Water Resources Research

RESEARCH ARTICLE

10.1029/2019WR026325

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

Key Points:

- Observations collected by volunteers can improve the performance of a semidistributed hydrological model
- The ensemble Kalman filter can integrate observations provided by volunteers into a semidistributed hydrological model
- For certain interarrival times of observations, the hydrological model reproduced the central tendency of streamflow and stream temperature

Supporting Information:

- Supporting Information S1

Correspondence to:
P. M. Avellaneda,
pmavellaneda@gmail.com

Citation:
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*, e2019WR026325. https://doi.org/10.1029/2019WR026325

Received 11 SEP 2019
Accepted 17 APR 2020
Accepted article online 22 APR 2020

Improving Hydrological Models With the Assimilation of Crowdsourced Data

P. M. Avellaneda[1], D. L. Ficklin[1], C. S. Lowry[2], J. H. Knouft[3], and D. M. Hall[4]

[1]Department of Geography, Indiana University, Bloomington, IN, USA, [2]Department of Geology, University at Buffalo, Buffalo, NY, USA, [3]Department of Biology, Saint Louis University, Saint Louis, MO, USA, [4]Department of Biomedical, Biological and Chemical Engineering; School of Natural Resources, University of Missouri, Columbia, MO, USA

Abstract Small streams often lack reliable hydrological data. Environmental agencies play a key role in providing such data; however, these agencies are often challenged by the growing monitoring needs and lack of funding. Given the spatial mismatch between observed data and small watersheds/headwaters, local volunteers can act as potentially valuable research partners. We examine how CrowdHydrology, a citizen science program that collects stream stage and stream temperature observations, improves a hydrologic model of the Boyne River, Michigan, USA. Volunteers provided observations at four calibration sites with different interarrival times of the observations. We tested whether stream stage and stream temperature observations (measured by volunteers) improved the performance of a Soil and Water Assessment Tool (SWAT) model of the Boyne River. Observations were integrated into the model using the ensemble Kalman filter. This framework allowed us to integrate observation error, track the variability of model parameters, and simulate daily streamflow and stream temperature across the watershed. Measures of daily model performance included the Nash-Sutcliffe efficiency, modified Nash-Sutcliffe efficiency ($E_{f\text{-}mod}$), refined index of agreement (d_r), and relative bias (*Bias*). For all calibration sites, estimates of streamflow improved after data assimilation compared to simulations based on initial/default SWAT parameters. Different measures of model performance emerged based on the interarrival times of the observations. Results demonstrate that observations collected by local volunteers, with a certain temporal resolution, can improve SWAT hydrological models and capture central tendency.

1. Introduction

Hydrologic and ecosystem models rely on observed data for both calibration and model validation. Collection of these observed data is largely focused on locations with urban/municipal needs (e.g., flooding, water supply), with smaller headwater streams often lacking information for the development of hydrological models. Government environmental agencies play a key role in providing reliable hydrologic data; however, these agencies are often challenged by the growing number of monitoring needs (Cosgrove & Loucks, 2015; Hannah et al., 2011). Moreover, evidence suggests a decline in available water monitoring information due to limited funding, weakening of infrastructure, and shifting government priorities (Hannah et al., 2011; Ruhi et al., 2018; Vorosmarty et al., 2001). Considering the need for continued water monitoring, citizen science applications represent a potentially promising resource for the collection of hydrological data.

Citizen science refers to the active participation of the general public in the generation of new scientific knowledge (Buytaert et al., 2014). In water resources, citizen science contributes to the public engagement in a scientific project via providing or analyzing data (Le Coz et al., 2016; McKinley et al., 2017; Yang & Ng, 2017). Local volunteers can provide a wide range of information for hydrologic monitoring purposes (Stepenuck & Genskow, 2018). For example, pictures and videos can help determine the extent of a flood event (Le Coz et al., 2016), volunteers can use simple methods to measure streamflow (Davids et al., 2019), and mobile phone text messages can be used to submit stream stage observations (Lowry et al., 2019; Lowry & Fienen, 2013; Weeser et al., 2018). Water level measurements can be derived from pictures as in CrowdWater (Seibert et al., 2019), stream level classes that represent a range of streamflow data (Etter et al., 2020; Strobl et al., 2019; van Meerveld et al., 2017), and community-based monitoring programs can gather hydrometeorological or water quality data to improve the knowledge of local water resources (Jollymore et al., 2017; Walker et al., 2016). Throughout this paper, the term “citizen” refers to a volunteer (community member) that provided an observation and does not refer to a citizenship status.

When volunteers provide stream stage observations, four characteristics are relevant when integrating these types of data into a hydrological model: temporal coverage, spatial coverage, quantity, and accuracy (Assumpção et al., 2018). These characteristics are relevant for model calibration, a phase in which these observations are assumed to have random accuracy and larger errors than professional observations (Aceves-Bueno et al., 2017; Cortes Arevalo et al., 2014; Etter et al., 2018). Methods to calibrate hydrological models include manual calibration, optimization algorithms, Bayesian inference, and data assimilation. Due to the characteristics of citizen science observations, data assimilation methods are suitable for integrating this type of data into hydrological models. For instance, data assimilation methods account for measurement errors and the temporal evolution of model parameters while allowing for model updates as new information becomes available (Mazzoleni et al., 2018; Moradkhani et al., 2005; Xie & Zhang, 2013). To explore how the accuracy of these observations influences model performance, streamflow measurement errors (observational error) can be generated from a probability distribution or a stochastic process (Etter et al., 2018; Mazzoleni et al., 2015, 2017, 2018). For example, Mazzoleni et al. (2017) improved flood prediction by considering observational errors to be normally distributed with zero mean and given standard deviation. Etter et al. (2018) reported success in model calibration only when the error standard deviation between observations (streamflow) and professional measurements was reduced by half. In their study, field surveys were conducted to determine the typical errors between observations and professional measurements.

We hypothesize that sparse, discontinuous, spatially distributed volunteer-provided observations (stream stage and stream temperature) can be used to improve a semidistributed hydrological model. The CrowdHydrology network (Lowry et al., 2019; Lowry & Fienen, 2013) was used to gather stream stage and stream temperature observations in the Boyne River, located in Northern Michigan, United States. Volunteers provided observations at four CrowdHydrology sites across the watershed and submitted the data in the form of text messages. The Soil and Water Assessment Tool (SWAT), a semidistributed watershed model, was chosen because of its ability to simulate daily streamflow and stream temperature at various locations within a watershed (Arnold et al., 2012; Ficklin et al., 2012). The SWAT model was calibrated and validated within a data assimilation approach that considered observational error, uncertainty of the atmospheric forcing (rainfall and air temperature), and the temporal variability of model parameters. In the following sections, an observation refers to a stream stage or stream temperature observation provided by volunteers and reported via a text message system (crowdsourced data).

2. Boyne River Watershed

The Boyne River (Figure 1) is the second largest tributary flowing into Lake Charlevoix, Michigan, an inland lake that drains into Lake Michigan. There are two branches of the Boyne River that split above the Boyne River Hydroelectric Dam, which supplies energy for the operation of a recreational area. The hydroelectric project consists of a reservoir with a storage capacity of 1.67×10^6 m^3 and an area of 3.56×10^5 m^2 at the average pool height of 4.70 m. The existing dam embankment is 290-m long and 7.6-m high (Boyne, 2017). The Boyne River is an excellent salmon and trout fishery: steelhead trout (*Oncorhynchus mykiss*) and salmon (*Oncorhynchus spp.*) are located downstream of the dam, while brook trout (*Salvelinus fontinalis*) and brown trout (*Salmo trutta*) can be found upstream of the dam (The Tip of the Mitt Watershed Council, 2012). The Boyne River watershed has an area of 184 km^2 and is dominated by forests (55%), agriculture (13%), and wetlands (10%). Glacial tills and outwash overlying Devonian age bedrock dominate the watershed. The mean annual rainfall was 1,015 mm, the mean annual snowfall was 2,737 mm, and the mean air temperature was 8 °C between June of 2014 and May of 2019 (period of analysis defined in section 5.1).

We selected the Boyne River watershed because of the active participation of two community groups, Friends of The Boyne River and Michigan Trout Unlimited. These two groups were essential for community engagement, participation in the installation of stream gauges, and solving maintenance issues with the CrowdHydrology infrastructure. The methods explained in a later section (section 5) could be applied to a different watershed where CrowdHydrology instrumentation is in operation or could be deployed.

3. Data Collected by Volunteers

We utilized CrowdHydrology (www.crowdhydrology.com), a citizen science network that collects hydrologic data throughout the United States (Lowry & Fienen, 2013), to obtain local stream stage and stream

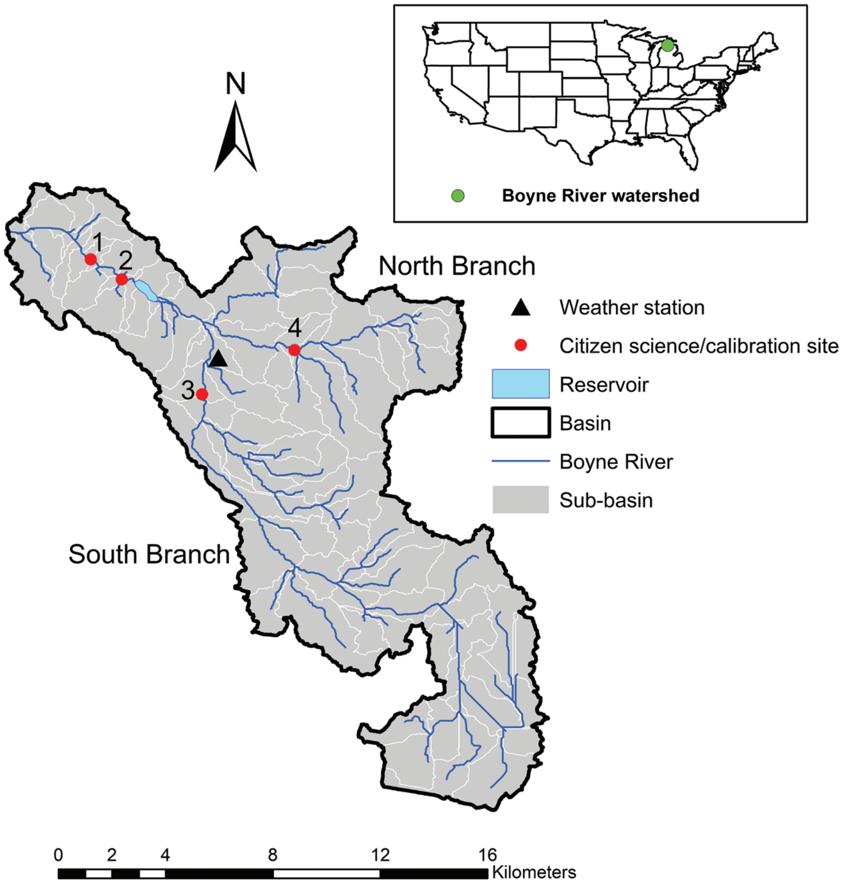

Figure 1. Map of the Boyne River watershed with the location of CrowdHydrology stations used for calibration and validation.

