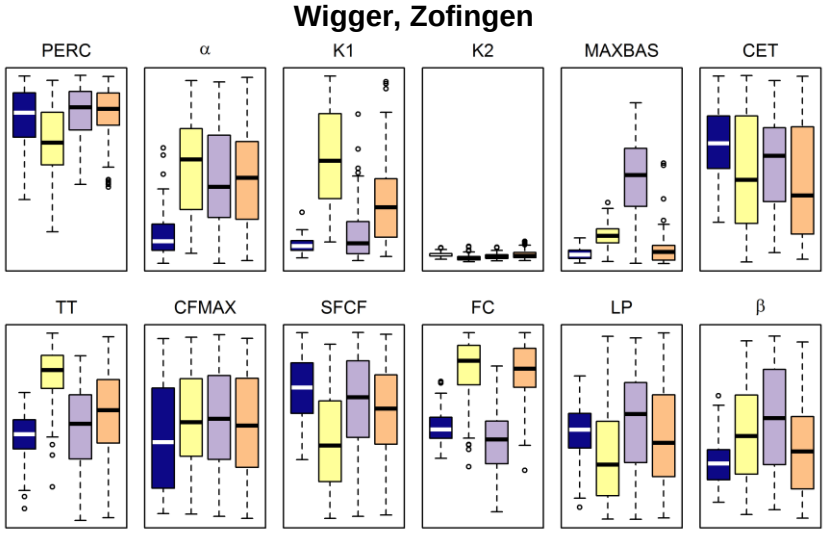
10.9 Distribution of parameter values




Alp, Einsiedeln
PERC
α
K1
K2
MAXBAS
CET
TT
CFMAX
SFCF
FC
LP
β


Kleine Emme, Werthenstein
PERC
α
K1
K2
MAXBAS
CET
TT
CFMAX
SFCF
FC
LP
β


Ova dal Fuorn, Zernez
PERC
α
K1
K2
MAXBAS
CET
TT
CFMAX
SFCF
FC
LP
β


Wigger, Zofingen
PERC
α
K1
K2
MAXBAS
CET
TT
CFMAX
SFCF
FC
LP
β

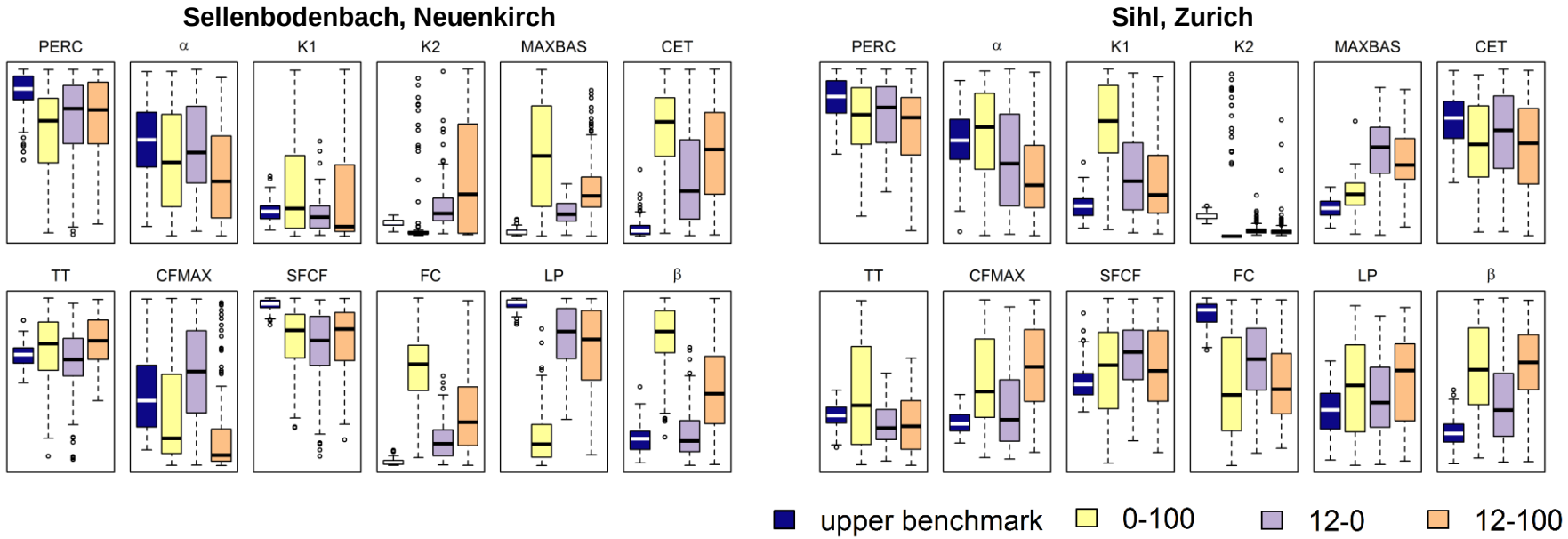


Figure 51: Distribution of the parameter values obtained in the 100 parameter sets resulting from the GAP calibration for the upper benchmark as well as in the top 100 parameter sets of the scenarios in the corners (0-100, 12-0 and 12-100) for all catchments except for the Kleine Emme in Emmen which was shown in the results. The y-axes cover the ranges that were allowed for each parameter.

