AGU100 ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2019WR025248

Special Section:
Advancing process representation in hydrologic models: Integrating new concepts, knowledge, and data

Rainfall-Runoff Modeling Using Crowdsourced Water Level Data

B. Weeser[1,2,3], S. Jacobs[1,2], P. Kraft[1], M. C. Rufino[3,4], and L. Breuer[1,2]

[1]Institute for Landscape Ecology and Resources Management (ILR), Research Centre for BioSystems, Land Use and Nutrition (iFZ), Justus Liebig University Giessen, Giessen, Germany, [2]Centre for International Development and Environmental Research (ZEU), Justus Liebig University, Giessen, Giessen, Germany, [3]Centre for International Forestry Research (CIFOR), C/O World Agroforestry Centre, Nairobi, Kenya, [4]Lancaster Environment Centre, Lancaster University, Lancaster, UK

Key Points:
- Crowdsourced water levels were used to calibrate a conceptual rainfall-runoff model
- Combining crowdsourced water levels with a water balance derived from remotely sensed evapotranspiration increases the model efficiency
- Time-variable-collected water levels can be converted into a continuous discharge time series using a simple model approach

Correspondence to:
B. Weeser,
bjoern.weeser@envr.jlug.de

Citation:
Weeser, B., Jacobs, S., Kraft, P., Rufino, M. C., & Breuer, L. (2019). Rainfall-Runoff modeling using crowdsourced Water level data. *Water Resources Research*, *55*, 10,856–10,871. https://doi.org/10.1029/2019WR025248

Received 28 MAR 2019
Accepted 26 NOV 2019
Accepted article online 11 DEC 2019
Published online 17 DEC 2019

Abstract Complex and costly discharge measurements are usually required to calibrate hydrological models. In contrast, water level measurements are straightforward, and practitioners can collect them using a crowdsourcing approach. Here we report how crowdsourced water levels were used to calibrate a lumped hydrological model. Using six different calibration schemes based on discharge or crowdsourced water levels, we assessed the value of crowdsourced data for hydrological modeling. As a benchmark, we used estimated discharge from automatically measured water levels and identified 2,500 parameter sets that resulted in the highest Nash-Sutcliffe-Efficiencies in a Monte Carlo-based uncertainty framework (*Q-NSE*). Spearman-Rank-Coefficients between crowdsourced water levels and modeled discharge (*CS-SR*) or observed discharge and modeled discharge (*Q-SR*) were used as an alternative way to calibrate the model. Additionally, we applied a filtering scheme (*F*), where we removed parameter sets, which resulted in a runoff that did not agree with the water balance derived from measured precipitation and publicly available remotely sensed evapotranspiration data. For the *Q-NSE* scheme, we achieved a mean NSE of 0.88, while NSEs of 0.43 and 0.36 were found for *Q-SR* and *CS-SR*, respectively. Within the filter schemes, NSEs approached the values achieved with the discharge calibrated model ($Q\text{-}SR_F$ 0.7, $CS\text{-}SR_F$ 0.69). Similar results were found for the validation period with slightly better efficiencies. With this study we demonstrate how crowdsourced water levels can be effectively used to calibrate a rainfall-runoff model, making this modeling approach a potential tool for ungauged catchments.

1. Introduction

Increasing human population and climate change increase the pressure on water resources, make society more dependent on this resource, and require practitioners to be better prepared to manage this scarce resource more efficiently (Montanari et al., 2013; Rodda, 2001). Sustainable and effective water resource management decisions can only be made if reliable spatial and temporal water balance information is available. The performance of hydrological models, which can offer central support in the decision-making process, depends on sound hydrometeorological input (Wagner et al., 2009). In contrast, the engagement and investment of environmental agencies in hydrological and meteorological monitoring effort is decreasing worldwide (van de Giesen et al., 2014; Vörösmarty et al., 2001). Particularly large tropical basins are suffering from this decrease, many remain poorly gauged or were never gauged, often due to poor accessibility (Getirana et al., 2009). In addition, data restriction policies can lead to a delay on data release (Vörösmarty et al., 2001), which makes data use for immediate water resources management difficult especially when recent information is needed (Wagner et al., 2009). This data gap prevents the investigation of temporal and spatial changes of relevant parameters for water resources management, which are critical to support decision-making and the design of for example mitigation actions to prevent natural disasters (Davids et al., 2017).

Hydrological models can be used to investigate land use or climate change impacts on basins and to predict and assess the effects of management decisions on water resources. The level of complexity and the required amount of input data vary between different models. Nevertheless, all models need input and calibration data and require a monitoring network, which can be difficult and costly to establish and maintain. In the recent past, attempts have been made to obtain necessary data using novel ways. The increasing

availability of remotely sensed data provides scientists with some of the important water balance variables in regions where monitoring networks are scarce (Montanari et al., 2013). While remote sensing provides spatial data of variable resolution, hydrologists are still looking for ways to obtain direct hydrometric information such as on water levels or discharge at higher temporal and spatial resolution. So far, spaceborne remote sensing methods provide information like water level data in a sufficient resolution for large to medium-sized catchments (Yan et al., 2015), but these methods are still not operational for narrow rivers (<100-m width) (Bandini et al., 2017).

Besides remotely sensed data, crowdsource approaches have recently become attractive in research and capacity building campaigns from nongovernment institutions and agencies to fill hydrometeorological monitoring gaps (Davids et al., 2017; Walker et al., 2016). Data collected by citizens can help to create new hydrological knowledge and may support the efforts to identify the human impacts on the water cycle (Buytaert et al., 2014; Njue et al., 2019). The fast developments of communication technology will further increase the potential for citizen scientists to collect, submit, store, and process relevant data more easily (Buytaert et al., 2012; Montanari et al., 2013). In order to ensure a smooth and widespread implementation, the tasks assigned to citizens should be quick to perform and should not require special equipment. Davids et al. (2019) showed that undergraduate researchers can conduct discharge measurements using, among others, the salt dilution streamflow measurement method within reasonable ranges when compared against professional measurements. Other studies revealed that collecting simple parameters such as water levels is straightforward and that citizens can perform this task successfully (Fienen & Lowry, 2012; Weeser et al., 2018) and over long periods (Lowry et al., 2019).

Recent studies proved that water level instead of discharge data can be used for model calibration by using the monotonic relationship between water level and discharge mapped by the Spearman-Rank-Coefficient (Jian et al., 2017; Seibert & Vis, 2016; van Meerveld et al., 2017). This step avoids the need to convert water levels to discharge and potentially reduces the uncertainty introduced by this conversion (Jian et al., 2017). However, this step can also lead to a systematic bias since no information on the total water volume is taken into account. Therefore, there have been different attempts to include additional data during model calibration, for example, by filtering acceptable model parameters using annual streamflow volume (Seibert & Vis, 2016) or regionalized runoff coefficients from similar catchments (Jian et al., 2017). In this study, we tested a simple Water-Balance-Filter, which does not rely on any previous hydrometric information other than measured precipitation and actual evapotranspiration derived from MODIS (Moderate Resolution Imaging Spectroradiometer) data.

Using crowdsourced data for hydrological modeling is still in its infancy, and the value of this data source has not been comprehensively tested yet. Data collected by citizens differ from traditionally collected data in being irregular and of unknown quality and uncertainty. A few studies investigated the impact of these before-mentioned characteristics on the model calibration process using synthetic data sets derived from traditionally measured discharge (Mazzoleni et al., 2017; Mazzoleni et al., 2018), water levels (Seibert & Vis, 2016), or discharge combined with an error term generated from discharge estimates by citizens (Etter et al., 2018). However, none of these studies used real crowdsourced data.

Besides the potential use of crowdsourced data, involving the community brings additional benefits. Locals, who are supporting citizen science projects are more likely to protect environmental resources and participate in community services or sociopolitical activities (Overdevest et al., 2004). Linking this to the fact that especially low-income countries face pressing challenges in the water sector, it is attractive to test the integration of crowdsourced data for water resources management. To address this need, we established a comprehensive monitoring network based on crowdsourcing in the Sondu-Miriu River basin in Kenya in 2016 (Weeser et al., 2018). To date, the implementation of this approach has yielded more than 5,000 records of water levels.

This study aimed at rigorously testing the potential use of crowdsourced data for hydrological modeling, which could support the assessment of management practices in tropical environments. It was designed to answer the question of (1) whether water level data collected by citizen scientists are suitable for calibrating a rainfall-runoff model with an uncertainty similar to the uncertainty resulting from a calibration with conventional data sources and (2) if the model uncertainties can be reduced by using a simple to obtain Water-Balance-Filter as an additional criterion.