temperature observations. CrowdHydrology provides an infrastructure for volunteers to send a text message with the current stream stage and stream temperature to a server located at the University at Buffalo (Lowry et al., 2019; Lowry & Fienen, 2013). A CrowdHydrology gauge station consists of a gauge plate mounted in the stream with a simple sign asking that observations be sent via a text message (Figure 2). Using Social.Water, the server transforms the format-varying text messages into a unified format and then inserts the resulting data points into a publicly available database (Fienen & Lowry, 2012). All sites (Figure 1) are also equipped with a digital stream temperature sensor that reports to a screen from which volunteers can read and submit an observation (Figure 2). We worked closely with the Friends of the Boyne River and Michigan Trout Unlimited, as well as other active stakeholders in Boyne City, Michigan, to install gauges and address maintenance issues. The two community groups led efforts to distribute information to the general public by circulating flyers by mail and in social media, posting project material to their websites, hosting public presentations to the community, and supplying information for newspaper articles/interviews. Volunteers started sending text messages in the summer of 2014 by providing stream stage observations. Stream temperature sensors were installed in the summer of 2017 near the stream gauge sites.

Observations provided by volunteers and stage-discharge relationships are uncertain. Observations may arrive with irregular frequency, random levels of precision, and be scarce across periods of analysis. In a previous study, Lowry and Fienen (2013) validated CrowdHydrology observations with a pressure transducer, which revealed root-mean-square error of participant versus research data of about 6 mm (0.02 feet) for stream stage, roughly the resolution of the installed Class A gauge. In-stream temperature sensors reported observations with a resolution of 0.06 °C (0.1 °F).

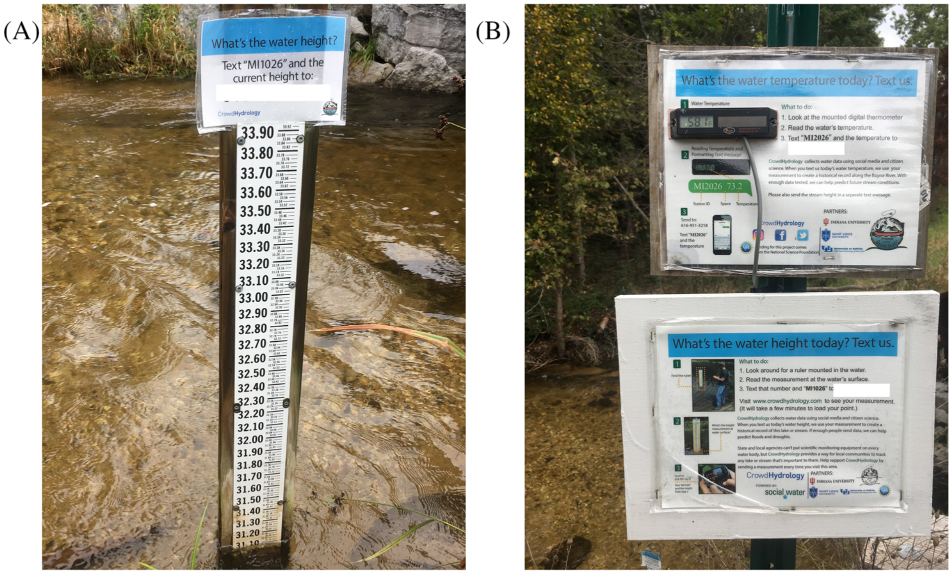
(A)
What's the water height?
Text "MI1026" and the current height to:
CrowdHydrology
33.90
33.80
33.70
33.60
33.50
33.40
33.30
33.20
33.10
33.00
32.90
32.80
32.70
32.60
32.50
32.40
32.30
32.20
32.10
32.00
31.90
31.80
31.70
31.60
31.50
31.40
31.30
31.20
(B)
What's the water temperature today? Text us.
Water Temperature
What to do:
1. Look at the mounted digital thermometer
2. Read the water's temperature.
3. Text "MI2026" and the temperature to
MI2026 73.2
Station ID
Space
Temperature
Send to: 616-951-3218
Text "MI2026" and the temperature
CrowdHydrology
PARTNERS:
What's the water height today? Text us.
What to do:
1. Look around for a ruler mounted in the water.
2. Read the measurement at the water's surface.
3. Text that number and "MI1026" to
Visit www.crowdhydrology.com to see your measurement. (It will take a few minutes to load your point.)
CrowdHydrology
social water

Figure 2. CrowdHydrology station for Calibration Site 4: (a) stream gauge and (b) digital temperature sensor screen and instructions to send an observation via a cell phone text message.

4. Hydrological Model

4.1. Model Description

We used the SWAT (version 2016/rev. 664) (Arnold et al., 1998, 2012) because it is relatively easy to setup, can adequately simulate hydrology, and is coupled with our previously developed stream temperature model. In the SWAT model, the water cycle is simulated based on a water balance equation where a change in soil water content is a function of precipitation, evapotranspiration, groundwater flow, infiltration, and surface runoff (Neitsch et al., 2011). Infiltrated water reaches the unsaturated zone and may move past the lowest depth of the soil profile to recharge a shallow and a deep aquifer. Baseflow from the two aquifers, lateral flow from the soil profile, and surface runoff contribute to streamflow. The described water balance is evaluated for each hydrologic response unit (HRU)—a spatial unit with specific land use, soil type, and surface slope—defined within a subbasin and then summed over all HRUs. Potential evapotranspiration was calculated from the Penman-Monteith equation (Monteith, 1965), and infiltration was estimated using the curve number method (USDA, 1986). Finally, streamflow estimated at a subbasin level is routed through the river system using a kinematic wave model (Neitsch et al., 2011). SWAT distinguishes solid and liquid precipitation based on near-surface temperature. When the mean daily air temperature of a subbasin is lower than a snowfall temperature threshold, precipitation is considered solid, and it is accumulated until snowmelt (Grusson et al., 2015; Neitsch et al., 2011). Snowmelt is controlled by air and snowpack temperatures. For stream temperature, we adopted the stream temperature model for SWAT developed by Ficklin et al. (2012). This model estimates stream temperature by considering temperature and amount of local water contribution within a subbasin (e.g., snowmelt, groundwater, lateral flow, and surface runoff), temperature and inflow volume from upstream subbasins, and heat transfer at the air-water interface during the streamflow travel time in a subbasin (Barnhart et al., 2014; Ficklin et al., 2012, 2013). Simulations were run at a daily time step.

4.2. Model Input Data

The Boyne River hydrological model was implemented using meteorological data, digital elevation models, and soil and land use maps. Precipitation and air temperature data consisted of daily records collected at a National Oceanic and Atmospheric Administration weather station located in Boyne Falls, Michigan (USC00200925) (Figure 1). The following spatial information was used to develop the hydrological model: a 1/3 arc-second USGS digital elevation model, a 30 × 30-m land use map from the National Land Cover Database 2011 (obtained from the MRLC Web site https://www.mrlc.gov/nlcd2011.php), and a 30 × 30-m soil map from the US Department of Agriculture (SSURGO database for Charlevoix County, Michigan, obtained from the USDA Web site https://sdmdataaccess.sc.egov.usda.gov).

4.3. Model Setup

Based on topography and the natural stream network, the Boyne River watershed was divided into 104 subbasins. To balance a reasonable resolution of soil properties, land use distribution, and computational time for calibration, these subbasins were divided into a total of 626 HRUs. HRUs were created based on land uses, soils, and slope areas with coverage greater than 15% within a subbasin. Preliminary SWAT model parameters (supporting information, Table S1) were generated using ArcSWAT (Winchell et al., 2007), which allowed for the creation of input files from the digital maps (e.g., soil and land use maps) and internal databases (e.g., Manning roughness coefficients, evapotranspiration coefficients). Meteorological inputs including solar radiation, wind speed, and relative humidity were automatically generated by the weather generator model within SWAT (Arnold et al., 2012). We adopted the simulated controlled outflow-target release scheme of SWAT to represent the reservoir of the Boyne River hydropower project (Jalowska & Yuan, 2019; Neitsch et al., 2011). The reservoir is manually operated to maintain a reservoir level of 4.7 m. This operation is performed by adjusting flow through a turbine and spillways as a means of maintaining the target reservoir volume (167.7×10^4 m^3). For a given day, the volume of water flowing out of the reservoir is a function of the difference between the current volume of water in the reservoir (estimated from the influent flow from upstream subbasins) and the target reservoir volume for a given day. Based on the operation of the reservoir, we assumed 2 days as the number of days required to reach the target storage. SWAT parameters for the reservoir configuration are displayed in Table S3. We assumed a warm-up period of 2 years to minimize the effect of the initial soil water content of the watershed.

5. Hydrological Model Calibration

5.1. Calibration and Validation Periods

The available observations were divided into periods for model calibration (data assimilation) and validation. The calibration phase for flow considered 4 years of data assimilation (discussed in the next section) and 1 year for validation (Table 1). For stream temperature, the dataset was split into one period for calibration and one period for validation. We improved stage-discharge relationships as more field data were gathered over time and performed maintenance of field equipment as soon as our research partners detected an issue.

5.2. Stage-Discharge Relationships

Streamflow was derived from stream stage observations using stage-discharge relationships. For each calibration site, field observations of discharge (Q) and stream stage (H) were fitted to the following power function (World Meteorological Organization, 2010):

$$Q = a\ (H - b)^c, \tag{1}$$

where a is a coefficient related to the channel conditions (e.g., flow resistance, flow energy slope, and wetted area), b is a reference level ($H \geq b$), and c is an exponent related to the type of hydraulic control. Before generating streamflow from stream stage observations and the fitted stage-discharge relationships, we removed unrealistic stream stage observations that fell outside a predefined range of values. For example, the range of expected values were defined between 0 and 1 m for Calibration Site 4 (Figure 2a).

