# The potential of proxy water level measurements for calibrating urban pluvial flood models

Matthew Moy de Vitry [a, b, *], João P. Leitão [a]

a: Eawag: Swiss Federal Institute of Aquatic Science and Technology, Überlandstrasse 144, 8600 Dübendorf, Switzerland

b: Institute of Civil, Environmental and Geomatic Engineering, ETH Zürich, Stefano-Franscini-Platz 5, 8093 Zurich, Switzerland

*Corresponding author: matthew.moydevitry@eawag.ch

**Keywords**: urban pluvial flooding; proxy measurements; flood monitoring; model calibration; measurement error; sensor placement

**Abstract:** Urban pluvial flood models need to be calibrated with data from actual flood events in order to validate and improve model performance. Due to the lack of conventional sensor solutions, alternative sources of data such as citizen science, social media, and surveillance cameras have been proposed in literature. Some of the methods proposed boast high scalability but without an on-site survey, they can only provide proxy measurements for physical flooding variables (such as water level). In this study, the potential value of such proxy measurements was evaluated by calibrating an urban pluvial flood model with data from experimental flood events conducted in a 25x25 meter facility, monitored with surveillance cameras and conventional sensors in parallel. Both ideal proxy data and actual image-based proxy measurements with noise were tested, and the effects of measurement location and measurement noise were investigated separately. The results with error-free proxy data confirm the

theoretic potential of such measurements, as in half of the calibration configurations tested, ideal proxy data increases model performance by at least 70% compared to sensor data. The presence of complex correlated errors, which has a complex but predominantly negative effect on performance.

# 1 Introduction

## 1.1 The need for flood monitoring data

Urban pluvial floods, also known as nuisance floods or flash floods, occur when a city's drainage system does not have the capacity to drain local rainfall during storms. As climate change increases the frequency of extreme rainfall events (Field et al., 2012), it is expected that the frequency and intensity of urban pluvial floods will also increase. Such flooding is further aggravated by the reduction of pervious surfaces due to urbanization (Skougaard Kaspersen et al., 2017). According to certain studies, the societal cost of urban pluvial flooding is comparable to that of coastal or river flooding, because these tend to occur less frequently (Jiang et al., 2018; ten Veldhuis, 2011). The issue of urban pluvial floods is of growing importance and demands adequate planning and mitigation tools.

To assess the risk of urban pluvial flooding and design prevention solutions, urban drainage experts use numerical models tools to simulate the flow and accumulation of water in urban catchments. These models can be of varying complexity depending mainly on how surface flow is represented, the completeness of the flow equations solved (if any), and the complexity of sewer-surface interactions (if any). Overviews of the different numerical models used for urban flood modelling can be found in literature (Ochoa-Rodriguez et al., 2015; Zoppou, 2001).

Regardless of the model used, there are often parameters with poorly defined values, which can introduce uncertainty in the modelling results. But despite the recognized necessity of calibration to reduce these uncertainties, calibration of urban drainage models is rarely performed in practice (Tscheikner-Gratl et al., 2016). The probable cause for this is a lack of monitoring data (Thorndahl et al., 2008), a problem that is aggravated when it comes to measuring surface flooding. Conventional sensors designed for sewer pipes or well-defined channels are not suited to the open urban environment with its complex geometries, moving objects, and risk of vandalism.

## 1.2 Proxy water level measurements from images

The difficulty in monitoring urban pluvial floods with conventional sensors has encouraged researchers to explore alternative sources of flooding data such as social media (Assumpção et al., 2018; Chaudhary et al., 2019; Hénonin et al., 2015; Wang et al., 2018) or surveillance cameras (Bhola et al., 2019; Jiang et al., 2019; Leitão et al., 2018; Lv