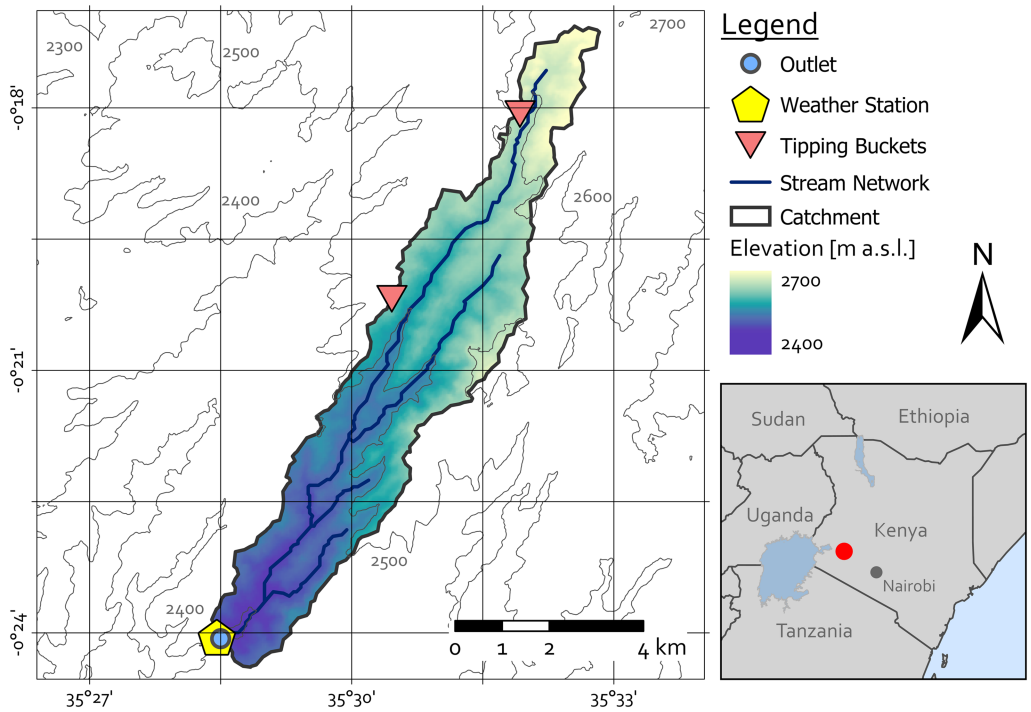

Figure 1. Location of the Sondu-Miriu-River Basin in Western Kenya (red dot in the overview map) and a map of the study area, including the stream network, outlet, weather station, and tipping buckets. The reference grid displays coordinates in WGS 1984.

2. Materials and Methods

2.1. Study Area

The study was conducted in a headwater catchment (27.4 km^2) in the northwestern part of the Sondu-Miriu-River Basin in Western Kenya (0.35°S, 35.5°E WGS1984) (Figure 1). Smallholder agriculture dominates the land use including annual crops, grazing lands, woodlands, and forests and led to a degraded land cover (Olang & Kundu, 2011). Increasing human population has resulted in rapid forest cover loss and forest degradation in the last decades (Brandt et al., 2018) and physical evidence has revealed a noticeable discharge decline for major rivers in the region (Olang & Kundu, 2011). The soils are in general deep and well drained classified as Humic Nitisols and Mollic Andosols (ISRIC - World Soil Information, 2007).

The climate is influenced by the Intertropical Convergence Zone, resulting in a bimodal rainfall pattern with a longer rainy season from April to July and a shorter rainy season between October and December (Figure 2). Temperature and precipitation (Table 1) were recorded by a weather station (ECRN-100 high-resolution rain gauge and VP-3 sensor, Decagon Devices, Pullman WA, USA) located 100 m northwest of the outlet measuring at a 10-min resolution. The installation of the instruments was carried out as far as possible according to the WMO guidelines, whereby local conditions had to be taken into account. The resolution of the rain gauge is 0.2 mm per tip, the accuracy of the VP-3 sensor depends on the temperature and humidity but lies in most cases within ~0.25 °C and 2–5% humidity. Precipitation was measured at two additional sites located in the center and the upper part of the catchment using tipping buckets (Theodor Friedrichs, Schenefeld, Germany). Thiessen-Polygons were used to calculate the area-weighted precipitation. If precipitation data gaps existed, the weights were adjusted by omitting the tipping bucket where no data was available. Data gaps in the temperature and precipitation time series were scarce (precipitation: tipping buckets <0.1%, precipitation from the weather station 5.5%; temperature: 7.2%) and filled with a linear interpolation after the data were aggregated to daily time steps. The yearly potential evapotranspiration (ET_{pot}) using grass as a reference crop was calculated based on the daily minimum, maximum and mean temperatures and the extraterrestrial radiation using the Hargreaves equation (Hargreaves & Samani,

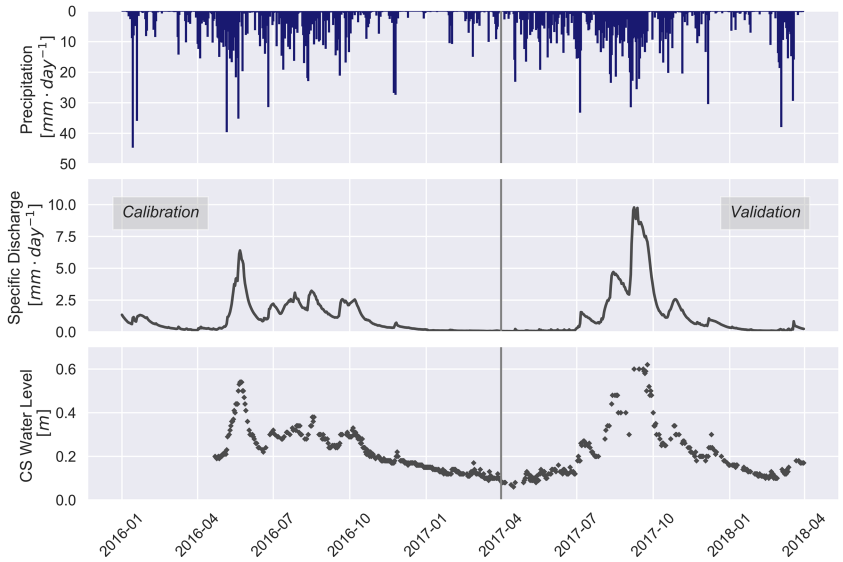

Figure 2. Daily mean areal weighted precipitation using Thiessen-Polygons (upper panel), specific discharge (middle panel), and crowdsourced (CS) reported water level data (bottom panel) of the catchment for January 2016 to April 2018.

1985). These values were in line with long term, altitude depending, ET_{pot} of 1,400 to 1,800 mm reported for this area (Krhoda, 1988; Nyenzi et al., 1981).

Water levels were measured in two different ways. A radar-based sensor (VEGAPULS WL61, VEGA Grieshaber KG, Schiltach, Germany) automatically collected water level data 20 m upstream of the outlet at 10-min intervals. Data collected by citizens (n = 271 during the calibration period) were recorded at the outlet (Weeser et al., 2018). The crowdsourcing-monitoring station was installed in April 2016 and equipped with a sign-board that explains to locals how they can participate in the monitoring. A small reward of 0.02 USD per measurement is paid automatically to compensate for the transmission costs. The coverage in the observation period was high with typically more than 16 observations per month covering 75% of all days during the calibration period. The data were not further filtered. Only one obvious outlier caused by a misinterpretation of the received text message was removed after checking the original text message associated with the doubtful data point. A comparison between the crowdsourced data and the automatically measured water levels showed a high agreement between both data types resulting in a Pearson correlation coefficient of 0.98 for the 271 measurements during the calibration period (Figure 3a). The high value of the correlation coefficient indicates that the crowdsourced data only differs slightly from professional measurements. Note that intercept and slope deviate from 0 and 1, respectively, due to the fact that readings from citizens were conducted 20 m upstream from where the professional reading took place resulting in slightly different cross sections.

A rating curve and the catchment area were used to convert the automatically measured water levels into daily specific discharge, which was the basis for model testing and evaluation (Figure 3b). To develop the rating curve (equation (1)), 86 manual discharge measurements using the salt dilution method (n = 82) and an Acoustic Doppler Current Profiler (RiverSurveyor S5, SonTek, San Diego CA, USA) (n = 4) over a

Table 1
Averaged Annual Hydrometeorological Data for the Study Area

Period	Specific Discharge (mm)	Precipitation (mm)	Mean daily temperature (°C)	ET_{pot} (mm)
01 April 2016 to 31 March 2017	413	1,287	14.9	1,596
01 April 2017 to 31 March 2018	485	1,557	14.4	1,522

Note. ET_{pot} = potential evapotranspiration calculated using the Hargreaves equation. Temperature represents the mean daily temperature measured at the weather station at the catchment outlet.