We inferred the parameters of the power function using the BaRatin method (Le Coz et al., 2014). This method applies a Markov chain Monte Carlo sampler to capture the probability distribution of the power function coefficients. These distributions can then be used to derive the stage-discharge relationship and 95% uncertainty ranges. Field observations and stage-discharge relationships for all sites are displayed in Figure 3. The uncertainty ranges consider errors made by an operator when measuring stream stage and velocity with an acoustic Doppler velocimeter, used to measure streamflow. The acoustic Doppler velocimeter (Flowtracker 2) was deployed 14 times at each calibration site during the period of analysis. At each calibration site, the cross section of the stream was divided into 25–35 stations to measure discharge. Two vertical discharge measurements (at 0.2 and 0.8 D, with D = water depth) were used to estimate discharge if the water level was deep enough ($D > 0.75$ m); otherwise, only one vertical measurement (at 0.6 D) was used.

Table 1
Time Periods Used for Calibration (Data Assimilation) and Validation of Stream Flow and Stream Temperature

		Number of observations				
Period	Phase	Site 1	Site 2	Site 3	Site 4	Total
6/1/2014–5/31/2015	Flow data assimilation (DA1)	21	41	19	8	89
6/1/2015–5/31/2016	Flow data assimilation (DA2)	36	24	18	4	82
6/1/2016–5/31/2017	Flow data assimilation (DA3)	63	19	8	7	97
6/1/2017–5/31/2018	Flow data assimilation (DA4)	58	38	28	15	139
6/1/2018–5/31/2019	Flow validation (VAL)	62	34	24	18	138
9/1/2017–8/31/2018	Temperature data assimilation (DA1)	37	31	31	13	112
9/1/2018–5/31/2019	Temperature validation (VAL)	30	16	26	14	86

Note. The number of observations is reported for each calibration site (Figure 1).

We assumed an error of ±6 mm for water level (the resolution of the stream gauge) and integrated discharge uncertainties reported by the instrument (vertical red error bars in Figure 3) (Cohn et al., 2013). In the field, discharge uncertainty varied between 2% and 9% for Sites 1 and 2 and 3% and 11% for Sites 3 and 4. Discharge uncertainty was reported by the acoustic Doppler velocimeter and was based on channel geometry and flow conditions experienced during the discharge measurement (Kiang et al., 2009). Larger errors were expected upstream (Sites 3 and 4) due to the shallow water conditions and less uniform streamflow beds, which prevented collection of two velocity observations in the vertical direction. Based on the described approach, discharges could have the following average variations from the estimated stage-discharge relationships (solid black lines in Figure 3): 6% for Site 1, 9% for Site 2, and 22% for the upstream sites. For example, for Site 1, $Q = 3.8 \pm 0.23$ m^3/s when $H = 8$ m (3.8 m^3/s × 0.06 = 0.23 m^3/s). Note that the uncertainty grows with discharge, but here we consider only average variation across the range of discharge values.

The accuracy of the observations (stream stage and stream temperature) contributes to the uncertainty of streamflow and stream temperature estimates. If a participant reported a stream stage with an error that

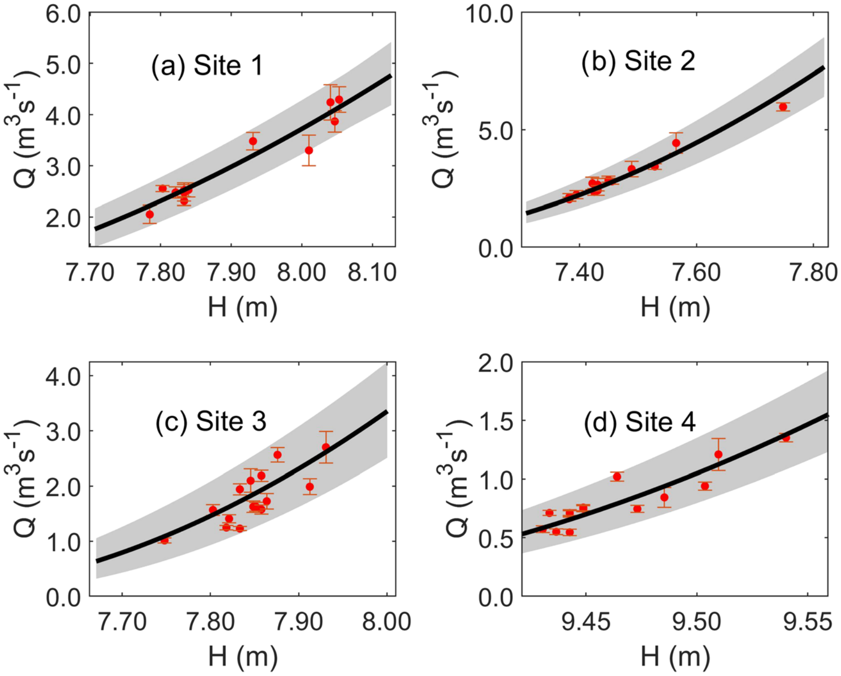

Figure 3. Estimated stage-discharge relationships (solid black line) at four calibration sites of the Boyne River. Field observations (solid red dots) of stream stage (*H*), discharge (*Q*), and measurement error (vertical red error bars) cover a wide range of flow conditions. The light gray area represents the 95% uncertainty ranges. The fitted coefficients of the estimated stage-discharge relationships are listed in Table S2.

is three orders of magnitude (±18 mm) higher than the stream gauge resolution, discharges would deviate up to 6% across sites. To consider this uncertainty, we assumed the observation errors to have a fixed and common standard deviation of 25% of an observation. This deviation includes both stage-discharge relationship uncertainty and the unpredictable accuracy of an observation. We adopted the same deviation for stream temperature observations from volunteers. A standard deviation of 10% is commonly used when using professional observations (Vrugt et al., 2005).

5.3. Data Assimilation of Observations

Data assimilation methods improve hydrologic models by combining predictions of an imperfect model with observations that arrive with irregular frequency, random levels of precision, and different levels of data collection across periods of analysis. We selected the ensemble Kalman filter (EnKF) because there is evidence of its potential as a data assimilation technique for uncertain crowdsourced data (Mazzoleni et al., 2017, 2018), ability to capture possible seasonal variations in parameters (Pathiraja et al., 2016; Wu & Johnston, 2007), and capacity for the model to update parameters as soon as a cluster of observations is received. EnKF is a well-tested data assimilation technique in the case of nonlinear systems and has been applied in the context of soil moisture modeling (Han et al., 2012; Patil & Ramsankaran, 2017; H. Zhang et al., 2016; Y. Zhang et al., 2017), streamflow routing and forecasting (Mazzoleni et al., 2018; Moradkhani et al., 2005; Vrugt et al., 2006; Xie & Zhang, 2013), and data collected by volunteers (Mazzoleni et al., 2015, 2017, 2018). Typically, the EnKF is based upon Monte Carlo or ensemble generations where a forecast of state variables (e.g., storage of water or energy) and model parameters are made by propagating an ensemble of n model states using the updated states and parameters from a previous time step (Moradkhani et al., 2005). In our approach, we observed streamflow and stream temperature (model output) and thus updated only model parameters (θ) according to

$$\theta_{t+1}^{i} = \theta_{t+1}^{i-} + \mathrm{K}_{\theta}\left(\widehat{\mathrm{y}}_{t+1}^{i} - \mathrm{y}_{t+1}^{i}\right), \tag{2}$$

where θ_{t+1}^{i} represents a $p \times 1$ array of updated parameters at time interval $t = t + 1$ and ith ensemble member, with $i = \{1, ...,n\}$, θ_{t+1}^{i-} denotes a $p \times 1$ array of forecast model parameters (approximations from a previous time step), $\widehat{\mathrm{y}}_{t+1}^{i}$ is a 4×1 array of observations (i.e., streamflow or stream temperature from the four calibration sites), y_{t+1}^{i} represents a 4×1 array of simulated streamflow or stream temperature values (streamflow is calculated during a first model run and then used to compute stream temperature on a second model run), and K_{θ} is the $p \times 4$ Kalman gain matrix (see below). p signifies the number of parameters for the streamflow model ($p = 27$) and the temperature model ($p = 6$). Two types of changes were applied to the SWAT parameters: a type 1 change means a replacement of an existing parameter value by a given value and a type 2 change means a multiplication of an existing parameter value by (1 + a given value). For example, a type 2 calibration applied to the soil hydraulic conductivity SOL_K = −0.25 would cause a global relative change of preliminary SOL_K values by multiplying each existing parameter value by a factor of (1–0.25) = 0.75. Type 1 changes are the same throughout the watershed, while type 2 changes result in spatially different parameter values throughout the watershed.

The Kalman gain matrix measures how much state variables or model parameters should change based on a given observation. The Kalman gain is computed as follows:

$$\mathrm{K}_{\theta} = \mathrm{C}_{\theta y}\left[\mathrm{C}_{yy} + \mathrm{R}\right]^{-1}, \tag{3}$$

where $\mathrm{C}_{\theta y}$ is a $p \times 4$ error covariance matrix of the forecast parameters θ_{t+1}^{i-} and simulated values y_{t+1}^{i}, C_{yy} represents a 4×4 error covariance matrix of the simulated values y_{t+1}^{i}, and R denotes a 4×4 observation error covariance matrix. A large value in the covariance matrix encourages an abrupt change of a parameter value as the combined uncertainty in parameter and simulated streamflow or stream temperature (represented by $\mathrm{C}_{\theta y}$) is larger than the uncertainty in simulated streamflow or temperature plus the uncertainty of an observation (represented by C_{yy}+R). A negligible change in parameter trajectory will occur when the uncertainty in the observation and simulated value surpasses that of a parameter and simulated value —the Kalman gain takes a value close to zero. In the assimilation process, $R = (\rho\ \mathrm{CS}_{\mathrm{obs}})^2$ where $\rho = 0.25$

and CS_{obs} denotes a 4 × 4 matrix whose diagonal elements are populated with observed discharges or stream temperatures and zero elsewhere.