10.10 Density plots NPE vs. volume error

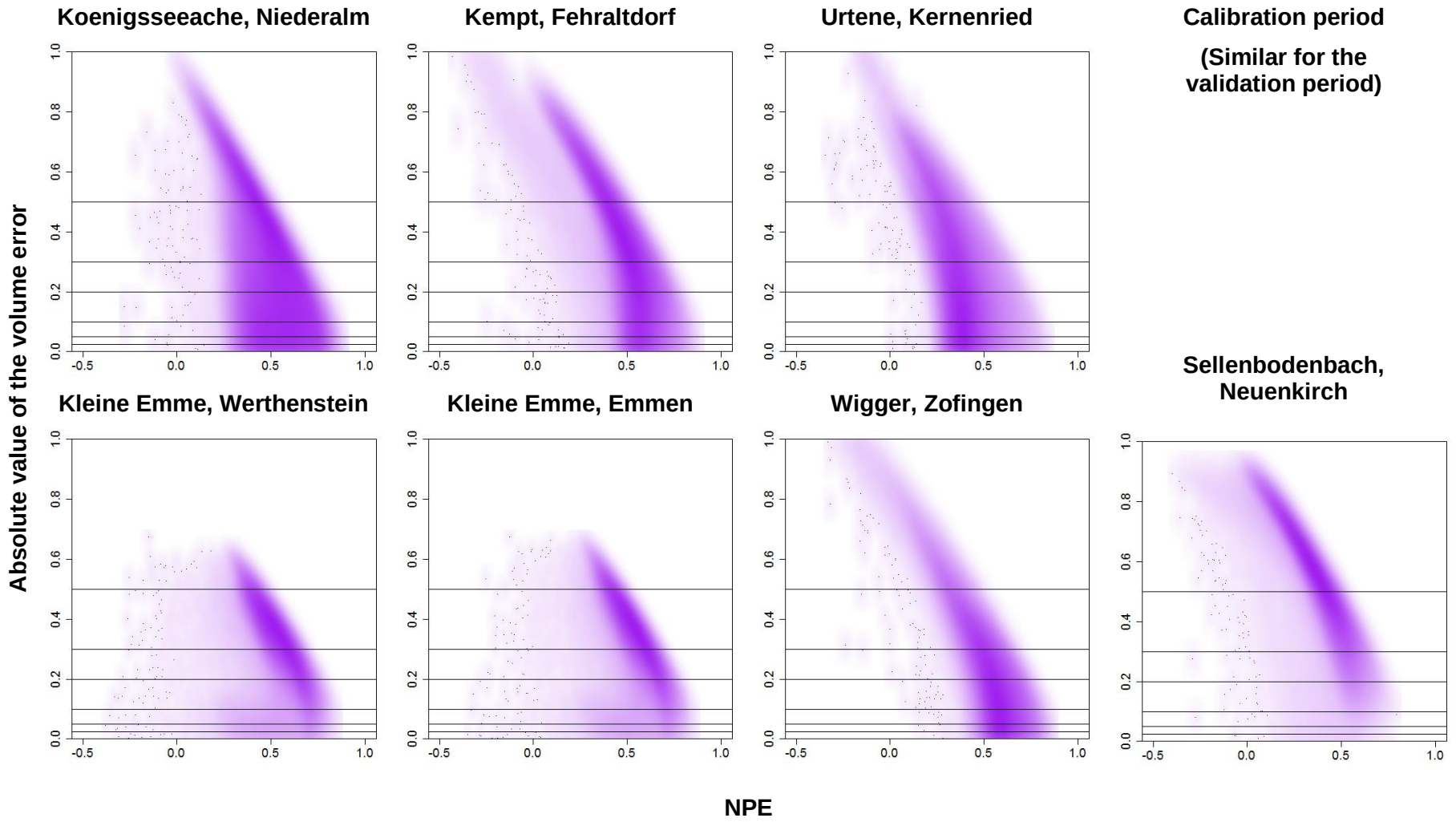
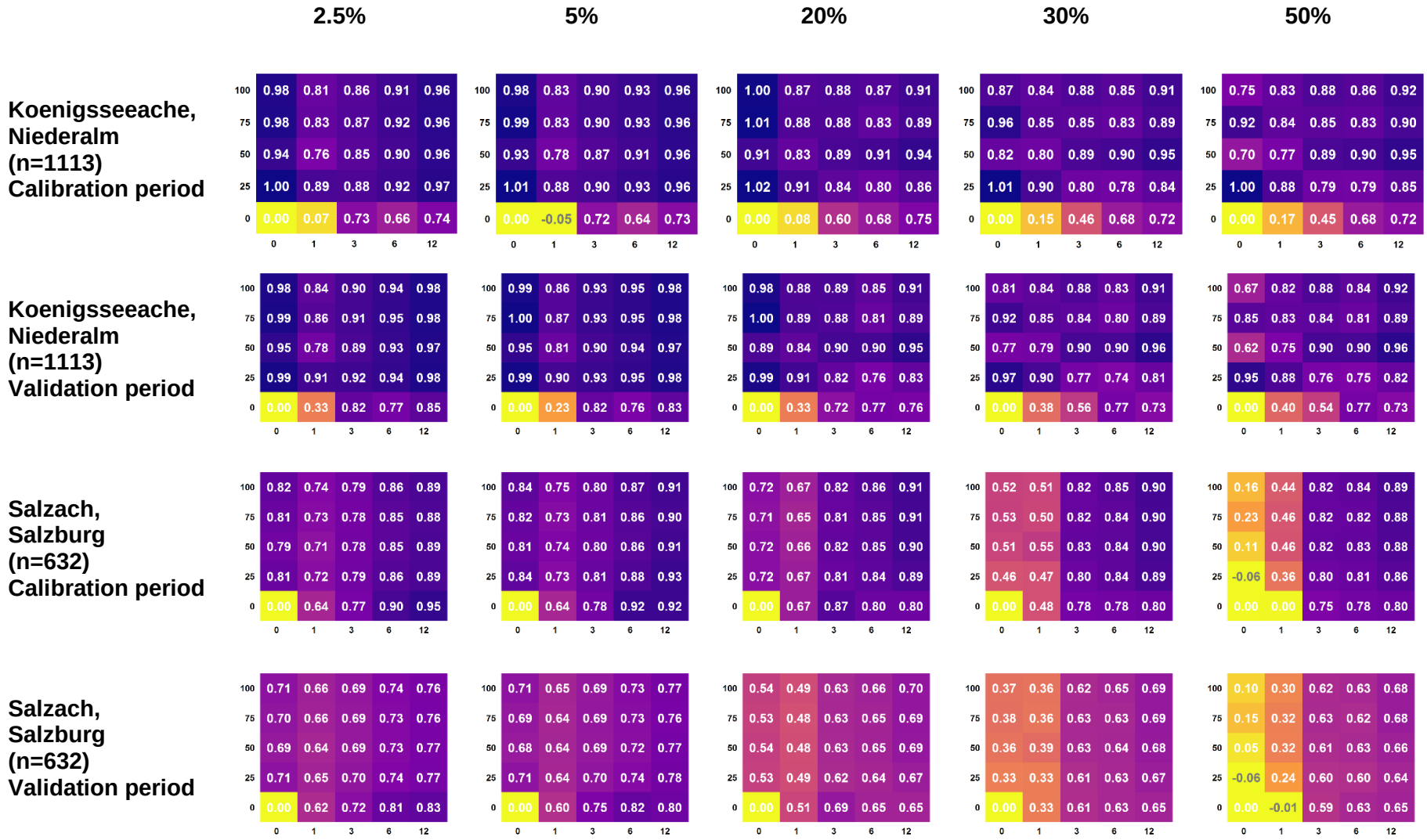