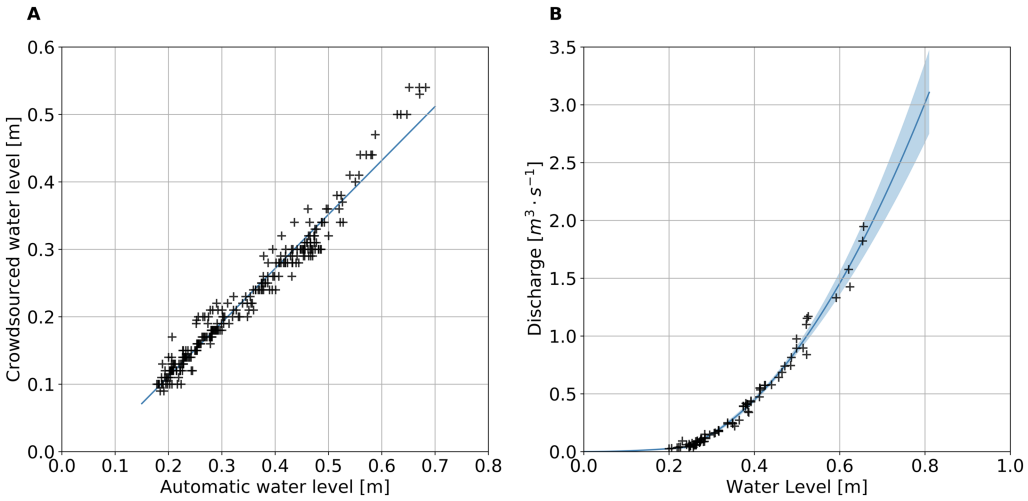

Figure 3. (a) Correlation between automatic water level measurements and crowdsourced water level data (Pearson correlation $r = 0.98$, $n = 271$) for the calibration period and (b) rating curve (solid blue line) with 95% confidence interval (blue shaded band) for the outlet of the study area based on 86 water level discharge pairs (black crosses).

wide range of water levels (h) were conducted. Extrapolation below the water level of 0.236 m was done using a quadratic function through the lowest measured discharge and zero discharge (Jacobs, Weeser, et al., 2018). For water levels above the highest measured water level used to develop the rating curve (0.66 m), we extrapolated the discharge using the same rating curve (3.3% of the time). To assess the discharge uncertainty we followed the procedure described in Jacobs, Weeser, et al. (2018), where the uncertainty was estimated based on the standard deviation (SD) of repeated measurements (SD water level: 1 mm, SD ADCP: 6.2%, SD Salt Dilution: 6.9%). We assumed that the true values were within 3*SD and generated 10,000 random samples for each water level/discharge combination. Figure 3b shows the uncertainty 95% confidence interval for the rating curve.

$$Q = \begin{cases} 0.0973 - 1.892 \cdot h + 6.923 \cdot h^2, & h \geq 0.236\ m \\ 0.651 \cdot h^2, & h < 0.236\ m \end{cases} \qquad R^2 = 0.98 \tag{1}$$

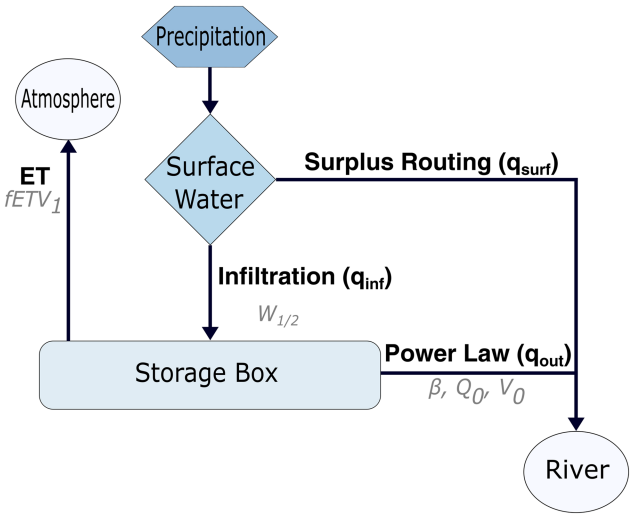

Figure 4. Schematic model structure. Catchment Modelling Framework processes are given in bold and their parameters in italic letters. Oval structures represent sinks, the hexagon an input flux, the box a storage, and the rhombus a distribution node without storage functionality.

2.2. Model Setup

We developed a lumped conceptual rainfall-runoff-model (Figure 4) using the Catchment Modelling Framework (Kraft et al., 2011) Version 1.4.1 (Kraft et al., 2018) for Python 3. Catchment Modelling Framework is a python based programing library to build hydrological models from building blocks (Jehn et al., 2018). This framework has been applied to a variety of catchments and is considered to describe the underlying hydrological processes sufficiently well (Jehn et al., 2018; Maier et al., 2017; Windhorst et al., 2014). The model structure represents the conceptual understanding of the rainfall-runoff processes reported by Jacobs, Timbe, et al. (2018). Daily precipitation and ET_{pot} are the only model inputs required.

In the model, water storage is represented as a single storage volume (V), which receives water from infiltration and loses water to the catchment outlet and actual evapotranspiration (equation (2)). Precipitation (P) is partitioned into infiltration (q_{inf}) and direct runoff (q_{surf}) by saturation excess (e_{sat}) (equation (3)). Since the soil saturation happens only in parts of the catchment area, depending on the stored volume, the saturated area is modeled with the Boltzmann sigmoidal function based on the soil water

Table 2
Model Parameters and A Priori Ranges

Name	Meaning	Unit	A priori parameter ranges	
			Min	Max
β	Kinematic flow curve shape exponent	—	1	6
Q_0	Reference runoff when storage contains the reference volume V_0	mm day^{-1}	0.01	1,000
V_o	Reference volume where storage runoff is equal to Q_0	mm	100	3,000
$fETV_1$	Scaling factor for potential evapotranspiration	—	0.01	0.8
$W_{1/2}$	Saturation, where half of the catchment area is saturated	—	0.1	0.9

storage and a parameter $W_{1/2}$. The parameter $W_{1/2}$ represents the saturation at which half of the incoming water infiltrates and the other half is routed to the outlet without time lag. Water in the storage box is released to the outlet using the normalized water volume raised to a power. This power law equation (equation (4)) determines the outflow (q_{out}) based on the actual water volume stored in the box (V) and the parameters V_0 (reference volume), Q_0 (outflow from the source when V equals V_0), and β (shape of the response curve). Water in the storage box is subject to evapotranspiration where the daily ET_{pot} is limited by the available water in the storage box. Actual evapotranspiration is assumed to be equal to the potential evapotranspiration if the water volume in the box is greater than the factor $fETV_1$ multiplied by the box capacity. If the volume is lower, the actual evapotranspiration is linearly scaled down to 0.

We used the implicit, error-controlled CVode solver (Hindmarsh et al., 2005) to integrate the differential equation (equation (2)) to prevent numerical problems (Kavetski & Clark, 2011). A priori model parameter ranges (Table 2) were chosen based on expert knowledge from previous applications of comparable model types and an exploratory analysis of a large parameter space.

$$\frac{dV}{dt} = q_{inf}(P,V) - q_{out}(V) - fET(V)\cdot ET_{pot}(t) \tag{2}$$

$$q_{surf}(V,P) = e_{sat}\left(VW_{\frac{1}{2}}\right)Pq_{inf} = P - q_{surf} = \left(1 - e_{sat}\left(VW_{1/2}\right)\right) \tag{3}$$

$$q_{out} = Q_0\left(\frac{V}{V_0}\right)^{\beta} \tag{4}$$

2.3. Model Calibration and Validation

The time series were split-up in a warm-up period (1 January 2016 to 31 March 2016), a calibration period (1 April 2016 to 31 March 2017), and a validation period (1 April 2017 to 31 March 2018). We followed a Monte Carlo-based calibration approach and quantified the model parameter uncertainty using the open-source python package SPOTPY (Houska et al., 2015). We evaluated the calibration efficiency using two objective functions, that is, Nash-Sutcliffe-Efficiency (NSE, equation (5), where e_i is the ith observation, s_i is the ith simulation, and $\bar{e}$ is the mean of the observations) (Nash & Sutcliffe, 1970) and percent bias (PBIAS; equation (6)). While the NSE is mainly influenced by peaks and therefore ensures an acceptable model fit during high flow conditions, the PBIAS indicates the tendency of overestimation or underestimation of the discharge through the model over the full period. In total, 10^6 parameter sets were generated for the calibration process within predefined (a priori) parameter ranges (Table 2). Instead of sampling the entire parameter space, we used Latin Hypercube Sampling (McKay et al., 1979).

$$NSE = 1 - \frac{\sum_{i=1}^{N}(e_i - s_i)^2}{\sum_{i=1}^{N}(e_i - \bar{e})^2} \tag{5}$$

$$PBIAS = 100^* \frac{\sum_{i=1}^{N}(e_i - s_i)}{\sum_{i=1}^{N}(e_i)} \tag{6}$$

2.3.1. Using the Spearman-Rank-Coefficient to Calibrate on Water Level Data

To calibrate the model on water level data, we took advantage of the fact that water levels are dynamically linked to discharge variation and that they can, therefore, be compared against modeled discharge by using

the Spearman rank correlation coefficient (R_{Spear}) (Seibert & Vis, 2016). Ranging from −1 to 1, an R_{Spear} close to 1 indicates that the simulated discharge and the measured water levels reproduce the same dynamics and that the water level and discharge values are strictly monotonically related (Seibert & Vis, 2016). The R_{Spear} is not affected if the data is transformed using a strictly monotonically increasing or decreasing function as done by the rating curve in this study. In this case, the R_{Spear} values will be similar regardless if the automatically measured water level data (which was converted into discharge data using the rating curve) or the discharge data itself is used. Consequently, we do not show a calibration based on the automatically measured water level data since the results are the same as obtained from a discharge-based calibration.