One key step of the EnKF algorithm is the estimation of forecast parameters θ_{t+1}^{i-}. New forecast parameters may suffer from overdispersion and loss of information between filter iterations. To prevent these issues, we generated a new forecast parameter vector by perturbing parameters from a previous filter iteration and adding random noise as follows (Liu, 2000; Moradkhani et al., 2005; Xie & Zhang, 2013):

$$\theta_{t+1}^{i-} = \alpha\theta_t^i + (1-\alpha)\overline{\theta_t} + \tau_t^i, \tag{4}$$

$$\tau_t^i \sim \mathcal{N}\left(0,\ (1-\alpha)^2 var\left(\theta_t^i\right)\right), \tag{5}$$

where θ_t^i represents a $p \times 1$ array of updated parameters at time interval t and ith ensemble member, α is a shrinkage factor, $\overline{\theta_t}$ is a $p \times 1$ array of parameter means across ensemble members and time interval t, and τ_t^i denotes a $p \times 1$ noise array. In our approach, we adopted $\alpha = 0.95$ (which means a higher weight to the value of the ith ensemble member in equation (4)) and noise values drawn from a p variate normal distribution with zero mean and variance proportional to the spread of the parameters (Liu, 2000; Moradkhani et al., 2005).

A graphical representation of the methodology is presented in Figure 4. Stream stage observations are used to estimate streamflow via a stage-discharge relationship (Figures 4a and 4b). We defined sequential periods of data assimilation during the calibration phase (Figure 4c). Each period had a different number of observations (Table 1) and allowed for model assessment. Stage-discharge relationships were updated over time as different field observations targeted lower and higher flows across seasons. Model assessment at the end of an assimilation period revealed sensitivity of model parameters, allowing for model improvement when necessary (Figure 4d). Within each assimilation period, the algorithm set a new filter iteration ($t + 1$) with the arrival of a new observation (Figures 4c and 4d). Note that an observation may occur at any time and at any of the four calibration sites. In our approach, $n = 90$ ensemble members defined error covariance matrices and parameter arrays. Parameters achieved on the most recent data assimilation period defined the most up-to-date calibrated model.

5.4. Measures of Model Performance

The discrepancy between simulated and observed values was measured using the Nash-Sutcliffe efficiency (E_f) (5), modified Nash-Sutcliffe efficiency ($E_{f\text{-}mod}$) ((6)), refined index of agreement (d_r) ((7)), and relative bias ($Bias$) ((8)):

$$E_f = 1 - \frac{\sum_{i=1}^{M} |y_i - \widehat{y}_i|^2}{\sum\limits_{i=1}^{M} |\widehat{y}_i - \overline{y}|^2}, \tag{6}$$

$$E_{f-mod} = 1 - \frac{\sum_{i=1}^{M} |y_i - \widehat{y}_i|^j}{\sum_{i=1}^{M} |\widehat{y}_i - \overline{y}|^j}, \tag{7}$$

$$d_r = \begin{cases} 1 - \dfrac{\sum_{i=1}^{M}(y_i - \widehat{y}_i)}{c\sum_{i=1}^{M}(\widehat{y}_i - \overline{y})}, & when\ \sum_{i=1}^{M}(y_i - \widehat{y}_i) \le c\sum_{i=1}^{M}(\widehat{y}_i - \overline{y}) \\ \dfrac{c\sum_{i=1}^{M}(\widehat{y}_i - \overline{y})}{\sum_{i=1}^{M}(y_i - \widehat{y}_i)} - 1, & when\ \sum_{i=1}^{M}(y_i - \widehat{y}_i) > c\sum_{i=1}^{M}(\widehat{y}_i - \overline{y}) \end{cases}, \tag{8}$$

$$Bias = \frac{\sum_{i=1}^{M}(y_i - \widehat{y}_i)}{\sum_{i=1}^{M}(\widehat{y}_i)}, \tag{9}$$

where y_i represents the ith simulated value, adopted as the mean across ensemble members, with $i = \{1, ..., M\}$, M is the total number of observations, $\widehat{y}_i$ is the ith observed value, $\overline{y}$ is the mean of observed values, $\widehat{y}_{max}$ is the maximum observed value, and $\widehat{y}_{min}$ is the minimum observed value. We adopted $c = 2$ for the refined index of agreement (Willmott et al., 2012).

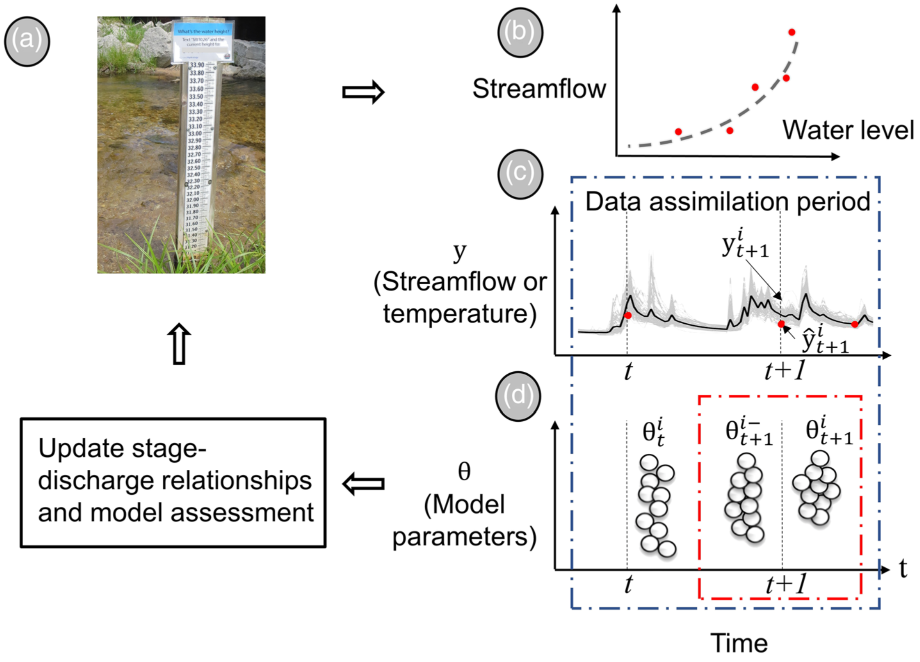

Figure 4. Schematic of the data assimilation process of observations. (a) Stream gauge at a calibration site. (b) A stage-discharge relationship (gray dashed line) is derived from uncertain field observations (red dots on gray area). (c) A new observation (red dot) at $t = t + 1$ triggers a filter iteration (box depicted with a red dash-dotted line). (d) Evolution of the distribution of model parameters. Simulated values generated for each ensemble member (gray solid lines) are derived from updated model parameters. The mean of the simulated values across ensemble members (black solid line) is estimated for model assessment.

The above measures of performance are dimensionless and target different features of the error term. The Nash-Sutcliffe efficiency (E_f) varies from $-\infty$ to 1, with higher values indicating better agreement between the model simulations and observations. One disadvantage of E_f is that large differences between simulated and observed values dominate the magnitude of the error metric (due to the summation of squared terms), thus diminishing the effect of lower observed values (Krause et al., 2005; Legates & McCabe, 1999). In contrast, $E_{f\text{-}mod}$ increases the sensitivity to lower observed values when $j = 1$, thus overcoming the higher sensitivity to large differences between simulated and observed values (Krause et al., 2005). $E_{f\text{-}mod}$ ranges from $-\infty$ *to* 1, with a value of one representing a perfect match between simulated and observed values. Values of d_r are bounded between −1 and 1, with poorly performing models identified when $d_r < 0$ and a perfect model when $d_r = 1$ (Willmott et al., 2015). Positive *Bias* values indicate model overestimation, whereas negative values indicate model underestimation. These measures were computed using HydroErr, an open source library that collects a wide range of error metric functions (Jackson et al., 2019).

6. Results

6.1. Streamflow Simulation Derived From Initial SWAT Parameters

We conducted streamflow simulations using initial/default SWAT parameters derived from available spatial information (e.g., soils, land use, and topography). These simulations, also known as open-loop simulations, served as a baseline to compare results after data assimilation. The open-loop time series simulations of streamflow exhibited sharp vertical jumps and steep declines that are typical of urban streams (Figure 5). This streamflow dynamic, however, does not represent the observed nature of the Boyne River, which is further illustrated by poor model performance (−40.9 [Site 1], −18.4 [Site 2], −21.3 [Site 3], and −34.5 [Site 4] for the Nash-Sutcliffe efficiency coefficient). This highlights the need for model calibration.

6.2. Assimilation of Streamflow Data

Trajectories for selected SWAT parameters show the ensemble spread from June 2014 to May 2018, using $n = 90$ ensemble members (Figure 6). Trajectories (gray solid lines) are presented for the available water capacity of the soil (SOL-AWC), saturated hydraulic conductivity (SOL-K), SCS runoff curve number (CN-F), and threshold depth of water in the shallow aquifer for return flow to occur (GWQMN), and the ensemble mean is

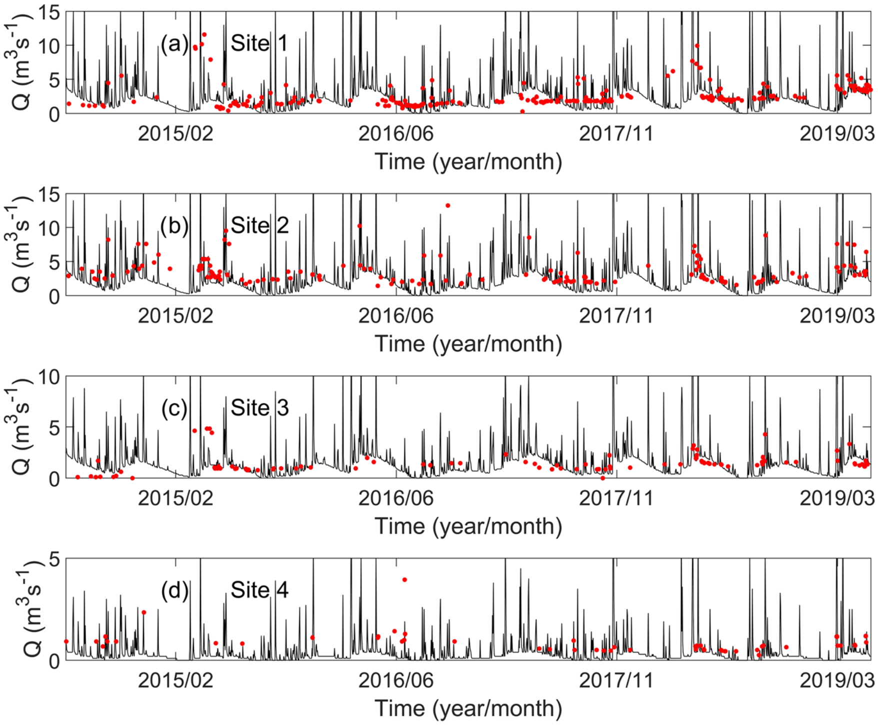

Figure 5. Simulated streamflow (solid black line) using initial SWAT parameters derived from available spatial information (e.g., soils, land use, and topography) and observations provided by volunteers (solid red dots).