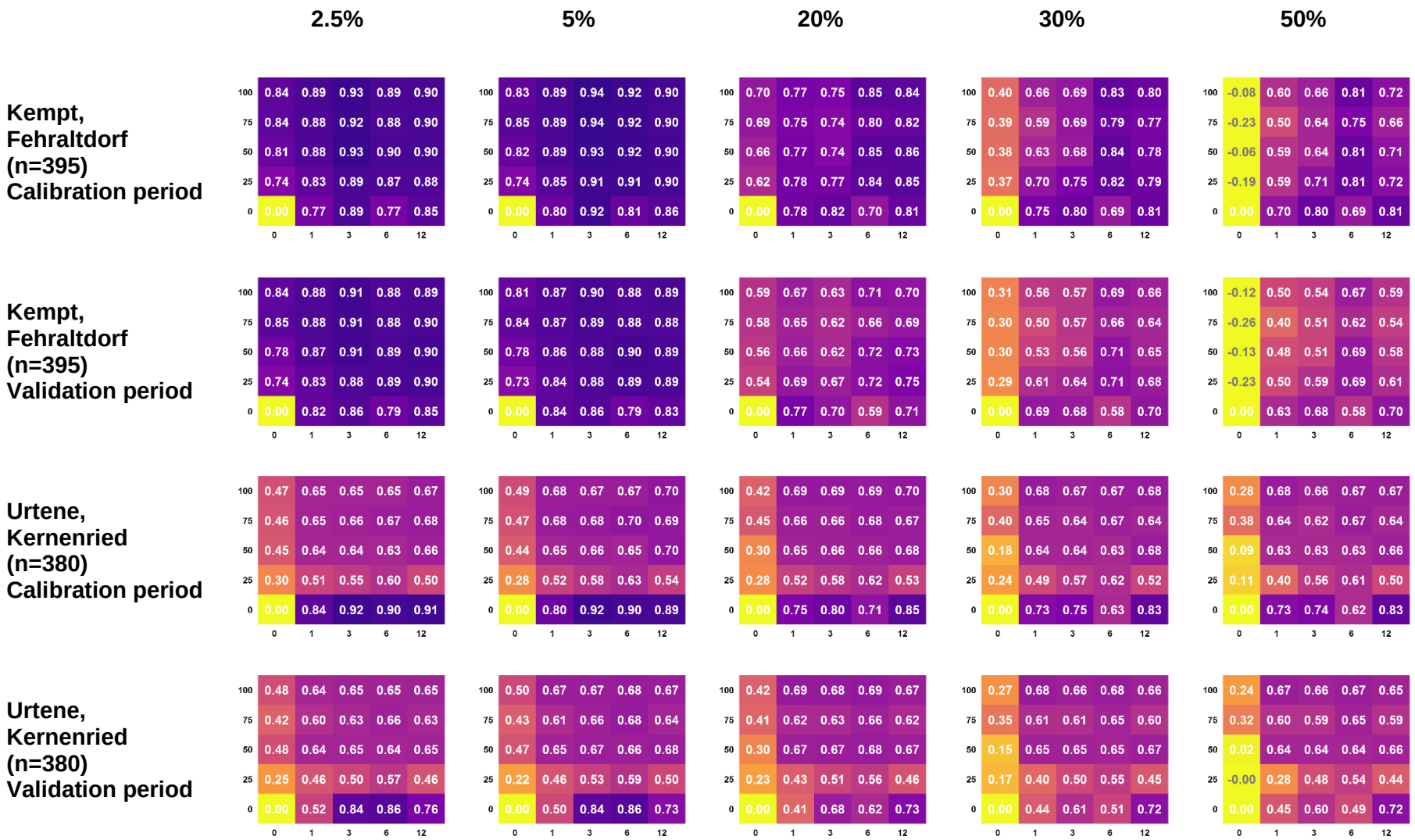
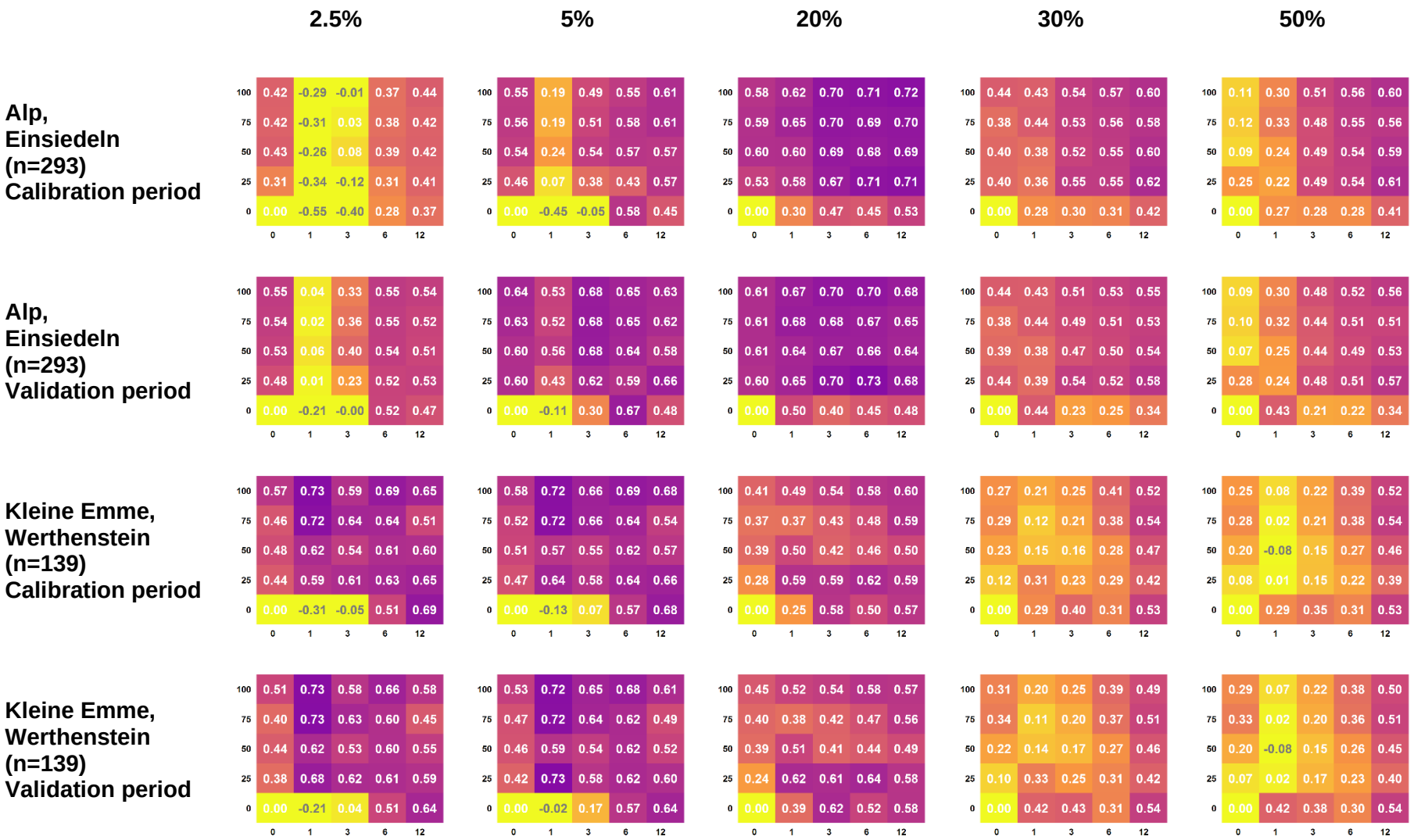