Since the R_{spear} only reflects the similarity of the dynamics between the observed discharge and water level data and does not reflect the absolute volumes, a value of 1 does not ensure a perfect fit (Seibert & Vis, 2016). Therefore, a threshold for behavioral parameter sets cannot be defined similarly to a calibration based on objective functions like the NSE. Instead, we propose to select behavioral parameter sets by ranking all model runs by their associated R_{Spear} value and take the top set. In this study, we defined behavioral parameter sets by taking the best 0.25% of the 10^6 runs, resulting in 2,500 parameter sets. The same procedure of taking the best 0.25% was applied when the model was calibrated on discharge data and the NSE to ensure the comparability of the different calibration schemes.

2.3.2. Water-Balance-Filter

As stated above, utilizing water level readings for calibrating a model to calculate discharge can lead to overestimation or underestimation. Therefore, we tested a simple annual Water-Balance-Filter to obtain acceptable model outputs by selecting only those parameters sets where model runs resulted in a high R_{Spear} and additionally matched a simplified water balance. The annual water balance was calculated from observed precipitation minus mean actual evapotranspiration (ET_{act}). ET_{act} can be retrieved from spaceborne remote sensing data sets obtained from the MODIS. For the study area, a mean ET_{act} of 1,055 mm yr^{-1} was derived from data provided by the MOD16A2 Collection 6 Global Evapotranspiration Product from MODIS imagery based on land surface temperature and albedo and the Penman-Monteith equation (Running et al., 2017) for the 2-yr simulation period. These values are close to the estimation from our measured data when we subtract runoff from precipitation assuming that possible storage changes can be considered small enough to be ignored for the 2-yr period and the remaining water consequently represents the ET_{act} (Senay et al., 2011). From our measured data we derived an ET_{act} of 973 mm on average for the 2 yr, which is 7.7% less than the value determined using the MODIS data set. The MOD16A2 Collection 6 data set, obtained from the satellite Terra, contains composite evapotranspiration data with 500-m pixel resolution for 8-day periods. In order to calculate the annual ET_{act} each satellite image was cropped to the catchment area, fill values without calculated ET were set to unavailable, the result was multiplied by 0.1 (scale factor after Running et al., 2017) and a mean value for the catchment area was calculated. In order to determine the annual value, all individual values were summed up. To compensate measurement errors, unknown uncertainties and possible storage changes we added a (subjective) confidence interval of ±30%, resulting in an ET_{act} between 738 and 1,371 mm yr^{-1} (ET_{act}/ET_{pot}-ratio = 48–88%) for the study area. This value is in line with a study of Velpuri et al. (2013), which reported mean uncertainties up to 25% for MOD16 data sets at basin scale. Given the average annual precipitation over the 2-yr observation period of 1,422 mm (Table 1), model runs were discarded if the simulated specific discharge was >684 or <51 mm yr^{-1}.

2.3.3. Calibration Schemes

Six independent calibration schemes were carried out to evaluate the value of crowdsourced water level data for model calibration. As a benchmark, we first calibrated the model on the discharge data using both the Nash-Sutcliffe-Efficiency and the Spearman-Rank coefficient (schemes *Q-NSE* and *Q-SR*). After that, we calibrated the model on the crowdsourced water level measurements and did not consider any automatically measured water level or discharge data (*CS-SR*). Finally, all accepted parameters from the different calibration schemes were filtered using the Water-Balance-Filter ($Q\text{-}NSE_F$, $Q\text{-}SR_F$, $CS\text{-}SR_F$).

2.4. Model Comparison (Benchmark)

The model was validated by conducting runs for the validation period using the a posteriori parameter sets, comparing the modeled with observed discharge. To compare the model efficiencies between the different calibration schemes, we defined a lower benchmark (R_{lower}) following an approach described by Seibert et al. (2018). For this, we run the model 2,500 times with random parameter sets within the a priori

Table 3
Relative Performance and Model Efficiency Measures Nash-Sutcliffe-Efficiency (NSE) and Percent Bias (PBIAS) During Calibration and Validation of the Different Calibration Schemes Using Discharge Observations (Q) and the Crowdsourced Data (CS) Without and With a Water-Balance-Filter (Filter) for the Best 0.25% of All 10^6 Model Runs Calibrated on the NSE or R_{Spear}

					NSE						PBIAS					
					Calibration			Validation			Calibration			Validation		
Data set	Calibrated with	Filter	ID	*n*-Runs	mean (-)	best (-)	$R_{relative}$ (%)	mean (-)	best (-)	$R_{relative}$ (%)	mean (%)	range (%)	$R_{relative}$ (%)	mean (%)	range (%)	$R_{relative}$ (%)
Q	NSE	No	Q-NSE	2500	0.88	0.91	96.6	0.86	0.93	91.3	−0.88	[−23,16]	99.1	6.88	[−17,29]	93.8
		Yes	Q-NSE$_F$	2500	0.88	0.91	96.6	0.86	0.93	91.3	−0.88	[−23,16]	99.1	6.88	[−17,29]	93.8
	R_{Spear}	No	Q-SR	2500	0.43	0.91	66.4	0.69	0.93	70.0	51.95	[−36,133]	46.7	51.8	[−28,76]	53.4
		Yes	Q-SR$_F$	1539	0.70	0.91	84.6	0.80	0.93	83.8	28.48	[−36,65]	70.8	30.4	[−28,76]	72.7
CS	R_{Spear}	No	CS-SR	2500	0.36	0.91	61.7	0.70	0.93	71.3	58.27	[−30,142]	40.2	53.9	[−27,124]	51.6
		Yes	CS-SR$_F$	1408	0.69	0.91	83.9	0.82	0.93	86.3	32.5	[−30,65]	66.7	30.9	[−27,70]	72.2

Note. Heat map indicates best (green) to worst (yellow) model performance.

parameter ranges (Table 2). From these 2,500 model runs, a mean discharge time series was calculated and compared against the observed discharge for both, the calibration and validation period. The upper benchmark (R_{upper}) was defined as the best efficiency obtained during the discharge-based calibration assuming that this value reflects the best possible calibration of the model for the given data set. The relative performance ($R_{Relative}$) of each calibration scheme can then be determined following equation (7), whereby R_x indicates the performance reached for each individual calibration scheme

$$R_{Relative} = \frac{R_x - R_{lower}}{R_{upper} - R_{lower}} \quad (7)$$

3. Results

3.1. Lower and Upper Benchmark

To compare the efficiencies of the different model calibration scenarios a lower benchmark was defined by randomly selecting 2,500 model runs and calculating a mean discharge time series, which was compared against the observed discharge values. A NSE of −0.56 and a PBIAS of 97.51% was found as a lower benchmark for the calibration period. For the validation using the same 2,500 random parameter sets, the NSE was 0.13 and the PBIAS 111.31%. The highest performance measure within the *Q-NSE* scheme defined the upper benchmark, resulting in an upper NSE benchmark of 0.93 and a PBIAS of 0%. All schemes resulted in at least one parameter set with similar best performance measures for all schemes.

3.2. Discharge-Based Calibration (Q-NSE and Q-SR)

The model simulated observed discharge reasonably well when calibrated against discharge using the Q-NSE scheme (Table 3). Under this scheme the model achieved a mean NSE of 0.88 and a relative NSE performance for the mean of all runs of 96.6% (relative performance of PBIAS 99.1%) when compared against the upper and lower benchmark of the *Q-NSE* scheme. The parameter sets which achieved the best 0.25% (equals 2500) NSE values (*Q-NSE*) or R_{Spear} values (*Q-SR*) were considered as behavioral and were accepted for further analysis. When testing the behavioral parameter sets of *Q-NSE* against the validation time series the model performance was only marginally lower achieving a mean NSE of 0.86 and a relative NSE performance 91.3% (relative performance of PBIAS 93.8%).