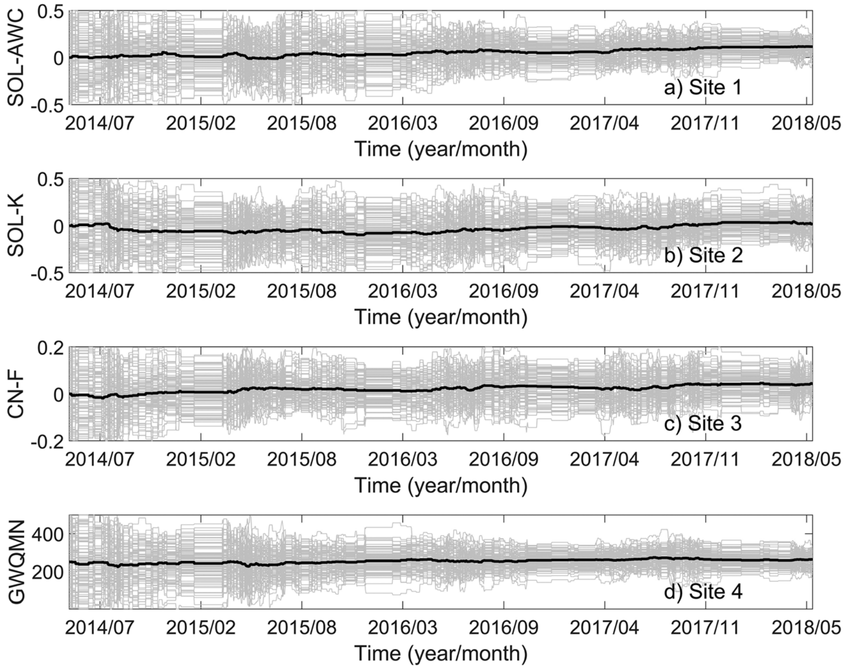

Figure 6. Trajectories (gray solid lines) of the ensemble members ($n = 90$) for the following SWAT parameters: available water capacity of the soil (SOL_AWC), saturated hydraulic conductivity (SOL_K), SCS runoff curve number (CN-F), and threshold depth of water in the shallow aquifer (GWQMN). The evolution of the ensemble mean is indicated using a solid black line. Note that SOL_AWC, SOL_K, and CN_F represent a fraction that was used to change an initial parameter value (type 2 change in Table S1) and GWQMN represents a fixed parameter value (type 1 change in Table S1).

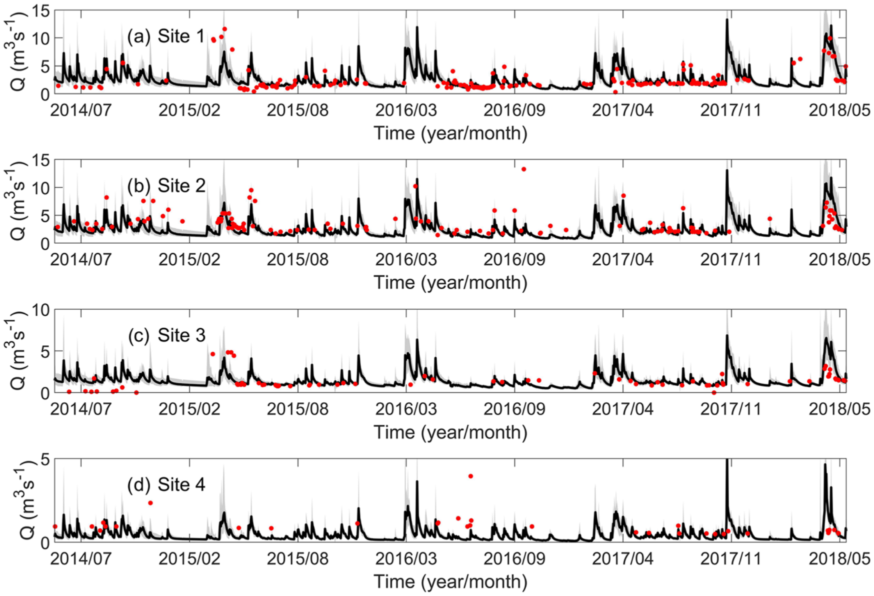

Figure 7. Simulated streamflow (solid black line), 95% uncertainty ranges (light gray areas), and observations provided by volunteers (solid red dots) at calibration sites for the data assimilation periods.

indicated with a solid black line. We tested a larger number of ensemble members, but 90 were deemed to cover the overall ensemble spread, which was computed in parallel on a high-performance computing cluster at Indiana University. The ensemble spread covered the prior distributions (uniform distributions defined by the lower and upper limits specified in Table S1) at the early stages of the data assimilation process but converged to the posterior distributions as more data were assimilated. Some parameters were more identifiable than others and reached steady posterior distributions at different times. For example, the distribution of GWQMN converged more quickly than the others, while SOL-K still displayed a larger uncertainty range at the end of the assimilation phase. This behavior was expected, as the same SOL-K change was applied across all HRUs, but soil hydraulic conductivity can vary tremendously in space. Periods of no observations resulted in undisturbed parameter trajectories during the winter season (December–March), when volunteers did not provide observations due severe weather conditions. It is important to note, though, that the Boyne River generally does not freeze during the winter due to the substantial amount of groundwater inflow.

Streamflow (*Q*) was simulated with the updated parameters during the assimilation phase. At each time interval, the SWAT model was rerun with the updated parameters using a 2-year warm-up period. Simulated stream flows (solid black lines) and observations (solid red dots) from June 2014 to May 2018 are reported in Figure 7. Because the observation error is proportional to the observed streamflow, the uncertainty ranges (light gray areas) are narrower during low flows and wider at higher flows. We selected the ensemble mean (solid black line) to represent the overall dynamics of discharge, which followed a distinctive pattern across years. In general, stream flows changed seasonally, with low flows during cold months and higher flows early in the spring and summer. Sites 1 and 2 had more observations relative to Sites 3 and 4, due to the proximity of Sites 1 and 2 to a more populated area (Boyne City) and accessibility to walking paths. For the validation period from June 2018 to May 2019, simulated flows are displayed in Figure 8, which did not consider data assimilation and used updated parameters from the last period of data assimilation (DA4, Table 1).

The uncertainty envelope covered 75% (Site 1) and 71% (Site 2) of the observations for the fourth assimilation period (DA4, from June 2017 to May 2018), while the envelope covered 46% (Site 3) and 40% (Site 4) of the observations for the same period (Figure 7). We also observed increases in coverage as more data were assimilated. For example, for Site 1, the coverage changed from 28% for the first assimilation period (DA1) to 75% for the last assimilation period (DA4). For the validation period, the coverage of the uncertainty

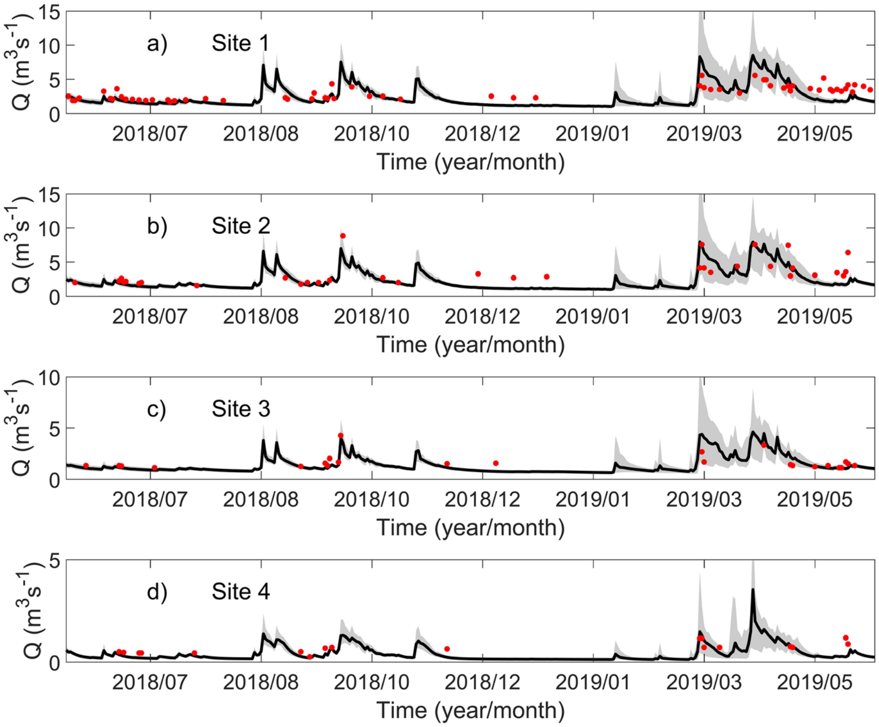

Figure 8. Simulated streamflow (solid black line), 95% uncertainty ranges (light gray areas), and observations provided by volunteers (solid red dots) for the validation period.

envelope was 46% (Site 1), 52% (Site 2), 62% (Site 3), and 44% (Site 4) (Figure 8). Accurate streamflow estimation is expected with a 95% coverage, with a lesser coverage indicating that the predictive uncertainty is either under or overestimated.

Model performance varied across sites (Figure 9), periods of data assimilation and validation (Table 1), and arrival frequency of observations. To characterize the frequency of the observations, we estimated the interarrival time of the observations (time interval between two consecutive observations) as the median of the distribution of interarrival times, which displayed a right-skewed shape. For example, the interarrival time was 2.1 days for Site 1, with a 25% percentile of 1.0 day and a 75% percentile of 6.8 days (Table 2). These arrival frequencies were expected as Sites 1 and 2 were more likely to be visited by volunteers.