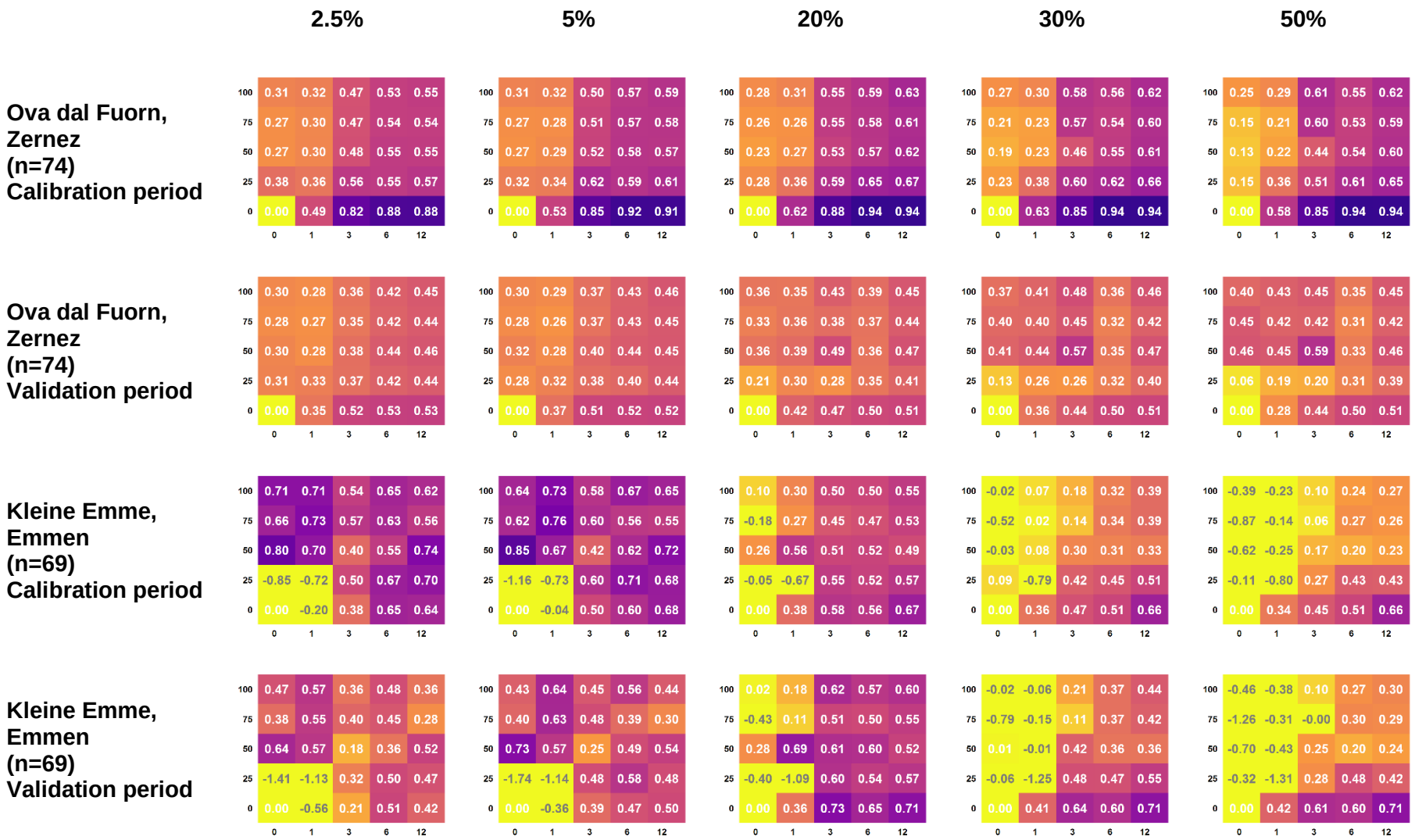
Figure 52: Plots showing the NPE performance against the volume error of each parameter set for the catchments that were not shown in the results. Results for the calibration period. The horizontal lines indicate the volume error filters that were tested, i.e., all parameter sets above a horizontal line were excluded if the corresponding filter was applied.