Calibrated on discharge but using 0.25% of all parameter sets with the highest R_{Spear} instead of NSE (*Q-SR*) the model performance decreased achieving a mean NSE of 0.43 and a relative performance of 66.4%. The mean PBIAS increased from –0.88% (*Q-NSE*) to 52% during calibration. The same trend was followed during validation with similar performance measures.

3.3. Crowdsourced Calibration (CS-SR)

The model predicted the observed discharge within acceptable ranges when calibrated and validated against the crowdsourced water level data without applying the Water-Balance-Filter. The mean NSE performance decreased by 34.9% during calibration in comparison to the *Q-NSE* scheme to similar values than the ones achieved with the *Q-SR* calibration scheme. However, the *CS-SR* scheme outperformed the lower benchmark

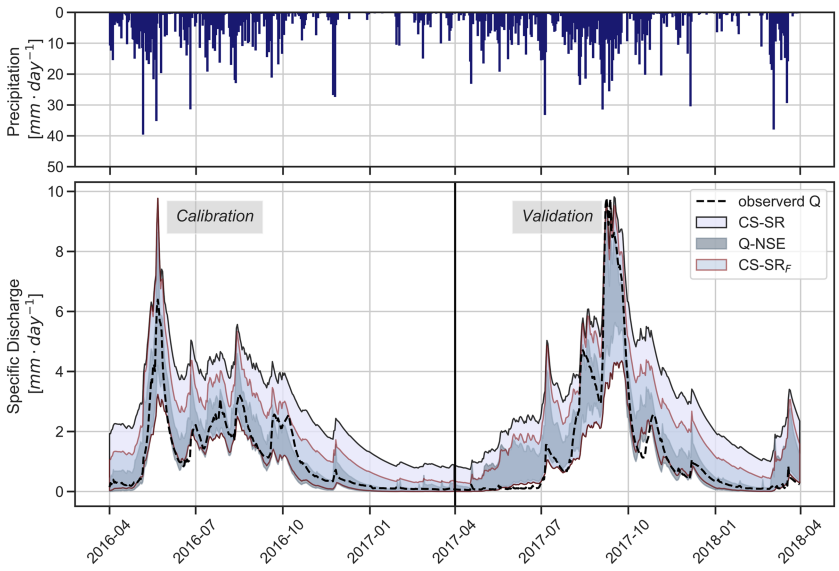

Figure 5. Observed precipitation (top) and discharge (black dashed line in the lower box) in the study area from April 2016 to March 2018. Simulated discharge for three different calibration schemes during calibration and validation (from light blue to dark blue: CS-SR, CS-SR$_F$, and Q-NSE), where Q-NSE indicates a traditional calibration against observed discharge data, CS-SR a calibration against 2,500 runs with the highest Spearman-Rank-Coefficient when calibrated against the crowdsourced water level data and CS-SR$_F$ a calibration using the same runs obtained from CS-SR but filtered for a maximum yearly runoff based on an estimated water balance using observed precipitation and actual evapotranspiration derived from MODIS.

model. The PBIAS revealed a decrease of the relative performance of 58.9% in relation to the relative performance of the PBIAS during calibration for the *Q-NSE* scheme. A comparable decrease could be observed during validation. Since the mean PBIAS is >0 in all cases, the CS-SR schema tends to overestimate the overall discharge.

3.4. Water-Balance-Filter Effects on the Calibration (Q-NSE$_F$, Q-SR$_F$, CS-SR$_F$)

No differences were observed between the *Q-NSE* and the *Q-NSE$_F$* scheme since all accepted parameter sets within the *Q-NSE* scheme already matched the water balance and subsequently no parameter set was discarded. For all R_{Spear}-based calibration schemes, the filter improved the model performance notably. This holds regardless of the data set used for both the discharge-based calibration (*Q-SR*) and the crowdsourced water level data calibration (*CS-SR*). The relative performance for these calibration schemes increased to comparable values between 84% and 86% during calibration and validation for NSE and between 66% and 72% for PBIAS. Hence, calibrated with crowdsourced water level data combined with the Water-Balance-Filter (*CS-SR$_F$*), the model predicted the discharge almost as well as if calibrated on the observed discharge (*Q-NSE*). This applies for the behavior of both model efficiency measures, the NSE and the PBIAS.

Figure 5 shows the modeled discharge time series during calibration and validation for the *Q-NSE* scheme and the crowdsourced-based calibration scheme (*CS-SR* and *CS-SR$_F$*). This figure underlines the similarities and differences between the different calibration methods. In general, all calibration schemes tended to slightly overestimate base flow conditions. Remarkably, all schemes resulted in similar lower discharge bands and only the upper discharge band deviated for the scenario *CS-SR* compared to the scenarios *CS-SR$_F$* and *Q-NSE,* which was also reflected in the PBIAS.

3.5. Comparison of Different Calibration Schemes

We analyzed specific flux components simulated by the model to further understand and evaluate the model behavior regarding the different calibration schemes. This allowed us to assess whether the simulated processes are within realistic boundaries and whether the different calibration schemes influence the hydrological fluxes. A large discrepancy between the individual fluxes would be questionable and indicate a mismatch between the model simulations and the underlying processes. The same applies to abnormally

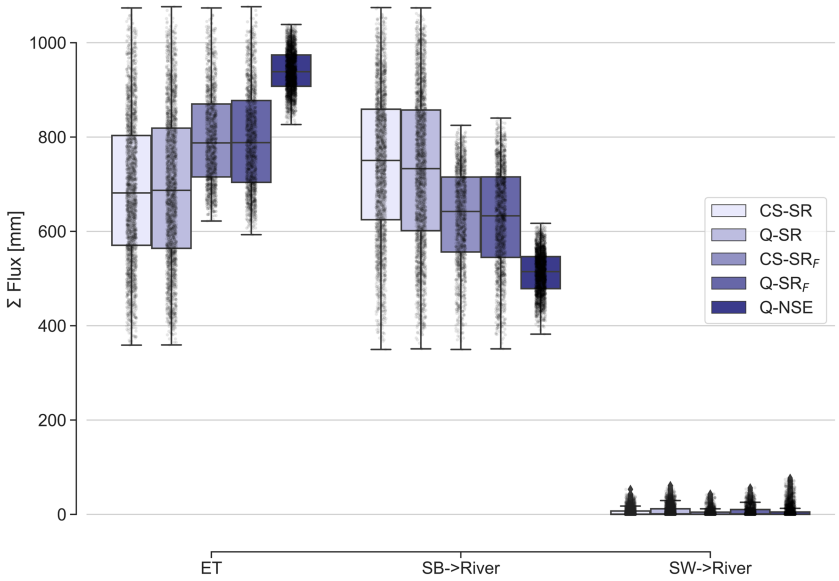

Figure 6. Boxplots of the sum of fluxes released by the different model components (ET = actual Evapotranspiration, SB-> River = Water released from the Storage Box to the Outlet, SW-> River = Water released from the Surface Water Storage to the Outlet) under different model calibration schemes (CS-SR = calibration based on crowdsourced water level data; Q-SR = calibration based on discharge and the Spearman-Rank-Coefficient; CS-SR$_F$ = calibration based on crowdsourced data in combination with a Water-Balance-Filter; Q-SR$_F$ = calibration based on discharge and the Spearman-Rank-Coefficient data in combination with a Water-Balance-Filter; Q-NSE = traditional calibration process based on discharge data and the Nash-Sutcliffe model efficiency coefficient) during the validation period.

large or small values for the actual evapotranspiration. In addition, the analysis provides insights into the range of the simulated flows under the various calibration schemes and thus into the related model uncertainties. Figure 6 shows the distribution of the sums of each flux for every model run within the calibration schemes for the validation period (the figure for the calibration is similar and not shown) excluding the Q-NSE$_F$ scheme, because of its redundancy to the Q-NSE scheme. The results reveal an equal distribution of the modeled flux components for all five calibration schemes. The variability in fluxes is smallest for the *Q-NSE* scheme and increases for the filtered (*Q-SR$_F$, CS-SR$_F$*) and unfiltered (*Q-SR, CS-SR)* schemes. For example, the range of simulated ET values under *Q-SR* was largest (359–1,076 mm), and the contribution to the total water balance was on average (mean 693 mm) lower than for *Q-NSE* (mean 940 mm). Consequently, more water left the system from the storage box to the outlet in the *Q-SR* scheme compared to *Q-NSE*. This can also be seen in the time series where the *Q-SR* scheme (similar to the *CS-SR* scheme) tends to overestimate the flow (Figure 5). The distributions within the unfiltered or filtered calibration schemes are comparable. Consequently, the R$_{Spear}$ calibrated data sets show a similar distribution regardless of whether they were calibrated to the discharge or the citizen-based water levels. The proportion of surface runoff (SW) was low for all three methods. This is in line with the general process understanding for this catchment and its environmental conditions (Jacobs, Timbe, et al., 2018). Surface runoff can occur during heavy rain events but remains low. A high fraction of surface runoff would, therefore, not be realistic.