For Sites 1 and 2 (Figure 9), $E_{f\text{-}mod}$ was −0.71 and −0.35 for the first period of data assimilation (DA1), then increased to 0.53 and 0.41 for the last period of data assimilation (DA4). A similar trend was observed for E_f, with a maximum value of 0.65 (Site 1) and 0.33 (Site 2) for DA4. Moreover, d_r was greater than 0 for Sites 1 and 2 across data assimilation and validation periods, which highlights the ability of the model to reproduce central tendency. For sites of lower interarrival times (Sites 3 and 4), the model performed poorly during the assimilation periods DA1, DA2, and DA3 but slightly improved for the assimilation period DA4 and validation. There was no consistent evolution of *Bias* across assimilation periods and sites. Bias was 4% (Site 1), 22% (Site 2), 51% (Site 3), and 41% (Site 4) for the last assimilation period, which shows a tendency of the model to overestimate streamflow during this period. However, the model underestimated streamflow for Sites 1, 2, and 4 during the validation period, with *Bias* ranging from −22% to −5%.

Annual rainfall varied a maximum of 10% relative to the total average rainfall of the period of analysis. However, snowfall increased by 27% during the 2018–2019 validation period relative to the average snowfall of the calibration period. Large errors (difference between simulated streamflow and observations) during the spring of 2019 (Figure 8) caused the decline in the performance metrics during the validation period (VAL in Figure 9). These errors are likely due to some parameters (SCS curve number, CN_F; surface runoff lag-time, SURLAG) not changing fast enough to accurately reproduce higher flows. However, these parameters could change faster with the assimilation of more observations. The interannual variability of

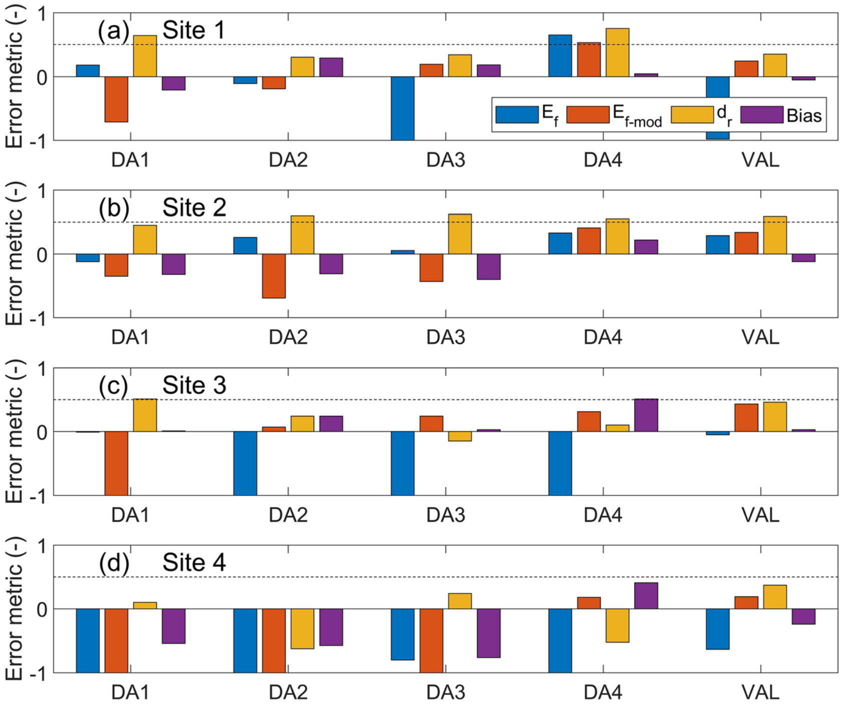

Figure 9. Measures of model performance for consecutive periods of data assimilation (DA1, DA2, DA3, and DA4) and validation period (VAL). The error metrics are the Nash-Sutcliffe efficiency (E_f), the modified Nash-Sutcliffe efficiency ($E_{f\text{-}mod}$), the refined index of agreement (d_r), and the relative bias (*Bias*). For visualization, a dashed horizontal line at 0.5 is displayed on each panel.

forcing events (e.g., rainfall and snowmelt events) also affected model performance during the data assimilation period. For example, the model did not capture heavy runoff events during April of 2015 and 2018. In particular, the largest errors occurred for a rainfall event of 93 mm (five consecutive days) in mid-April of 2018. The smallest errors mostly occurred during the summer and fall months.

6.3. Assimilation of Stream Temperature Data

Simulated stream temperatures were calculated after the model updated watershed parameters and generated stream flows, as accurate stream temperatures are dependent on accurate flows. Simulated stream temperature (solid black lines) and observations (solid red dots) are reported in Figure 10, from September 2017 to August 2018. During the winter, simulated stream temperatures were near 0.1 °C and volunteers collected only a few observations (Site 3). Most observations were collected between March and November, and stream temperatures increased in the spring and reached a maximum in the summer. Sites 1 and 2 show more observations relative to Sites 3 and 4. Overall, the uncertainty ranges and temperature ensemble mean match the observations. Simulated stream temperatures for the validation period are displayed in Figure 11 from September 2018 to May 2019.

Interarrival times in the stream temperature network were consistent with those of the streamflow network (Table 2). Despite having only one period of data assimilation, the uncertainty ranges and ensemble mean quickly captured the observations. For the assimilation period, the uncertainty envelope covered 81% (Site 1), 74% (Site 2), 64% (Site 3), and 84% (Site 4). However, the uncertainty envelopes covered a lower percentage of observations during the validation period: 33% (Site 1), 50% (Site 2), 42% (Site 3), and 57% (Site 4).

Metrics of performance show the ability of the model to estimate stream temperature across sites (Figure 12). Despite having different interarrival times across sites, results suggest that model performance

Table 2
Interarrival Time of the Observations

	Interarrival times	
Calibration site	Streamflow	Stream temperature
Site 1	2.1 (1.0, 6.8)	1.0 (0.5, 3.4)
Site 2	4.6 (1.0, 12.1)	1.9 (0.7,6.0)
Site 3	5.6 (1.0, 20.1)	2.8 (0.8, 10.8)
Site 4	7.1 (1.0, 37.7)	5.2 (1.0, 22.0)

Note. The 25% and 75% percentiles are in parentheses.

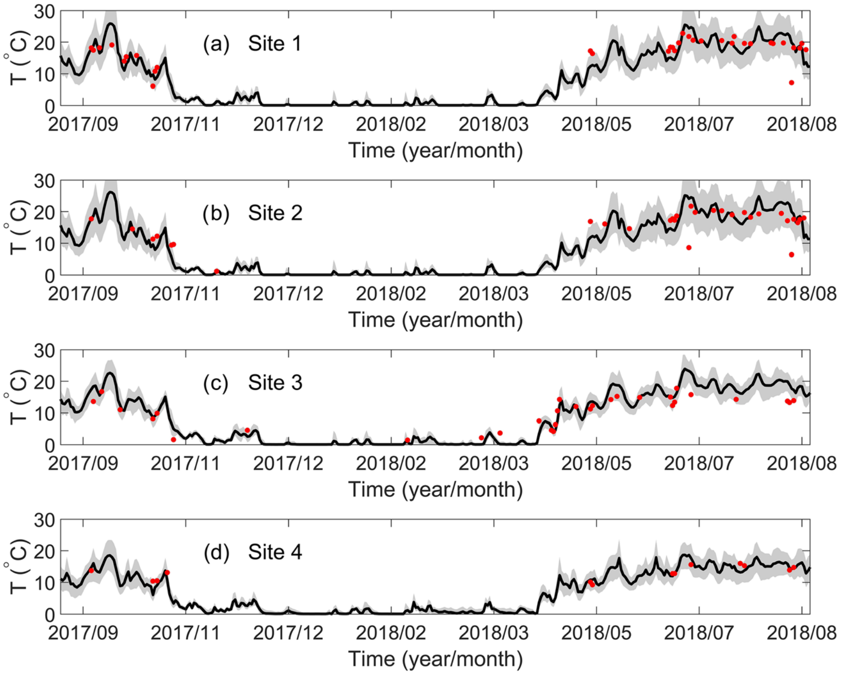

Figure 10. Simulated stream temperature (solid black line), 95% uncertainty ranges (light gray areas), and observations provided by volunteers (solid red dots) at calibration sites and period of data assimilation.

was similar across sites. For example, $d_r > 0$ for all sites and *Bias* ranged between −0.14 and 0.10. For site 1, $E_{f\text{-}mod}$ was −0.17 and −0.05 for the assimilation and validation periods; however, d_r was 0.55 and 0.46 for the same periods, which highlights the ability of the model to reproduce central tendency. For Sites 3 and 4, d_r

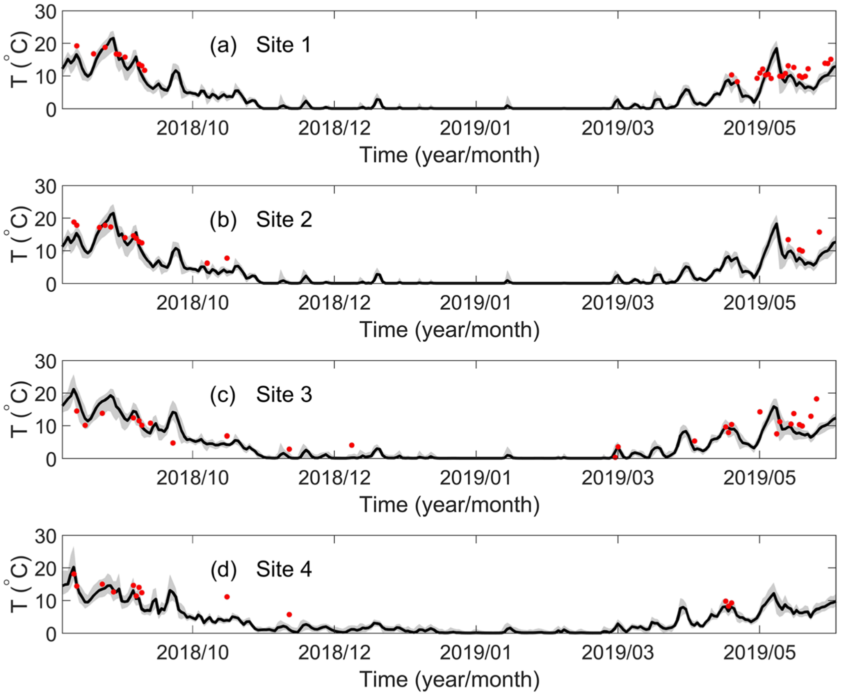

Figure 11. Simulated stream temperature (solid black line), 95% uncertainty ranges (light gray areas), and observations provided by volunteers (solid red dots) for the validation period.

was 0.70 and 0.57 during the assimilation period but was slightly lower during the validation period. There was no consistent trend in *Bias*, but results suggest less overestimation for stream temperature compared to streamflow.