10.11 Results from other filters




2.5%
5%
20%
30%
50%
Kempt, Fehraltdorf (n=395) Calibration period
100 0.84 0.89 0.93 0.89 0.90
75 0.84 0.88 0.92 0.88 0.90
50 0.81 0.88 0.93 0.90 0.90
25 0.74 0.83 0.89 0.87 0.88
0 0.00 0.77 0.89 0.77 0.85
0 1 3 6 12
100 0.83 0.89 0.94 0.92 0.90
75 0.85 0.89 0.94 0.92 0.90
50 0.82 0.89 0.93 0.92 0.90
25 0.74 0.85 0.91 0.91 0.90
0 0.00 0.80 0.92 0.81 0.86
0 1 3 6 12
100 0.70 0.77 0.75 0.85 0.84
75 0.69 0.75 0.74 0.80 0.82
50 0.66 0.77 0.74 0.85 0.86
25 0.62 0.78 0.77 0.84 0.85
0 0.00 0.78 0.82 0.70 0.81
0 1 3 6 12
100 0.40 0.66 0.69 0.83 0.80
75 0.39 0.59 0.69 0.79 0.77
50 0.38 0.63 0.68 0.84 0.78
25 0.37 0.70 0.75 0.82 0.79
0 0.00 0.75 0.80 0.69 0.81
0 1 3 6 12
100 -0.08 0.60 0.66 0.81 0.72
75 -0.23 0.50 0.64 0.75 0.66
50 -0.06 0.59 0.64 0.81 0.71
25 -0.19 0.59 0.71 0.81 0.72
0 0.00 0.70 0.80 0.69 0.81
0 1 3 6 12
Kempt, Fehraltdorf (n=395) Validation period
100 0.84 0.88 0.91 0.88 0.89
75 0.85 0.88 0.91 0.88 0.90
50 0.78 0.87 0.91 0.89 0.90
25 0.74 0.83 0.88 0.89 0.90
0 0.00 0.82 0.86 0.79 0.85
0 1 3 6 12
100 0.81 0.87 0.90 0.88 0.89
75 0.84 0.87 0.89 0.88 0.88
50 0.78 0.86 0.88 0.90 0.89
25 0.73 0.84 0.88 0.89 0.89
0 0.00 0.84 0.86 0.79 0.83
0 1 3 6 12
100 0.59 0.67 0.63 0.71 0.70
75 0.58 0.65 0.62 0.66 0.69
50 0.56 0.66 0.62 0.72 0.73
25 0.54 0.69 0.67 0.72 0.75
0 0.00 0.77 0.70 0.59 0.71
0 1 3 6 12
100 0.31 0.56 0.57 0.69 0.66
75 0.30 0.50 0.57 0.66 0.64
50 0.30 0.53 0.56 0.71 0.65
25 0.29 0.61 0.64 0.71 0.68
0 0.00 0.69 0.68 0.58 0.70
0 1 3 6 12
100 -0.12 0.50 0.54 0.67 0.59
75 -0.26 0.40 0.51 0.62 0.54
50 -0.13 0.48 0.51 0.69 0.58
25 -0.23 0.50 0.59 0.69 0.61
0 0.00 0.63 0.68 0.58 0.70
0 1 3 6 12
Urtene, Kernenried (n=380) Calibration period
100 0.47 0.65 0.65 0.65 0.67
75 0.46 0.65 0.66 0.67 0.68
50 0.45 0.64 0.64 0.63 0.66
25 0.30 0.51 0.55 0.60 0.50
0 0.00 0.84 0.92 0.90 0.91
0 1 3 6 12
100 0.49 0.68 0.67 0.67 0.70
75 0.47 0.68 0.68 0.70 0.69
50 0.44 0.65 0.66 0.65 0.70
25 0.28 0.52 0.58 0.63 0.54
0 0.00 0.80 0.92 0.90 0.89
0 1 3 6 12
100 0.42 0.69 0.69 0.69 0.70
75 0.45 0.66 0.66 0.68 0.67
50 0.30 0.65 0.66 0.66 0.68
25 0.28 0.52 0.58 0.62 0.53
0 0.00 0.75 0.80 0.71 0.85
0 1 3 6 12
100 0.30 0.68 0.67 0.67 0.68
75 0.40 0.65 0.64 0.67 0.64
50 0.18 0.64 0.64 0.63 0.68
25 0.24 0.49 0.57 0.62 0.52
0 0.00 0.73 0.75 0.63 0.83
0 1 3 6 12
100 0.28 0.68 0.66 0.67 0.67
75 0.38 0.64 0.62 0.67 0.64
50 0.09 0.63 0.63 0.63 0.66
25 0.11 0.40 0.56 0.61 0.50
0 0.00 0.73 0.74 0.62 0.83
0 1 3 6 12
Urtene, Kernenried (n=380) Validation period
100 0.48 0.64 0.65 0.65 0.65
75 0.42 0.60 0.63 0.66 0.63
50 0.48 0.64 0.65 0.64 0.65
25 0.25 0.46 0.50 0.57 0.46
0 0.00 0.52 0.84 0.86 0.76
0 1 3 6 12
100 0.50 0.67 0.67 0.68 0.67
75 0.43 0.61 0.66 0.68 0.64
50 0.47 0.65 0.67 0.66 0.68
25 0.22 0.46 0.53 0.59 0.50
0 0.00 0.50 0.84 0.86 0.73
0 1 3 6 12
100 0.42 0.69 0.68 0.69 0.67
75 0.41 0.62 0.63 0.66 0.62
50 0.30 0.67 0.67 0.68 0.67
25 0.23 0.43 0.51 0.56 0.46
0 0.00 0.41 0.68 0.62 0.73
0 1 3 6 12
100 0.27 0.68 0.66 0.68 0.66
75 0.35 0.61 0.61 0.65 0.60
50 0.15 0.65 0.65 0.65 0.67
25 0.17 0.40 0.50 0.55 0.45
0 0.00 0.44 0.61 0.51 0.72
0 1 3 6 12
100 0.24 0.67 0.66 0.67 0.65
75 0.32 0.60 0.59 0.65 0.59
50 0.02 0.64 0.64 0.64 0.66
25 -0.00 0.28 0.48 0.54 0.44
0 0.00 0.45 0.60 0.49 0.72
0 1 3 6 12