4. Discussion

To date, integrating crowdsourced data into hydrological monitoring remains uncommon with only a few studies that examined the value of these data for model calibration (Etter et al., 2018; Mazzoleni et al., 2017; Starkey et al., 2017). All these studies, however, lack a direct comparison between a model driven by real-world crowdsourced data and a model calibrated using conventionally collected discharge measurements. Some of the studies used synthetically generated data to mimic crowdsourcing. In contrast, our study compares the efficiencies of a model calibrated on crowdsourced water levels, combined with or without a Water-Balance-Filter, against results from a model calibrated on measured daily discharge. Projects like *CrowdHydrology*, which resulted in more than 16,000 observations since 2010 ongoing, proved that it is

possible to gather this type of data, also over extended periods (Lowry et al., 2019). We found that the model calibrated with crowdsourced water levels combined with a Water-Balance-Filter performs similarly in terms of model efficiencies but results in greater model uncertainties and overestimations of discharge.

4.1. Assessing Model Performance Through Spearman-Rank-Coefficient

The Spearman-Rank-Coefficient (R_{Spear}) used to identify behavioral parameter sets during calibration allowed a comparison of the dynamics of the simulated discharge and measured water levels but revealed no information on total discharge volumes (Jian et al., 2015). There are no previous reports as to which parameter sets should be declared as behavioral when using the R_{Spear} as an objective function. Consequently, we ranked all parameter sets by their R_{Spear} value and chose a certain percentage (0.25% of 10^6 runs) as behavioral. An advantage of using the R_{Spear} is that for example no extra Inverse Rating Curve function with additional uncertainties needs to be estimated (Jian et al., 2015) and that water levels can directly be used to calibrate the model. This avoids the tedious process of discharge measurements, which also requires special equipment and expert knowledge, often not available particularly in remote places in low-income countries. This might change in the future when new methods to determine discharge like particle-image-velocimetry (Adrian, 1991) produce sound results, especially when such approaches become operational on consumer electronic devices like smartphones as shown by Lüthi et al. (2014). We believe, however, that water level measurements carried out by citizen scientists remains easier, reduces measurement errors, requires no expert knowledge, delivers reliable measurements, and can promote community participation (Lowry & Fienen, 2013; Weeser et al., 2018).

4.2. The Value of a Water-Balance-Filter

As stated above, a risk exists that a hydrological model might be biased when only calibrated with water level data. Seibert and Vis (2016) addressed this issue when calibrating a model for more than 600 catchments in the United States using daily water level and discharge data. Their study revealed that models that were calibrated on water levels performed well in wet catchments where the precipitation input was higher than the potential evapotranspiration. Seibert and Vis (2016) related this to the fact that the actual evapotranspiration in these catchments was close to the potential evapotranspiration which diminished the influence of different parameter sets on this term of the water balance. Our results confirmed these findings by showing acceptable results in a catchment with precipitation values close to potential evapotranspiration. At the same time, these results indicate that a more intense testing of the approach under different environmental conditions is needed. The hydrological behavior of different catchments might or might not have a further impact on the transferability of our approach, which we finally cannot decide based on a single catchment study. Seibert and Vis (2016) indicated that some volume information might improve the results for drier catchments and the authors stressed the need for further research on this field. In our study, we tested the added value of a Water-Balance-Filter on the parameter set selection to reduce the risk of selecting parameter sets that result in biased model calibration. However, we have to point out that the uncertainty of the actual evapotranspiration derived from the MODIS data set cannot be determined precisely since it depends on various local factors. Mu et al. (2011) identified uncertainties in the used algorithm input data (such as the daily meteorological data), inaccuracy of the measured eddy covariance flux tower data, the scaling from the flux tower point measurements to the landscape and algorithm limitations as main factors, which influence the bias between estimated and measured ET_{act}. When we compared the derived ET_{act} from MODIS with our measured precipitation minus the measured discharge and neglected storage changes (Senay et al., 2011), we found an overestimation for the remotely sensed actual evapotranspiration of 7.7%. After applying the uncertainty compensation of ±30%, the resulting Water-Balance-Filter range falls within the measured ET_{act} value. Consequently, our *CS-SR*$_F$ and *Q-SR*$_F$ results showed that the model efficiencies improved when those parameter sets, which were selected as behavioral in the first step using the R_{Spear}, were further filtered. The filter effectively removed model runs that resulted in a discharge overestimation. In contrast to that, these runs were accepted within the unfiltered R_{Spear}-based calibration schemes (*CS-SR*, *Q-SR*) since no volume information was considered.

All schemes resulted in fluxes that were in line with the general process representation. The analysis of the individual fluxes showed that the different schemes did not change the general process understanding of the model. Evapotranspiration was calibrated differently, which resulted in more water draining into the river in

the crowdsource-based model schemes. Having in mind that the approach should be applicable under remote conditions or in understudied catchments, we developed a filter that can be easily derived from publicly available data sources rather than aiming for a high precision of the filter itself. The uncertainty factor (30%) we used to define the Water-Balance-Filter based on the measured precipitation and remotely sensed evapotranspiration might deviate for other input data or catchments. We, therefore, argue that a wide range should be chosen. Since the filter only reduces the previously selected parameter sets but does not affect the calibration process itself, the filter has no negative influence on the results. Our results show that including such a simple filter in the a posteriori model selection process reduces effectively the bias that is inherent when calibrating the model using R_{Spear} as an objective function.

In general, the increasing availability of remotely sensed data brings new opportunities to obtain relevant water balance variables, particularly in regions where in situ monitoring networks are sparse (Montanari et al., 2013), although the spatial resolution is coarse and ground-truthing often is required. The sparse repeat cycle of satellite data hampers the measurement of daily or weekly changes further (Jian et al., 2017) making it impossible to detect or quantify short events which are typical for tropical catchments. Therefore, the combination of crowdsourced observations with remotely sensed data could be a way to support hydrological modeling in areas where no or only limited hydrometric information is available.

4.3. The Role of Input Data and Innovative Input Data Sources

Besides water levels, we used precipitation and temperature-based calculated evapotranspiration as inputs for our model. The quality and resolution of these data influence the model performance. We used precipitation and temperature data from automatic meteorological stations, with a controlled quality to demonstrate the feasibility of calibrating a model using crowdsourced water levels. However, these data might not be available in all cases and can become an additional error source. For larger catchments, where the spatial resolution might be less important which can lead to smoothing effects, these data could be derived from remote sensing or interpolated using measurements from existing meteorological stations. Beyond that, it is possible that precipitation and temperature measurements are performed by citizen scientists. Starting in 1998 as a local project the CoCoraHS (the Community Collaborative Rain, Hail, and Snow network) became the largest provider of daily manual rainfall measurements in the United States with 37,500 participants and over 31 million crowdsourced daily precipitation reports (Reges et al., 2016). A study by Walker et al. (2016) showed that an Ethiopian community monitored precipitation sufficiently for 18 months resulting in a high correlation between the crowdsourced data and data from a national station. By using a community-based rain gauge network, a high spatial resolution might compensate a potentially lower data precision since local rain events can be captured, which cannot be detected by coarser professional networks (Kirchner, 2006). Technical development opens the potential for new and alternative data collection methods which could contribute to improved availability of data. Overeem et al. (2013), for example, showed the possibility to estimate daily mean air temperatures from smartphone battery temperatures, while Messer et al. (2006) described a method how the signal levels of cellular networks can represent precipitation amounts. Gosset et al. (2016) claimed that this technique is particularly suitable for areas that lack precipitation measurement infrastructure including large parts in Africa. Linking data from different and innovative methods together may have great potential for hydrological modeling.