et al., 2018; Moy de Vitry et al., 2019a). Of the different methods proposed, that of Moy de Vitry et al. (2019a) stands out because the measurements obtained do not correspond directly to a physical flooding variable, but instead contain information on how the water level changes over time. The method exploits the assumption that the amount of visible water in the images of a static surveillance camera is associated with the actual flooding water level. Concretely, the method uses a deep convolutional neural network (DCNN) to periodically segment water in images from a static surveillance camera. The area covered by water in each image is expressed as a fraction of the whole image and called the static observer flood index (SOFI). When the method was proposed, it was tested on a range of different surveillance video qualities and contexts (outdoors and indoors, during night and day, etc.). The initial results showed that SOFI data had an average correlation of 75% with the actual water level, although the variability between videos was high (minimum of around 35%, maximum around 90%). This correlation, though imperfect, established SOFI as a proxy measurement for flood water level. The original publication describes the method and results in detail and lists measures by which the quality of the SOFI measurements can be improved.

Although the information type provided by SOFI is non-standard, the method has a potential advantage in scalability that justifies further study: in principle, the assumption SOFI relies on (association of water level with visible water) is so general that SOFI can be applied to any static surveillance camera footage without the need for any on-site measurements. Thus, SOFI is well suited to very large networks of surveillance cameras, especially those for which privacy is an issue and human access to the footage is restricted.

## 1.3 Hydraulic and hydrologic model calibration with proxy measurements

A measurement can be considered a proxy for a physical variable if it has a correlation with that variable. In urban pluvial flood modelling, proxy measurements could be obtained via surveillance cameras, as described in Section 1.2, or from conventional sensors that were not correctly calibrated. Proxy measurements differ from conventional sensor data in that they contain information not in the absolute values of the data but in the shape of the data series. As proxy measurements cannot be directly compared with the values of a model simulation, metrics like the Nash-Sutcliffe efficiency, which is commonly used for model calibration, cannot be used. Despite this challenge,

the lack of suitable sensing techniques motivates the investigation of how proxy measurements could help calibrate urban pluvial flood models.

While there are no known examples from urban flood modelling, model calibration with proxy measurements has been performed in stream hydrology, where van Meerveld et al. (2017) investigated the value of stream level classes reported by citizen scientists. The authors successfully used the Spearman rank correlation coefficient (Spearman, 1904) as an objective function for calibration, to maximize correlation between observed stream level classes and the linked but not equivalent stream discharge represented as a state variable. The Spearman rank correlation has also been used in another study (Jian et al., 2017) to calibrate a hydrologic model with water level data without a rating curve, which is an equivalent problem.

## 1.4 Objective of the current study

The objective of this study is to assess the value of proxy measurements, as compared to data from conventional sensors, when calibrating an urban pluvial flood model. The motivation to investigate proxy measurements from alternative data sources is the need for monitoring data for numeric model calibration and the lack of conventional sensor solutions. In particular, the study uses the SOFI method (Moy de Vitry et al., 2019a) as a possible source of proxy measurements. Using a large lab-like setup, the study first investigated the potential value of ideal, noise-free proxy data under different measurement configurations. Second, the study investigated the impact that noise in proxy measurements have on model performance after calibration. Noise with a complex correlated structure and with a Gaussian distribution were considered separately. Due to the restricted dimensions of the setup and the experimental nature of the flooding events simulated, the aim of this study was primarily of exploratory nature, to assess the novel concept of urban pluvial flood model calibration with proxy water level measurements.

# 2 Material and methods

## 2.1 Experimental urban catchment for flood events

The data used in this study was collected in a 25x25 meter lab-like facility. The facility is a simplified urban catchment with a drainage network and a small building with a basement that can be flooded (Fig. 1). Designed for training civil protection forces for flood events, the facility was temporarily outfitted with sensors and cameras to collect unique datasets that were documented and shared in previous work (Moy de Vitry et al., 2017). The datasets provide both conventional monitoring data and surveillance video documentation for multiple flash flood events. Despite the facility's limited size, it was possible to reproduce different phenomena common to urban pluvial flooding such as shallow overland flow, manhole overflow, ponding, and basement flooding. However, rainfall-runoff processes could not be reproduced in the catchment.