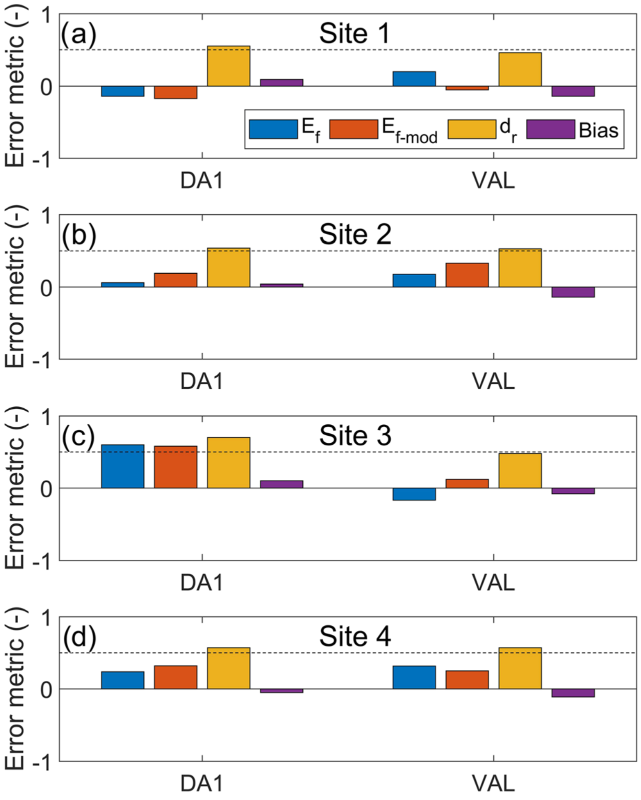

Figure 12. Measures of model performance for the data assimilation (DA1) and validation periods (VAL). The metrics are the Nash-Sutcliffe efficiency (E_f), the modified Nash-Sutcliffe efficiency ($E_{f\text{-}mod}$), the refined index of agreement (d_r), and the relative bias (*Bias*). For visualization, a dashed horizontal line at 0.5 is displayed on each panel.

7. Discussion

We trained a hydrological model for the Boyne River using distributed observations of stream stage and stream temperature collected by volunteers. Observations were collected through CrowdHydrology.com, a platform designed to gather, store, and process text messages containing stream stage and stream temperature observations, at four calibration sites of the Boyne River. Stage-discharge relationships were derived at each calibration site to transform stream stage observations into flows. We used the ensemble Kalman filter to assimilate a 4-year dataset of streamflow and a 1-year dataset of stream temperature, then used a 1-year period for model validation. The novelty of this work lies in the assimilation of sparse, discontinuous, spatially distributed observations collected by volunteers (stream stage and stream temperature) to improve a semidistributed hydrological model.

Our results indicate that observations collected by volunteers improved streamflow and stream temperature simulations when compared to those of an uncalibrated SWAT model (using initial/default SWAT parameters derived from available spatial information). However, different model performance metrics arose at each of the calibration sites. We posit that the interarrival times of the observations contributed to model performance. For example, sites located downstream performed better than sites located upstream, where the median arrival frequency of an observation was nearly double that of sites further down the river. Our results are consistent with numerical experiments using synthetically generated observations (Etter et al., 2018; Mazzoleni et al., 2015, 2017). For example, Etter et al. (2018) reported scenarios in which few synthetically generated observations (12 and 52), distributed throughout a year, were not informative for model calibration of a simple bucket-type runoff model for six Swiss catchments. In their study, the median performance of the models calibrated with such observations was not significantly better than the median performance of the models with random parameter values. We used a different model in this study, with a range of 8–28 (Site 3) and 4–18 (Site 4) stream stage observations per year delivering poor model performance.

Measures of model performance improved, relative to simulations with initial/default SWAT parameters, when more observations per year were used for calibration (Sites 1 and 2). Although performance metrics continued to vary as more observations were assimilated, model performance never reached the level expected for a streamflow model calibrated with continuous daily professional (e.g., US Geological Survey) observations. We speculate that the number of observations to reach a maximum level of performance is site specific due to different arrival frequencies, concentration of observations during a specific season, bias toward a range of streamflow and stream temperature values, and environmental settings (e.g., perennial versus intermittent streams). For example, the Boyne River model was likely trained with low to average streamflow conditions, so the model reproduced central tendency. Had more observations around peak flows been available, model parameters would have changed over time and required more observations to achieve a steady distribution. However, it is not reasonable to expect volunteers to collect streamflow observations during a flood event or icy conditions due to safety reasons. The question of which streamflow observations are more informative for model calibration is still under debate, with some researchers arguing that samples near high flows are more valuable for model calibration (Pool et al., 2017) and others not observing an abrupt change in model performance when considering an increased number of high flow observations (Etter et al., 2018).

It is difficult to assess whether a richer dataset would have generated better performance metrics for the less visited calibration sites. We would need to increase the frequency of observations at these sites and monitor model performance. Our current efforts to increase volunteer engagement include hosting public lectures of project updates to stakeholders, sending a reply text message to participants confirming that each observation was received, exploring motivations for participants to visit a site via interviews and focus groups, partnering with the Chamber of Commerce to promote the project using stickers to reward participation, and developing an interactive website where volunteers can see and tailor visualizations of model results.

Several limitations of our study are important to highlight. First, this study considered only one catchment, which prevented a broader testing of the methods across different hydrological characteristics and social settings. A significant portion of this project involved regular stakeholder engagement with the Boyne River community members in social research as to improve the meaningfulness of modeling products to the community (Hall et al., 2014, 2016). To build this relationship, we conducted interviews, disseminated the project through presentations to members of the community, and promoted the development of a strong social network. As CrowdHydrology transitions to a more mature and sustainable citizen science-based research program (Lowry et al., 2019), we expect to test various hydrological models and calibration methods across different catchments. Second, the data splitting approach for model calibration and validation may have a significant impact on model performance. Although the ensemble Kalman Filter allows for the parameters to change through time, we only tested one subset of data for calibration and a subsequent dataset for validation. Several data splitting methods have been evaluated, focusing on the skewness associated with runoff for the selection of the splitting approach (Zheng et al., 2018). Finally, it will be interesting to explore how a different model would perform when explicitly considering stream stage observations for calibration, rather than estimating streamflow through a stage-discharge relationship. Some studies have demonstrated that stream stage data can be informative for calibration, as well as the use of water level class observations (Etter et al., 2020; Seibert & Vis, 2016).

8. Conclusion and Recommendations

Our findings suggest that observations collected by volunteers can improve the performance of complex hydrologic models. Through a citizen science network (CrowdHydrology), volunteers submitted stream stage and stream temperature observations to guide the calibration of a SWAT model of the Boyne River watershed. We implemented the ensemble Kalman filter to sequentially assimilate uncertain observations and assess the temporal variability of model parameters. After data assimilation, the hydrological model reproduced central tendency for streamflow and stream temperature, although the model missed heavy rainfall or snowmelt events during the spring. Across four calibration sites, better performance metrics emerged in locations with higher interarrival times of the observations. In this citizen science project, observations arrived more frequently at calibration sites located downstream, likely because of the proximity to a more populated area (Boyne City) and accessibility to walking paths.

Due to the nature of the observations, stream flows likely covered low to average flow conditions. Questions remain regarding to what extent the produced models can capture higher streamflow conditions, which could only be evaluated by measuring discharge at the highest point of the hydrograph. Although it is unreasonable to expect volunteers to provide these observations due to safety concerns, alternative low-cost monitoring techniques can be explored and such data integrated with observations collected by volunteers. Future citizen science projects and modeling efforts can provide empirical evidence regarding finer temporal resolutions, coupling with other types of observations (e.g., pictures, videos) and integration with common hydrologic data sources (e.g., government-funded environmental agencies). We expect a continuous growth of the number of observations collected by volunteers due to the active participation of our research partners; thus, our current approach can be further tested over a longer time frame.

Acknowledgments
This research was financially supported by National Science Foundation (NSF) and Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute. We also acknowledge NSF under Grants ABI-1661156 and ABI-1661324. We acknowledge the Friends of the Boyne River (https://boyneriver.org/) and Michigan Trout Unlimited (http://www.michigantu.org/) for their role in community engagement, support during field visits, and maintenance of the CrowdHydrology network (http://www.crowdhydrology.com/). The authors gratefully acknowledge the active participation of volunteers near the Boyne River. The data and SWAT model are available on HydroShare at https://doi.org/10.4211/hs.8b5d02c85c714c639d00667d44fb8625.

References

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. *The Bulletin of the Ecological Society of America*, *98*(4), 278–290. https://doi.org/10.1002/bes2.1336

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

Arnold, JG, Kiniry, J., Srinivasan, R., Williams, J., Haney, E., & Neitsch, S. (2012). Soil and water assessment tool input/output file documentation version 2012. Texas Water Resources Institute Technical Report No. 439, Texas A&M University System, College Station, TX: Texas Water Resources Institute Technical Report No. 439, Texas A&M University System, College Station, TX.

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

Barnhart, B. L., Whittaker, G. W., & Ficklin, D. (2014). Improved stream temperature simulations in SWAT using NSGA-II for automatic multi-site calibration. *Transactions of the ASABE*, 517–530. https://doi.org/10.13031/trans.57.10472

Boyne USA. (2017). Boyne River Hydroelectric Project Pre-Application Document (PAD) FERC PROJECT no. *3409*. Boyne City, MI.