4.4. Model Structure and Data Resolution

The conceptual model used in this study involved only five parameters, which allowed a consistent calibration and avoided over-parameterization (Kirchner, 2006). Furthermore, since few parameters are involved, the model can be easily applied in data scarce regions. However, a more complex physically based and/or spatially distributed model might have benefits by providing the opportunity to use observed data from various sources and locations and integrate them into the model approach (Starkey et al., 2017). Mazzoleni et al. (2017) demonstrated the use of synthetically generated crowdsourced streamflow observations in a spatially distributed model to improve flood predictions. These authors showed that the temporal variability of data influenced the results less than their accuracy, which confirms the usefulness of crowdsourced data given that their accuracy is assured. However, even the resolution of the water level scale (vertical resolution) is not an exclusion criterion. For example, van Meerveld et al. (2017) demonstrated that the vertical resolution of water level measurements is less critical. These authors used a time series of only two stream level classes to calibrate a conceptual model successfully. These findings may further increase the applicability of

crowdsourced data as it allows the use of data with reduced vertical resolution and hence reduced accuracy and temporal resolution. A study by Seibert et al. (2019) showed that virtual water level gauges, generated by a mobile application, can be used to monitor water levels in any stream without physical installations, which can make the approach scalable. These results indicate a promising way to increase the spatial coverage of crowdsourced measurements in future.

4.5. Crowdsourced Versus a Discharge-Based Calibration

The often expressed concern that data irregularity induces problems can therewith be mitigated. Our study confirms this assumption since no evidence was found that data irregularity within the crowdsourced data affected the model performance and the model could be calibrated using the crowdsourced data which had a variable temporal resolution and measurement uncertainty. The crowdsourced-based calibration schemes led to comparable results as the discharge-based calibration when using the R_{Spear} performance measure. The increased uncertainty is therefore mainly induced by using the R_{Spear} and only marginally by the crowdsourced data itself. The crowdsourced data only led to a decrease of the relative performance of around 5% for both the NSE and PBIAS during calibration (*CS-SR*) in comparison to the discharge-based calibration (*Q-SR*). Compared to the NSE-based calibration (*Q-NSE*) the relative performance decreased by 30–35% under the R_{Spear}-based schemes regardless of the model was calibrated on discharge or crowdsourced water level data (*Q-SR* and *CS-SR*).

5. Conclusions

Based on our results, we suggest crowdsourced monitoring approaches as an additional tool for water resources management, particularly in ungauged or poorly gauged catchments and under limited financial resources. Combining simple measurement carried out by citizen scientists with a modeling approach could be a way to improve our knowledge of available water resources and process understanding in catchments that have so far been understudied. This approach may be an alternative in places where observational gaps are caused by a lack of hydrometeorological gauging networks. However, some limitations are worth noting. Although our findings provide evidence that crowdsourced water levels can be used to calibrate hydrological models, the outcome also depends on the quality of other input data and the general catchment behavior. Our study area only had a few flooding events and the water balance seems to be fairly simple. Consequently, the observed discharge could be modeled well using a simple model structure. The crowdsourced data we used in this study are from outstanding quality with high temporal coverage and low measurement errors. Future work should, therefore, investigate the behavior of crowdsourced calibrated models for catchments of different land use and climatic conditions, test the implementation of crowdsourced climate data and investigate the impact of crowdsourced data of various quality.

Based on our evidence, we provide the following answers to the research questions raised in section 1:

1. Are water levels collected by citizen scientists suitable for calibrating a rainfall-runoff model with an uncertainty similar to the uncertainty resulting from a calibration with conventional data sources?

A conceptual rainfall-runoff model can be calibrated on crowdsourced water level data. The combination of crowdsourced data and a rainfall-runoff-model might solve an often raised critical point when using crowdsourcing in hydrology, that is, data irregularity. After a 1-yr calibration, the model transforms the community-based collected data into a continuous time series. This is particularly valuable when one considers that only water levels were used for calibration and no discharge measurements had to be carried out, which would have required special equipment and training of the citizen scientists. The model, which was only calibrated against water level data, predicts the observed discharge in acceptable ranges, but the efficiencies were lower than the efficiencies of a model that was calibrated on conventional discharge data. However, the lower efficiencies are mainly introduced by using the R_{Spear} as a performance measure, which leads to an overestimation of the discharge.

2. Can the model uncertainties be reduced by using a simple to obtain Water-Balance-Filter as an additional criterion?

By applying a simple Water-Balance-Filter it was possible to achieve model efficiencies similar to those obtained from traditional calibration against streamflow. We used a parsimonious water balance derived

from measured precipitation and remotely sensed evapotranspiration data, avoiding data-intense estimates. Similar water balances can be established in other data-sparse regions. Combining the filter with the rainfall-runoff model increased the model reliability. The filter can compensate the effect of keeping parameter sets that result in unrealistic high or low volumes, which can occur when using the Spearman-Rank-Correlation as an objective function. To achieve this effect, the uncertainty of the remotely sensed actual evapotranspiration data should not have led to values that are too far from the real actual evapotranspiration in the respective catchment.

Acknowledgments
All data and scripts (model and data analysis) used in the study can be obtained from Zenodo (DOI: 10.5281/zenodo.3341592). We would like to express our thanks to the citizens of the Sondu-Miriu River basin who participated in our program. We also thank the Deutsche Forschungsgemeinschaft DFG (BR2238/23-1), the Deutsche Gesellschaft für Internationale Zusammenarbeit GIZ, and the German Federal Ministry for Economic Cooperation and Development (Grants 81195001 "Low Cost methods for monitoring water quality to inform upscaling of sustainable water management in forested landscapes in Kenya" and 81206682 "The Water Towers of East Africa: policies and practices for enhancing cobenefits from joint forest and water conservation") for generously providing financial support. This work was partially funded by the CGIAR program on Forest, Trees and Agroforestry led by the Centre for International Forestry Research (CIFOR).

References

Adrian, R. J. (1991). Particle-imaging techniques for experimental fluid mechanics. *Annual Review of Fluid Mechanics*, *23*(1), 261–304. https://doi.org/10.1146/annurev.fl.23.010191.001401

Bandini, F., Butts, M., Jacobsen, T. V., & Bauer-Gottwein, P. (2017). Water level observations from unmanned aerial vehicles for improving estimates of surface water-groundwater interaction. *Hydrological Processes*, *31*(24), 4371–4383. https://doi.org/10.1002/hyp.11366

Brandt, P., Hamunyela, E., Herold, M., de Bruin, S., Verbesselt, J., & Rufino, M. C. (2018). Sustainable intensification of dairy production can reduce forest disturbance in Kenyan montane forests. *Agriculture, Ecosystems & Environment*, *265*, 307–319. https://doi.org/10.1016/j.agee.2018.06.011

Buytaert, W., Baez, S., Bustamante, M., & Dewulf, A. (2012). Web-based environmental simulation: Bridging the gap between scientific modeling and decision-making. *Environmental Science & Technology*, *46*(4), 1971–1976. https://doi.org/10.1021/es2031278

Buytaert, W., Zulkafli, Z., Grainger, S., Acosta, L., Alemie, T. C., Bastiaensen, 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*. https://doi.org/10.3389/feart.2014.00026

Davids, J. C., Rutten, M. M., Pandey, A., Devkota, N., van Oyen, W. D., 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

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

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

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

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

Gosset, M., Kunstmann, H., Zougmore, F., Cazenave, F., Leijnse, H., Uijlenhoet, R., et al. (2016). Improving rainfall measurement in gauge poor regions thanks to mobile telecommunication networks. *Bulletin of the American Meteorological Society*, *97*(3), ES49–ES51. https://doi.org/10.1175/BAMS-D-15-00164.1

Hargreaves, G. H., & Samani, Z. A. (1985). Reference crop evapotranspiration from temperature. *Applied Engineering in Agriculture*, *1*(2), 96–99. https://doi.org/10.13031/2013.26773

Hindmarsh, A. C., Brown, P. N., Grant, K. E., Lee, S. L., Serban, R., Shumaker, D. E., & Woodward, C. S. (2005). SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers. *ACM Transactions on Mathematical Software*, *31*(3), 363–396. https://doi.org/10.1145/1089014.1089020

Houska, T., Kraft, P., Chamorro-Chavez, A., Breuer, L., & Hui, D. (2015). SPOTting model parameters using a ready-made Python package. *PLoS ONE*, *10*(12), e0145180. https://doi.org/10.1371/journal.pone.0145180

ISRIC - World Soil Information (2007), Soil and Terrain Database for Kenya (KENSOTER) *version 2.0*.