From the original experiments, four sensors and one camera were selected for inclusion in this study. A magnetic inductive flow meter provides data about the flow of water into the system and three ultrasonic water level sensors provide flood level data at three separate locations of the catchment. The ultrasonic sensors and camera provide the calibration data for this study as described in Sections 2.2, 2.3, and 2.6.

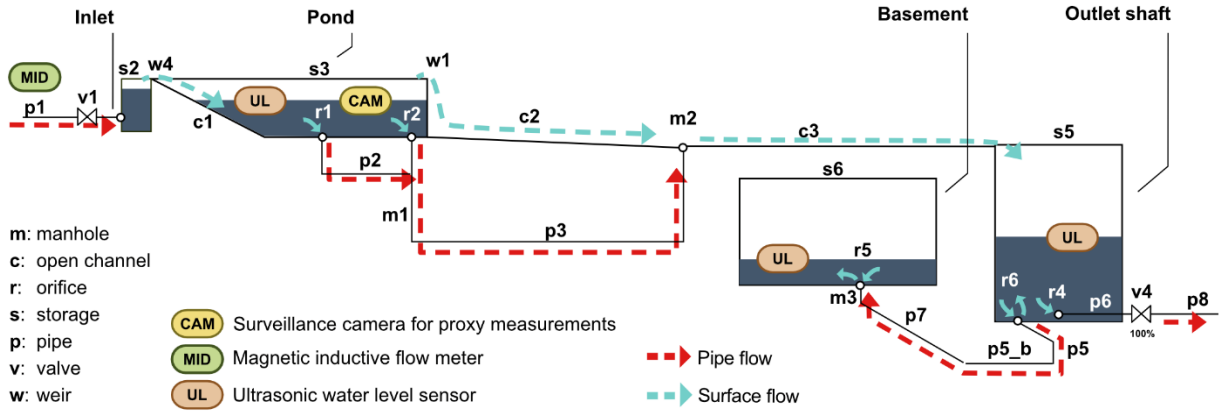


**Figure 1: Hydraulic diagram of the experimental facility in which this study was conducted, including the locations of sensors used and component codes used for modelling.**

Each flood event consisted in manually regulating a valve to let water from a reservoir into the facility. For the present study, five flood events with different characteristics were used (Tab. 1 and Fig. 2). The flood events were selected based on sensor availability, lighting conditions for the cameras, and the occurrence of flooding in the basement.

**Table 1: Characteristics of the flood events used in this study.**

| | Duration | Volume [$m^3$] | Flooding in basement | original event ID[1] |
|---|---|---|---|---|
| Event 1 | 00:16:00 | 33.3 | Yes | 20 |
| Event 2 | 00:17:00 | 26.5 | No | 21 |
| Event 3 | 00:20:40 | 34.5 | Yes (minimal) | 22 |
| Event 4 | 00:19:00 | 31.4 | No | 23 |
| Event 5 | 00:26:00 | 77.1 | Yes | 24[1] |

1: Identifier used in published datasets (Moy de Vitry et al., 2017)
2: Experiment 24 is not listed for use in published datasets (Moy de Vitry et al., 2017) because a plastic cone rolled in front of surface flow sensor during the experiment. The affected surface flow data is not used in the current study.

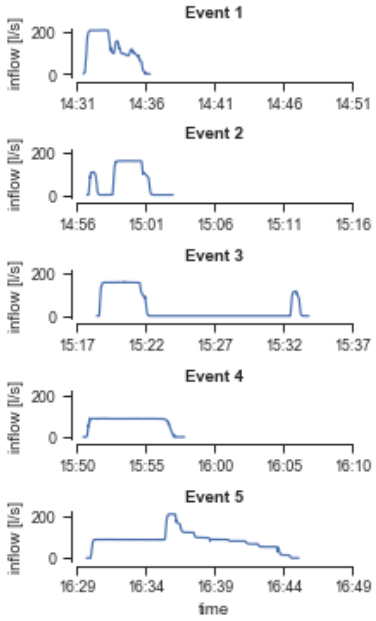


**Figure 2: Hydrographs of the five flood events used in study. The hydrographs represent the flow into the facility measured with a magnetic-inductive flow meter.**