2.5%
5%
20%
30%
50%
Alp, Einsiedeln (n=293) Calibration period
100 0.42 -0.29 -0.01 0.37 0.44
75 0.42 -0.31 0.03 0.38 0.42
50 0.43 -0.26 0.08 0.39 0.42
25 0.31 -0.34 -0.12 0.31 0.41
0 0.00 -0.55 -0.40 0.28 0.37
0 1 3 6 12
100 0.55 0.19 0.49 0.55 0.61
75 0.56 0.19 0.51 0.58 0.61
50 0.54 0.24 0.54 0.57 0.57
25 0.46 0.07 0.38 0.43 0.57
0 0.00 -0.45 -0.05 0.58 0.45
0 1 3 6 12
100 0.58 0.62 0.70 0.71 0.72
75 0.59 0.65 0.70 0.69 0.70
50 0.60 0.60 0.69 0.68 0.69
25 0.53 0.58 0.67 0.71 0.71
0 0.00 0.30 0.47 0.45 0.53
0 1 3 6 12
100 0.44 0.43 0.54 0.57 0.60
75 0.38 0.44 0.53 0.56 0.58
50 0.40 0.38 0.52 0.55 0.60
25 0.40 0.36 0.55 0.55 0.62
0 0.00 0.28 0.30 0.31 0.42
0 1 3 6 12
100 0.11 0.30 0.51 0.56 0.60
75 0.12 0.33 0.48 0.55 0.56
50 0.09 0.24 0.49 0.54 0.59
25 0.25 0.22 0.49 0.54 0.61
0 0.00 0.27 0.28 0.28 0.41
0 1 3 6 12
Alp, Einsiedeln (n=293) Validation period
100 0.55 0.04 0.33 0.55 0.54
75 0.54 0.02 0.36 0.55 0.52
50 0.53 0.06 0.40 0.54 0.51
25 0.48 0.01 0.23 0.52 0.53
0 0.00 -0.21 -0.00 0.52 0.47
0 1 3 6 12
100 0.64 0.53 0.68 0.65 0.63
75 0.63 0.52 0.68 0.65 0.62
50 0.60 0.56 0.68 0.64 0.58
25 0.60 0.43 0.62 0.59 0.66
0 0.00 -0.11 0.30 0.67 0.48
0 1 3 6 12
100 0.61 0.67 0.70 0.70 0.68
75 0.61 0.68 0.68 0.67 0.65
50 0.61 0.64 0.67 0.66 0.64
25 0.60 0.65 0.70 0.73 0.68
0 0.00 0.50 0.40 0.45 0.48
0 1 3 6 12
100 0.44 0.43 0.51 0.53 0.55
75 0.38 0.44 0.49 0.51 0.53
50 0.39 0.38 0.47 0.50 0.54
25 0.44 0.39 0.54 0.52 0.58
0 0.00 0.44 0.23 0.25 0.34
0 1 3 6 12
100 0.09 0.30 0.48 0.52 0.56
75 0.10 0.32 0.44 0.51 0.51
50 0.07 0.25 0.44 0.49 0.53
25 0.28 0.24 0.48 0.51 0.57
0 0.00 0.43 0.21 0.22 0.34
0 1 3 6 12
Kleine Emme, Werthenstein (n=139) Calibration period
100 0.57 0.73 0.59 0.69 0.65
75 0.46 0.72 0.64 0.64 0.51
50 0.48 0.62 0.54 0.61 0.60
25 0.44 0.59 0.61 0.63 0.65
0 0.00 -0.31 -0.05 0.51 0.69
0 1 3 6 12
100 0.58 0.72 0.66 0.69 0.68
75 0.52 0.72 0.66 0.64 0.54
50 0.51 0.57 0.55 0.62 0.57
25 0.47 0.64 0.58 0.64 0.66
0 0.00 -0.13 0.07 0.57 0.68
0 1 3 6 12
100 0.41 0.49 0.54 0.58 0.60
75 0.37 0.37 0.43 0.48 0.59
50 0.39 0.50 0.42 0.46 0.50
25 0.28 0.59 0.59 0.62 0.59
0 0.00 0.25 0.58 0.50 0.57
0 1 3 6 12
100 0.27 0.21 0.25 0.41 0.52
75 0.29 0.12 0.21 0.38 0.54
50 0.23 0.15 0.16 0.28 0.47
25 0.12 0.31 0.23 0.29 0.42
0 0.00 0.29 0.40 0.31 0.53
0 1 3 6 12
100 0.25 0.08 0.22 0.39 0.52
75 0.28 0.02 0.21 0.38 0.54
50 0.20 -0.08 0.15 0.27 0.46
25 0.08 0.01 0.15 0.22 0.39
0 0.00 0.29 0.35 0.31 0.53
0 1 3 6 12
Kleine Emme, Werthenstein (n=139) Validation period
100 0.51 0.73 0.58 0.66 0.58
75 0.40 0.73 0.63 0.60 0.45
50 0.44 0.62 0.53 0.60 0.55
25 0.38 0.68 0.62 0.61 0.59
0 0.00 -0.21 0.04 0.51 0.64
0 1 3 6 12
100 0.53 0.72 0.65 0.68 0.61
75 0.47 0.72 0.64 0.62 0.49
50 0.46 0.59 0.54 0.62 0.52
25 0.42 0.73 0.58 0.62 0.60
0 0.00 -0.02 0.17 0.57 0.64
0 1 3 6 12
100 0.45 0.52 0.54 0.58 0.57
75 0.40 0.38 0.42 0.47 0.56
50 0.39 0.51 0.41 0.44 0.49
25 0.24 0.62 0.61 0.64 0.58
0 0.00 0.39 0.62 0.52 0.58
0 1 3 6 12
100 0.31 0.20 0.25 0.39 0.49
75 0.34 0.11 0.20 0.37 0.51
50 0.22 0.14 0.17 0.27 0.46
25 0.10 0.33 0.25 0.31 0.42
0 0.00 0.42 0.43 0.31 0.54
0 1 3 6 12
100 0.29 0.07 0.22 0.38 0.50
75 0.33 0.02 0.20 0.36 0.51
50 0.20 -0.08 0.15 0.26 0.45
25 0.07 0.02 0.17 0.23 0.40
0 0.00 0.42 0.38 0.30 0.54
0 1 3 6 12


2.5%
5%