Jacobs, S. R., Timbe, E., Weeser, B., Rufino, M. C., Butterbach-Bahl, K., & Breuer, L. (2018). Assessment of hydrological pathways in East African montane catchments under different land use. *Hydrology and Earth System Sciences*, *22*(9), 4981–5000. https://doi.org/10.5194/hess-22-4981-2018

Jacobs, S. R., Weeser, B., Guzha, A. C., Rufino, M. C., Butterbach-Bahl, K., Windhorst, D., & Breuer, L. (2018). Using high-resolution data to assess land use impact on nitrate dynamics in East African Tropical Montane catchments. *Water Resources Research*, *54*, 1812–1830. https://doi.org/10.1002/2017WR021592

Jehn, F. U., Breuer, L., Houska, T., Bestian, K., & Kraft, P. (2018). Incremental model breakdown to assess the multi-hypotheses problem. *Hydrology and Earth System Sciences*, *22*(8), 4565–4581. https://doi.org/10.5194/hess-22-4565-2018

Jian, J., D. Ryu, J. F. Costelloe, and C.-H. Su (Eds.) (2015), Towards reliable hydrological model calibrations with river level measurements, 21st International Congress on Modelling and Simulation, *Modelling and Simulation Society of Australia and New Zealand*.

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

Kavetski, D., & Clark, M. P. (2011). Numerical troubles in conceptual hydrology: Approximations, absurdities and impact on hypothesis testing. *Hydrological Processes*, *25*(4), 661–670. https://doi.org/10.1002/hyp.7899

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, *42*, W0s3S04. https://doi.org/10.1029/2005WR004362

Kraft, P., Vaché, K. B., Frede, H.-G., & Breuer, L. (2011). CMF: A hydrological programming language extension for integrated catchment models. *Environmental Modelling & Software*, *26*(6), 828–830. https://doi.org/10.1016/j.envsoft.2010.12.009

Kraft, P., C. Weber, K. Bestian, F. Jehn, T. Houska, A. Karlson, and D. Windhorst (2018), Cmf 1.4, doi:https://doi.org/10.5281/zenodo.1402063.

Krhoda, G. O. (1988). The impact of resource utilization on the hydrology of the Mau Hills Forest in Kenya. *Mountain Research and Development*, *8*(2/3), 193. https://doi.org/10.2307/3673447

Lowry, C. S., & Fienen, M. N. (2013). CrowdHydrology: Crowdsourcing hydrologic data and engaging citizen scientists. *Ground Water*, *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*, 36. https://doi.org/10.3389/feart.2019.00128

Lüthi, B., T. Philippe, and S. Peña-Haro (2014), Mobil device app for small open-channel flow measurement, 7th International Congress on Environmental Modelling and Software, San Diego, California.

Maier, N., Breuer, L., & Kraft, P. (2017). Prediction and uncertainty analysis of a parsimonious floodplain surface water-groundwater interaction model. *Water Resources Research*, *53*, 7678–7695. https://doi.org/10.1002/2017WR020749

Mazzoleni, M., Cortes Arevalo, V. J., Wehn, U., Alfonso, L., Norbiato, D., Monego, M., et al. (2018). Exploring the influence of citizen involvement on the assimilation of crowdsourced observations: A modelling study based on the 2013 flood event in the Bacchiglione catchment (Italy). *Hydrology and Earth System Sciences*, *22*(1), 391–416. https://doi.org/10.5194/hess-22-391-2018

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

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, *21*(2), 239–245. https://doi.org/10.1080/00401706.1979.10489755

Messer, H., Zinevich, A., & Alpert, P. (2006). Environmental monitoring by wireless communication networks. *Science (New York, N.Y.)*, *312*(5774), 713. https://doi.org/10.1126/science.1120034

Montanari, A., Young, G., Savenije, H. H. G., Hughes, D., Wagener, T., Ren, L. L., et al. (2013). "Panta Rhei—Everything Flows": Change in hydrology and society—The IAHS Scientific Decade 2013–2022. *Hydrological Sciences Journal*, *58*(6), 1256–1275. https://doi.org/10.1080/02626667.2013.809088

Mu, Q., Zhao, M., & Running, S. W. (2011). Improvements to a MODIS global terrestrial evapotranspiration algorithm. *Remote Sensing of Environment*, *115*(8), 1781–1800. https://doi.org/10.1016/j.rse.2011.02.019

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

Njue, N., Kroese, J. S., Gräf, J., Jacobs, S. R., Weeser, B., Breuer, L., & Rufino, M. C. (2019). Citizen science in hydrological monitoring and ecosystem services management: State of the art and future prospects. *Science of the Total Environment*, *693*. https://doi.org/10.1016/j.scitotenv.2019.07.337

Nyenzi, B. S., Kiangi, P. M. R., & Rao, N. N. P. (1981). Evaporation values in East Africa. *Archives for Meteorology, Geophysics, and Bioclimatology Series B*, *29*(1-2), 37–55. https://doi.org/10.1007/BF02278189

Olang, L., and P. Kundu (2011), Land degradation of the Mau Forest complex in Eastern Africa: A review for management and restoration planning, in *Environmental Monitoring*, edited by E. Ekundayo, InTech.

Overdevest, C., Orr, C., & Stepenuck, K. (2004). Volunteer stream monitoring and local participation in natural resource issues. *Research in Human Ecology*, *11*(2), 177–185.

Overeem, A., Robinson, J. C. R., Leijnse, H., Steeneveld, G. J., Horn, B. K. P., & Uijlenhoet, R. (2013). Crowdsourcing urban air temperatures from smartphone battery temperatures. *Geophysical Research Letters*, *40*, 4081–4085. https://doi.org/10.1002/grl.50786

Reges, H. W., Doesken, N., Turner, J., Newman, N., Bergantino, A., & Schwalbe, Z. (2016). CoCoRaHS: The evolution and accomplishments of a volunteer rain gauge network. *Bulletin of the American Meteorological Society*, *97*(10), 1831–1846. https://doi.org/10.1175/BAMS-D-14-00213.1

Rodda, J. C. (2001). Water under pressure. *Hydrological Sciences Journal*, *46*(6), 841–854. https://doi.org/10.1080/02626660109492880

Running, S. W., Q. Mu, M. Zhao, and A. Moreno (2017), User's guide MODIS Global Terrestrial Evapotranspiration (ET) product (NASA MOD16A2/A3) NASA Earth Observing System MODIS Land Algorithm, http://files.ntsg.umt.edu/data/NTSG_Products/MOD16/MOD16UsersGuide2016_V1.5_2017Jul24.pdf

Seibert, J., Strobl, B., Etter, S., Hummer, P., & van Meerveld, H. J. (2019). Virtual staff gauges for crowd-based stream level observations. *Frontiers in Earth Science*, *7*, 18. https://doi.org/10.3389/feart.2019.00070

Seibert, J., & Vis, M. J. P. (2016). How informative are stream level observations in different geographic regions?: Value of stream level observations. *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 modeling. *Hydrological Processes*, *32*, 1120–1125. https://doi.org/10.1002/hyp.11476

Senay, G. B., Leake, S., Nagler, P. L., Artan, G., Dickinson, J., Cordova, J. T., & Glenn, E. P. (2011). Estimating basin scale evapotranspiration (ET) by water balance and remote sensing methods. *Hydrological Processes*, *25*(26), 4037–4049. https://doi.org/10.1002/hyp.8379

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

van de Giesen, N., Hut, R., & Selker, J. (2014). The Trans-African Hydro-Meteorological Observatory (TAHMO). *WIREs Water*, *1*(4), 341–348. https://doi.org/10.1002/wat2.1034

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

Velpuri, N. M., Senay, G. B., Singh, R. K., Bohms, S., & Verdin, J. P. (2013). A comprehensive evaluation of two MODIS evapotranspiration products over the conterminous United States: Using point and gridded FLUXNET and water balance ET. *Remote Sensing of Environment*, *139*, 35–49. https://doi.org/10.1016/j.rse.2013.07.013

Vörösmarty, C., Askew, A., Grabs, W., Barry, R. G., Birkett, C., Doll, P., et al. (2001). Global water data: A newly endangered species, *Eos. Transactions of the American Geophysical Union*, *82*(5), 54. https://doi.org/10.1029/01EO00031

Wagner, S., Kunstmann, H., Bárdossy, A., Conrad, C., & Colditz, R. R. (2009). Water balance estimation of a poorly gauged catchment in West Africa using dynamically downscaled meteorological fields and remote sensing information. *Physics and Chemistry of the Earth, Parts A/B/C*, *34*(4-5), 225–235. https://doi.org/10.1016/j.pce.2008.04.002

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

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

Windhorst, D., Kraft, P., Timbe, E., Frede, H.-G., & Breuer, L. (2014). Stable water isotope tracing through hydrological models for disentangling runoff generation processes at the hillslope scale. *Hydrology and Earth System Sciences*, *18*(10), 4113–4127. https://doi.org/10.5194/hess-18-4113-2014

Yan, K., Di Baldassarre, G., Solomatine, D. P., & Schumann, G. J.-P. (2015). A review of low-cost space-borne data for flood modelling: Topography, flood extent and water level. *Hydrological Processes*, *29*(15), 3368–3387. https://doi.org/10.1002/hyp.10449