## 2.2 Conventional sensor data

Data from five conventional sensors were used to support this study. Flow into the facility was measured with a magnetic-inductive flow meter situated in the pipe leading to the inlet of the flood facility. This inflow data was used as input data for the SWMM model. In addition, three ultrasonic rangefinders provide water level at three locations where accumulation would occur: the pond, basement, and outlet shaft. The error of these sensors is assumed to be negligible. Ultrasonic water level data is both used as reference data and is degraded into ideal and noisy proxy data. The characteristics of each sensor are listed in Tab. 2.

**Table 2: Conventional sensors used to collect data for calibrating the urban pluvial flooding model.**

| Location | Sensor type | Variable | Frequency | Purpose |
|---|---|---|---|---|
| Inflow | Magnetic-inductive flow meter[1] | Discharge [l/s] | 1 Hz | Model input |
| Pond | Ultrasonic rangefinder[2] | Water depth [m] | ~5 Hz[3] | Calibration / validation |
| Basement | Ultrasonic rangefinder[2] | Water depth [m] | ~5 Hz[3] | Calibration / validation |
| Outlet shaft | Ultrasonic rangefinder[2] | Water depth [m] | ~5 Hz[3] | Calibration / validation |

1: Endress+Hauser Proline Promag 53P
2: Maxbotix MB7369
3: Sampling frequency is variable

## 2.3 Proxy water level measurements

In this study, the SOFI measurement method introduced by Moy de Vitry et al. (2019a) and briefly described in Section 1.2 was investigated as a proxy for water level. Different qualities of SOFI data, both real and hypothetical, are explored. Table 3 provides an overview of the data series created, which are explained in the following sections.

141 **Table 3: Proxy data used to calibrate pluvial flood model.**

| Name | Quality[1] | Error type | Data source | Locations |
|---|---|---|---|---|
| IDEAL-100 | 100% | Negligible | normalized sensor data | Pond, basement, outlet |
| RAW[2] | 40-80% | Complex correlated | surveillance footage[3] | pond |
| GAU-60 | 60% | Gaussian | synthetic[4] | pond |
| GAU-70 | 70% | Gaussian | synthetic[4] | pond |
| GAU-80 | 80% | Gaussian | synthetic[4] | pond |
| GAU-90 | 90% | Gaussian | synthetic[4] | Pond |
| COR-60 | 60% | Complex correlated | synthetic[5] | pond |
| COR-70 | 70% | Complex correlated | synthetic[5] | pond |
| COR-80 | 80% | Complex correlated | synthetic[5] | pond |
| COR-90 | 90% | Complex correlated | synthetic[5] | pond |

142
143 1: Quality is measured by the Spearman rank correlation with the sensor data. There are slight variations (<2%) between events due to the
144 method for generating synthetic proxy data (see Section S1 in the supporting information for details).
145 2: The RAW SOFI data is not used for calibration directly since its quality is highly variable (see supporting information for details).
146 3: Extracted from footage after Moy de Vitry et al. (2019)
147 4: Linear combination of IDEAL-100 and Gaussian noise
148 5: Linear combination of IDEAL-100 and RAW

### 149 *2.3.1 Raw proxy measurements from surveillance cameras*

150 Raw proxy measurements (RAW) were obtained from surveillance footage of the pond, using the SOFI method
151 described in Section 1.2 and by Moy de Vitry et al. (2019a). As can be seen in the example frame provided in Fig. 4,
152 the image quality did not always allow clear differentiation between flooded and wet surfaces. This ambiguity is
153 reflected in the results of the DCNN prediction and in the proxy measurements, where noise with a complex
154 correlation structure is visible (Fig. 3f). The Spearman correlation of the raw proxy measurements with the sensor
155 data varies between 40% and 80% from event to event (see Tab. S1 in the supporting information).