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

Cohn, T. A., Kiang, J. E., & Mason, R. R. (2013). Estimating discharge measurement uncertainty using the interpolated variance estimator. *Journal of Hydraulic Engineering*, *139*(5), 502–510. https://doi.org/10.1061/(ASCE)HY.1943-7900.0000695

Cortes Arevalo, V. J., Charrière, M., Bossi, G., Frigerio, S., Schenato, L., Bogaard, T., et al. (2014). Evaluating data quality collected by volunteers for first-level inspection of hydraulic structures in mountain catchments. *Natural Hazards and Earth System Sciences*, *14*(10), 2681–2698. https://doi.org/10.5194/nhess-14-2681-2014

Cosgrove, W. J., & Loucks, D. P. (2015). Water management: Current and future challenges and research directions. *Water Resources Research*, *51*(6), 4823–4839. https://doi.org/10.1002/2014WR016869

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

Etter, S., Strobl, B., Seibert, J., & Meerveld, H. J. (2020). Value of crowd-based water level class observations for hydrological model calibration. *Water Resources Research*, *56*(2). https://doi.org/10.1029/2019WR026108

Etter, S., Strobl, B., Seibert, J., & van Meerveld, I. (2018). Value of uncertain streamflow observations for hydrological modelling. *Hydrology and Earth System Sciences Discussions*, (July), 1–26. https://doi.org/10.5194/hess-2018-355

Ficklin, D. L., Luo, Y., Stewart, I. T., & Maurer, E. P. (2012). Development and application of a hydroclimatological stream temperature model within the Soil and Water Assessment Tool. *Water Resources Research*, *48*(1). https://doi.org/10.1029/2011WR011256

Ficklin, D. L., Stewart, I. T., & Maurer, E. P. (2013). Effects of climate change on stream temperature, dissolved oxygen, and sediment concentration in the Sierra Nevada in California. *Water Resources Research*, *49*(5), 2765–2782. https://doi.org/10.1002/wrcr.20248

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

Grusson, Y., Sun, X., Gascoin, S., Sauvage, S., Raghavan, S., Anctil, F., & Sáchez-Pérez, J. (2015). Assessing the capability of the SWAT model to simulate snow, snow melt and streamflow dynamics over an alpine watershed. *Journal of Hydrology*, *531*, 574–588. https://doi.org/10.1016/j.jhydrol.2015.10.070

Hall, D. M., Gilbertz, S. J., Anderson, M. B., & Ward, L. C. (2016). Beyond “buy-in”: Designing citizen participation in water planning as research. *Journal of Cleaner Production*, *133*, 725–734. https://doi.org/10.1016/j.jclepro.2016.05.170

Hall, D. M., Lazarus, E. D., & Swannack, T. M. (2014). Strategies for communicating systems models. *Environmental Modelling & Software*, *55*, 70–76. https://doi.org/10.1016/j.envsoft.2014.01.007

Han, E., Merwade, V., & Heathman, G. C. (2012). Implementation of surface soil moisture data assimilation with watershed scale distributed hydrological model. *Journal of Hydrology*, *416-417*, 98–117. https://doi.org/10.1016/j.jhydrol.2011.11.039

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

Jackson, E. K., Roberts, W., Nelsen, B., Williams, G. P., Nelson, E. J., & Ames, D. P. (2019). Introductory overview: Error metrics for hydrologic modelling—A review of common practices and an open source library to facilitate use and adoption. *Environmental Modelling & Software*, *119*, 32–48. https://doi.org/10.1016/j.envsoft.2019.05.001

Jalowska, A. M., & Yuan, Y. (2019). Evaluation of SWAT impoundment modeling methods in water and sediment simulations. *JAWRA Journal of the American Water Resources Association*, *55*(1), 209–227. https://doi.org/10.1111/1752-1688.12715

Jollymore, A., Haines, M. J., Satterfield, T., & Johnson, M. S. (2017). Citizen science for water quality monitoring: Data implications of citizen perspectives. *Journal of Environmental Management*, *200*, 456–467. https://doi.org/10.1016/j.jenvman.2017.05.083

Kiang, J. E., Cohn, T. A., & Mason, R. R. Jr. (2009). Quantifying uncertainty in discharge measurements: A new approach. In *World environmental and water resources congress 2009*, (pp. 1–8). Reston, VA: American Society of Civil Engineers. https://doi.org/10.1061/41036(342)599

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

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*, 766–777. https://doi.org/10.1016/j.jhydrol.2016.07.036

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

Legates, D. R., & McCabe, G. J. (1999). Evaluating the use of “goodness-of-fit” Measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, *35*(1), 233–241. https://doi.org/10.1029/1998WR900018

Liu, F. (2000). Bayesian time series: Analysis methods using simulation-based computation. Phd Thesis. Durham, North Carolina: Duke University.

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*, 7. https://doi.org/10.3389/feart.2019.00128

Mazzoleni, M., Alfonso, L., Chacon-Hurtado, J., & Solomatine, D. (2015). Assimilating uncertain, dynamic and intermittent streamflow observations in hydrological models. *Advances in Water Resources*, *83*, 323–339. https://doi.org/10.1016/j.advwatres.2015.07.004

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

McKinley, D. C., Miller-Rushing, A. J., Ballard, H. L., Bonney, R., Brown, H., Cook-Patton, S. C., et al. (2017). Citizen science can improve conservation science, natural resource management, and environmental protection. *Biological Conservation*, *208*, 15–28. https://doi.org/10.1016/j.biocon.2016.05.015

Monteith, J. L. (1965). Evaporation and environment. *Symposia of the Society for Experimental Biology*, *19*, 205–34. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/5321565

Moradkhani, H., Sorooshian, S., Gupta, H. V., & Houser, P. R. (2005). Dual state-parameter estimation of hydrological models using ensemble Kalman filter. *Advances in Water Resources*, *28*(2), 135–147. https://doi.org/10.1016/j.advwatres.2004.09.002

Neitsch, S., Arnold, J., Kiniry, J., & Williams, J. (2011). Soil & Water Assessment Tool theoretical documentation version 2009. *Texas Water Resources Institute*, 1–647. https://doi.org/10.1016/j.scitotenv.2015.11.063

Pathiraja, S., Marshall, L., Sharma, A., & Moradkhani, H. (2016). Hydrologic modeling in dynamic catchments: A data assimilation approach. *Water Resources Research*, *52*(5), 3350–3372. https://doi.org/10.1002/2015WR017192

Patil, A., & Ramsankaran, R. (2017). Improving streamflow simulations and forecasting performance of SWAT model by assimilating remotely sensed soil moisture observations. *Journal of Hydrology*, *555*, 683–696. https://doi.org/10.1016/j.jhydrol.2017.10.058

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

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

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*, 7. 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? *Hydrological Processes*, *30*(14), 2498–2508. https://doi.org/10.1002/hyp.10887

Stepenuck, K. F., & Genskow, K. D. (2018). Characterizing the breadth and depth of volunteer water monitoring programs in the United States. *Environmental Management*, *61*(1), 46–57. https://doi.org/10.1007/s00267-017-0956-7

Strobl, B., Etter, S., van Meerveld, I., & Seibert, J. (2019). Accuracy of crowdsourced streamflow and stream level class estimates. *Hydrological Sciences Journal*, *65*(5), 823–841. https://doi.org/10.1080/02626667.2019.1578966

The Tip of the Mitt Watershed Council. (2012). Lake Charlevoix Watershed Management Plan. Petoskey, MI.

USDA. (1986). Urban hydrology for small watersheds. *SCS Technical Release 55*. Washington, D.C.

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

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

Vrugt, J. A., Diks, C. G. H., Gupta, H. V., Bouten, W., & Verstraten, J. M. (2005). Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation. *Water Resources Research*, *41*(1), 1–17. https://doi.org/10.1029/2004WR003059

Vrugt, J. A., Gupta, H. V., Nualláin, B., & Bouten, W. (2006). Real-time data assimilation for operational ensemble Streamflow forecasting. *Journal of Hydrometeorology*, *7*(3), 548–565. https://doi.org/10.1175/JHM504.1

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

Willmott, C. J., Robeson, S. M., & Matsuura, K. (2012). A refined index of model performance. *International Journal of Climatology*, *32*(13), 2088–2094. https://doi.org/10.1002/joc.2419

Willmott, C. J., Robeson, S. M., Matsuura, K., & Ficklin, D. L. (2015). Assessment of three dimensionless measures of model performance. *Environmental Modelling & Software*, *73*, 167–174. https://doi.org/10.1016/j.envsoft.2015.08.012

Winchell, M., Srinivasan, R., Di Luzio, M., & Arnold, J. (2007). ArcSWAT interface for SWAT 2005: User's guide.

World Meteorological Organization. (2010). Manual on stream gauging. *Fieldwork WMO-No. 1044*. Geneva, Switzerland.

Wu, K., & Johnston, C. A. (2007). Hydrologic response to climatic variability in a Great Lakes Watershed: A case study with the SWAT model. *Journal of Hydrology*, *337*(1–2), 187–199. https://doi.org/10.1016/j.jhydrol.2007.01.030

Xie, X., & Zhang, D. (2013). A partitioned update scheme for state-parameter estimation of distributed hydrologic models based on the ensemble Kalman filter. *Water Resources Research*, *49*(11), 7350–7365. https://doi.org/10.1002/2012WR012853

Yang, P., & Ng, T. L. (2017). Gauging through the crowd: A crowd-sourcing approach to urban rainfall measurement and storm water modeling implications. *Water Resources Research*, *53*(11), 9462–9478. https://doi.org/10.1002/2017WR020682

Zhang, H., Hendricks Franssen, H.-J., Han, X., Vrugt, J., & Vereecken, H. (2016). State and parameter estimation of two land surface models using the ensemble Kalman filter and particle filter. *Hydrology and Earth System Sciences Discussions*, 1–39. https://doi.org/10.5194/hess-2016-42

Zhang, Y., Hou, J., Gu, J., Huang, C., & Li, X. (2017). SWAT-based hydrological data assimilation system (SWAT-HDAS): Description and case application to river basin-scale hydrological predictions. *Journal of Advances in Modeling Earth Systems*, *9*(8), 2863–2882. https://doi.org/10.1002/2017MS001144

Zheng, F., Maier, H. R., Wu, W., Dandy, G. C., Gupta, H. V., & Zhang, T. (2018). On lack of robustness in hydrological model development due to absence of guidelines for selecting calibration and evaluation data: Demonstration for data-driven models. *Water Resources Research*, *54*(2), 1013–1030. https://doi.org/10.1002/2017WR021470

Legates, D. R., & McCabe, G. J. (1999). Evaluating the use of "goodness-of-fit" Measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, *35*(1), 233–241. https://doi.org/10.1029/1998WR900018