20%
30%
50%
Ova dal Fuorn, Zernez (n=74) Calibration period
100 0.31 0.32 0.47 0.53 0.55
75 0.27 0.30 0.47 0.54 0.54
50 0.27 0.30 0.48 0.55 0.55
25 0.38 0.36 0.56 0.55 0.57
0 0.00 0.49 0.82 0.88 0.88
0 1 3 6 12
100 0.31 0.32 0.50 0.57 0.59
75 0.27 0.28 0.51 0.57 0.58
50 0.27 0.29 0.52 0.58 0.57
25 0.32 0.34 0.62 0.59 0.61
0 0.00 0.53 0.85 0.92 0.91
0 1 3 6 12
100 0.28 0.31 0.55 0.59 0.63
75 0.26 0.26 0.55 0.58 0.61
50 0.23 0.27 0.53 0.57 0.62
25 0.28 0.36 0.59 0.65 0.67
0 0.00 0.62 0.88 0.94 0.94
0 1 3 6 12
100 0.27 0.30 0.58 0.56 0.62
75 0.21 0.23 0.57 0.54 0.60
50 0.19 0.23 0.46 0.55 0.61
25 0.23 0.38 0.60 0.62 0.66
0 0.00 0.63 0.85 0.94 0.94
0 1 3 6 12
100 0.25 0.29 0.61 0.55 0.62
75 0.15 0.21 0.60 0.53 0.59
50 0.13 0.22 0.44 0.54 0.60
25 0.15 0.36 0.51 0.61 0.65
0 0.00 0.58 0.85 0.94 0.94
0 1 3 6 12
Ova dal Fuorn, Zernez (n=74) Validation period
100 0.30 0.28 0.36 0.42 0.45
75 0.28 0.27 0.35 0.42 0.44
50 0.30 0.28 0.38 0.44 0.46
25 0.31 0.33 0.37 0.42 0.44
0 0.00 0.35 0.52 0.53 0.53
0 1 3 6 12
100 0.30 0.29 0.37 0.43 0.46
75 0.28 0.26 0.37 0.43 0.45
50 0.32 0.28 0.40 0.44 0.45
25 0.28 0.32 0.38 0.40 0.44
0 0.00 0.37 0.51 0.52 0.52
0 1 3 6 12
100 0.36 0.35 0.43 0.39 0.45
75 0.33 0.36 0.38 0.37 0.44
50 0.36 0.39 0.49 0.36 0.47
25 0.21 0.30 0.28 0.35 0.41
0 0.00 0.42 0.47 0.50 0.51
0 1 3 6 12
100 0.37 0.41 0.48 0.36 0.46
75 0.40 0.40 0.45 0.32 0.42
50 0.41 0.44 0.57 0.35 0.47
25 0.13 0.26 0.26 0.32 0.40
0 0.00 0.36 0.44 0.50 0.51
0 1 3 6 12
100 0.40 0.43 0.45 0.35 0.45
75 0.45 0.42 0.42 0.31 0.42
50 0.46 0.45 0.59 0.33 0.46
25 0.06 0.19 0.20 0.31 0.39
0 0.00 0.28 0.44 0.50 0.51
0 1 3 6 12
Kleine Emme, Emmen (n=69) Calibration period
100 0.71 0.71 0.54 0.65 0.62
75 0.66 0.73 0.57 0.63 0.56
50 0.80 0.70 0.40 0.55 0.74
25 -0.85 -0.72 0.50 0.67 0.70
0 0.00 -0.20 0.38 0.65 0.64
0 1 3 6 12
100 0.64 0.73 0.58 0.67 0.65
75 0.62 0.76 0.60 0.56 0.55
50 0.85 0.67 0.42 0.62 0.72
25 -1.16 -0.73 0.60 0.71 0.68
0 0.00 -0.04 0.50 0.60 0.68
0 1 3 6 12
100 0.10 0.30 0.50 0.50 0.55
75 -0.18 0.27 0.45 0.47 0.53
50 0.26 0.56 0.51 0.52 0.49
25 -0.05 -0.67 0.55 0.52 0.57
0 0.00 0.38 0.58 0.56 0.67
0 1 3 6 12
100 -0.02 0.07 0.18 0.32 0.39
75 -0.52 0.02 0.14 0.34 0.39
50 -0.03 0.08 0.30 0.31 0.33
25 0.09 -0.79 0.42 0.45 0.51
0 0.00 0.36 0.47 0.51 0.66
0 1 3 6 12
100 -0.39 -0.23 0.10 0.24 0.27
75 -0.87 -0.14 0.06 0.27 0.26
50 -0.62 -0.25 0.17 0.20 0.23
25 -0.11 -0.80 0.27 0.43 0.43
0 0.00 0.34 0.45 0.51 0.66
0 1 3 6 12
Kleine Emme, Emmen (n=69) Validation period
100 0.47 0.57 0.36 0.48 0.36
75 0.38 0.55 0.40 0.45 0.28
50 0.64 0.57 0.18 0.36 0.52
25 -1.41 -1.13 0.32 0.50 0.47
0 0.00 -0.56 0.21 0.51 0.42
0 1 3 6 12
100 0.43 0.64 0.45 0.56 0.44
75 0.40 0.63 0.48 0.39 0.30
50 0.73 0.57 0.25 0.49 0.54
25 -1.74 -1.14 0.48 0.58 0.48
0 0.00 -0.36 0.39 0.47 0.50
0 1 3 6 12
100 0.02 0.18 0.62 0.57 0.60
75 -0.43 0.11 0.51 0.50 0.55
50 0.28 0.69 0.61 0.60 0.52
25 -0.40 -1.09 0.60 0.54 0.57
0 0.00 0.36 0.73 0.65 0.71
0 1 3 6 12
100 -0.02 -0.06 0.21 0.37 0.44
75 -0.79 -0.15 0.11 0.37 0.42
50 0.01 -0.01 0.42 0.36 0.36
25 -0.06 -1.25 0.48 0.47 0.55
0 0.00 0.41 0.64 0.60 0.71
0 1 3 6 12
100 -0.46 -0.38 0.10 0.27 0.30
75 -1.26 -0.31 -0.00 0.30 0.29
50 -0.70 -0.43 0.25 0.20 0.24
25 -0.32 -1.31 0.28 0.48 0.42
0 0.00 0.42 0.61 0.60 0.71
0 1 3 6 12