### 156 *2.3.2 Synthetic ideal proxy data*

157 Ideal proxy data (IDEAL-100) was created by normalizing water level sensor data, thus retaining 100% correlation
158 with the water level but losing all absolute water level information. It represents “perfect” proxy measurements
159 that could be obtained if the image segmentation was improved to be error-free and if moving visual obstructions
160 could be avoided (Fig. 3a). Although proxy data of such high quality may not be obtainable from surveillance
161 cameras and automatic image analysis, such data could come from an uncalibrated water level sensor within the

sewer network. For the purpose of this study, these ideal synthetic proxy data allow the theoretic potential of proxy measurements for model calibration to be assessed.

### *2.3.3 Synthetic proxy data with complex correlated noise*

To assess the impact of noise on the utility of proxy data for calibration, the level of noise in the raw proxy measurements was modified to predefined levels (COR-60 to COR-90, see Tab. 3). The modification was performed by linear combination of the raw proxy measurements with the ideal proxy data to achieve synthetic proxy data with the same noise structure as the raw proxy measurements (Fig. 3g-j). The method for combining the raw and ideal proxy measurements is described in the supporting information (Section S1).

### *2.3.4 Synthetic proxy data with Gaussian noise*

Synthetic proxy data with pure Gaussian noise were created to provide further insight into the role of error structure on model calibration (GAU-60 to GAU-90, see Tab. 3). These synthetic proxy data (Fig. 3b-e) were created by linear combinations of Gaussian distributed noise with the ideal proxy data (as described in the supporting information in Section S1). The synthetic proxy data Gaussian noise were designed to have the same Spearman correlation with the sensor data as the synthetic proxy data with complex correlated noise.