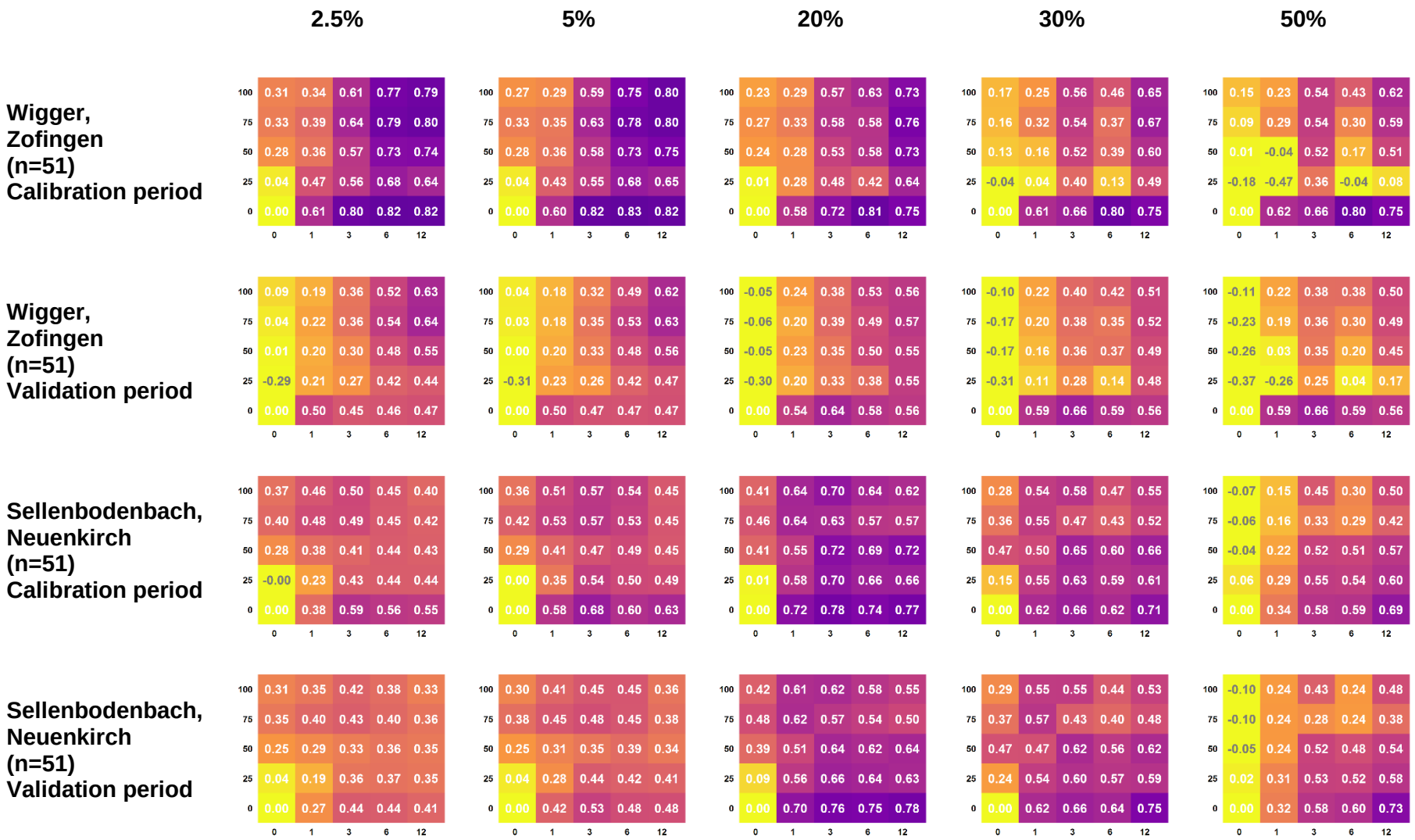

2.5%
5%
20%
30%
50%
Wigger, Zofingen (n=51) Calibration period
100 0.31 0.34 0.61 0.77 0.79
75 0.33 0.39 0.64 0.79 0.80
50 0.28 0.36 0.57 0.73 0.74
25 0.04 0.47 0.56 0.68 0.64
0 0.00 0.61 0.80 0.82 0.82
0 1 3 6 12
100 0.27 0.29 0.59 0.75 0.80
75 0.33 0.35 0.63 0.78 0.80
50 0.28 0.36 0.58 0.73 0.75
25 0.04 0.43 0.55 0.68 0.65
0 0.00 0.60 0.82 0.83 0.82
0 1 3 6 12
100 0.23 0.29 0.57 0.63 0.73
75 0.27 0.33 0.58 0.58 0.76
50 0.24 0.28 0.53 0.58 0.73
25 0.01 0.28 0.48 0.42 0.64
0 0.00 0.58 0.72 0.81 0.75
0 1 3 6 12
100 0.17 0.25 0.56 0.46 0.65
75 0.16 0.32 0.54 0.37 0.67
50 0.13 0.16 0.52 0.39 0.60
25 -0.04 0.04 0.40 0.13 0.49
0 0.00 0.61 0.66 0.80 0.75
0 1 3 6 12
100 0.15 0.23 0.54 0.43 0.62
75 0.09 0.29 0.54 0.30 0.59
50 0.01 -0.04 0.52 0.17 0.51
25 -0.18 -0.47 0.36 -0.04 0.08
0 0.00 0.62 0.66 0.80 0.75
0 1 3 6 12
Wigger, Zofingen (n=51) Validation period
100 0.09 0.19 0.36 0.52 0.63
75 0.04 0.22 0.36 0.54 0.64
50 0.01 0.20 0.30 0.48 0.55
25 -0.29 0.21 0.27 0.42 0.44
0 0.00 0.50 0.45 0.46 0.47
0 1 3 6 12
100 0.04 0.18 0.32 0.49 0.62
75 0.03 0.18 0.35 0.53 0.63
50 0.00 0.20 0.33 0.48 0.56
25 -0.31 0.23 0.26 0.42 0.47
0 0.00 0.50 0.47 0.47 0.47
0 1 3 6 12
100 -0.05 0.24 0.38 0.53 0.56
75 -0.06 0.20 0.39 0.49 0.57
50 -0.05 0.23 0.35 0.50 0.55
25 -0.30 0.20 0.33 0.38 0.55
0 0.00 0.54 0.64 0.58 0.56
0 1 3 6 12
100 -0.10 0.22 0.40 0.42 0.51
75 -0.17 0.20 0.38 0.35 0.52
50 -0.17 0.16 0.36 0.37 0.49
25 -0.31 0.11 0.28 0.14 0.48
0 0.00 0.59 0.66 0.59 0.56
0 1 3 6 12
100 -0.11 0.22 0.38 0.38 0.50
75 -0.23 0.19 0.36 0.30 0.49
50 -0.26 0.03 0.35 0.20 0.45
25 -0.37 -0.26 0.25 0.04 0.17
0 0.00 0.59 0.66 0.59 0.56
0 1 3 6 12
Sellenbodenbach, Neuenkirch (n=51) Calibration period
100 0.37 0.46 0.50 0.45 0.40
75 0.40 0.48 0.49 0.45 0.42
50 0.28 0.38 0.41 0.44 0.43
25 -0.00 0.23 0.43 0.44 0.44
0 0.00 0.38 0.59 0.56 0.55
0 1 3 6 12
100 0.36 0.51 0.57 0.54 0.45
75 0.42 0.53 0.57 0.53 0.45
50 0.29 0.41 0.47 0.49 0.45
25 0.00 0.35 0.54 0.50 0.49
0 0.00 0.58 0.68 0.60 0.63
0 1 3 6 12
100 0.41 0.64 0.70 0.64 0.62
75 0.46 0.64 0.63 0.57 0.57
50 0.41 0.55 0.72 0.69 0.72
25 0.01 0.58 0.70 0.66 0.66
0 0.00 0.72 0.78 0.74 0.77
0 1 3 6 12
100 0.28 0.54 0.58 0.47 0.55
75 0.36 0.55 0.47 0.43 0.52
50 0.47 0.50 0.65 0.60 0.66
25 0.15 0.55 0.63 0.59 0.61
0 0.00 0.62 0.66 0.62 0.71
0 1 3 6 12
100 -0.07 0.15 0.45 0.30 0.50
75 -0.06 0.16 0.33 0.29 0.42
50 -0.04 0.22 0.52 0.51 0.57
25 0.06 0.29 0.55 0.54 0.60
0 0.00 0.34 0.58 0.59 0.69
0 1 3 6 12
Sellenbodenbach, Neuenkirch (n=51) Validation period
100 0.31 0.35 0.42 0.38 0.33
75 0.35 0.40 0.43 0.40 0.36
50 0.25 0.29 0.33 0.36 0.35
25 0.04 0.19 0.36 0.37 0.35
0 0.00 0.27 0.44 0.44 0.41
0 1 3 6 12
100 0.30 0.41 0.45 0.45 0.36
75 0.38 0.45 0.48 0.45 0.38
50 0.25 0.31 0.35 0.39 0.34
25 0.04 0.28 0.44 0.42 0.41
0 0.00 0.42 0.53 0.48 0.48
0 1 3 6 12
100 0.42 0.61 0.62 0.58 0.55
75 0.48 0.62 0.57 0.54 0.50
50 0.39 0.51 0.64 0.62 0.64
25 0.09 0.56 0.66 0.64 0.63
0 0.00 0.70 0.76 0.75 0.78
0 1 3 6 12
100 0.29 0.55 0.55 0.44 0.53
75 0.37 0.57 0.43 0.40 0.48
50 0.47 0.47 0.62 0.56 0.62
25 0.24 0.54 0.60 0.57 0.59
0 0.00 0.62 0.66 0.64 0.75
0 1 3 6 12
100 -0.10 0.24 0.43 0.24 0.48
75 -0.10 0.24 0.28 0.24 0.38
50 -0.05 0.24 0.52 0.48 0.54
25 0.02 0.31 0.53 0.52 0.58
0 0.00 0.32 0.58 0.60 0.73
0 1 3 6 12

Figure 53: Results from filters constraining the volume error to 2.5%, 5%, 20%, 30%, 50% for all study catchments. For each catchment, the first row shows the results for the calibration period, while the second row shows the results for the validation period. As in the heatmaps in the results section, the number of discharge measurements per hydrological year used for calibration are stated on the x-axis and the percentage of citizen science data used for calibration is stated on the y-axis.

11 Declaration of Independence

I hereby declare that the submitted thesis is the result of my own, independent work. All external sources are explicitly acknowledged in the thesis.

Zurich, June 2022

Franziska Schwarzenbach