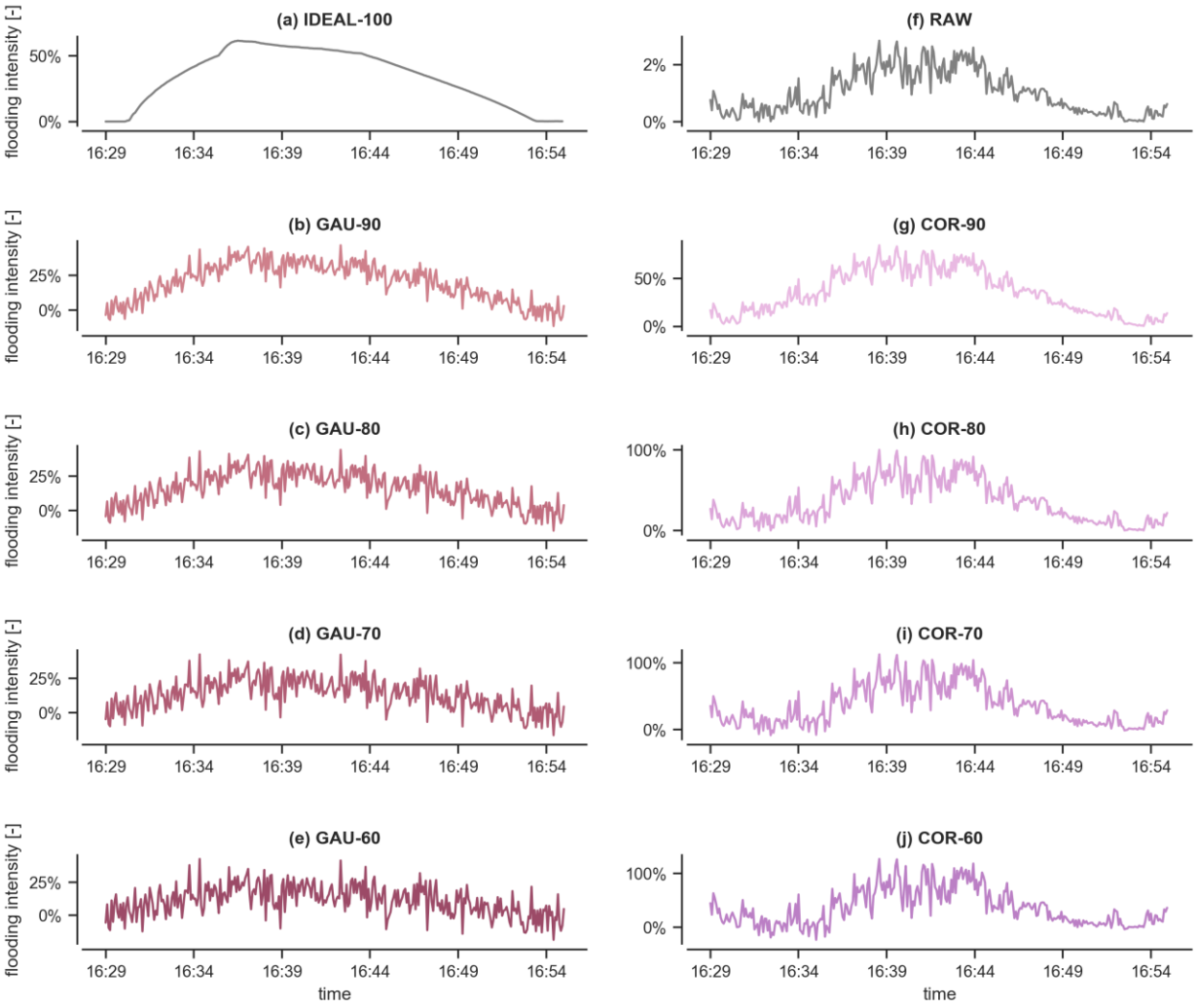


**Figure 3: Proxy water level measurements for flood event 5. For example, GAU-90 is synthetic proxy data generated by combining the ideal proxy data with Gaussian noise such that the resulting signal has a 90% correlation with the sensor data (see Tab. 3). (a) Ideal synthetic proxy data, (b-e) synthetic proxy data with Gaussian noise, (f) raw proxy measurements derived from surveillance footage after Moy de Vitry et al. (2019a), and (g-j) synthetic proxy data with the same correlated noise as the raw proxy measurements. The percentages on the y-axis indicate the Spearman correlation with sensor data. Data series for flood events 1-4 can be found in the supporting information.**


**Figure 4: Surveillance camera video frame of the pond. The wet ground and lighting make it difficult to distinctly identify floodwater.**

## 2.4 Hydraulic modelling with EPA SWMM

To model the flood events, the EPA SWMM 5.1 software (Rossman, 2010) was used for both the drainage system and the surface flows. EPA SWMM (aka SWMM) is an established urban drainage modelling software that has the additional advantage of being open source, which facilitates dissemination of this work. SWMM only allows creation of one-dimensional models, but this was not critical thanks to the simplicity of the flooding in the experimental catchment. The surface flows, which had mostly been channeled with sandbags, were approximated as conduit links in the model. Depressions where water could pond were modeled as storage nodes with appropriate depth-area curves. The high similarity between simulations and measurements confirmed that the use of a 1D model was acceptable.

## 2.5 Model calibration and evaluation

### *2.5.1 Calibration parameters*

Seven parameters were selected for calibration based on the uncertainty or heuristic nature of their values. These parameters, including their respective value ranges, are provided in Tab. 4. These parameters were selected for calibration because their values were not measured or could not be directly measured during the experiments. The roughness of pipe p3 is considered separately from that of other pipes because pipe p3 has sharp bends that could increase flow resistance.

**Table 4: SWMM calibration parameters and value ranges.**

| Component | Property | Unit | Lower limit | Upper limit |
|---|---|---|---|---|
| Weir w1 | Height | m | 0.4 | 0.6 |
| Weir w1 | Discharge coefficient | $m^3\ s^{-1}\ m^{-1}$ | 1.1 | 2.1 |
| Pipe p3 | Manning's roughness coefficient | $s\ m^{-1/3}$ | 0.009 | 0.03 |
| All other pipes | Manning's roughness coefficient | $s\ m^{-1/3}$ | 0.005 | 0.02 |
| Orifice r4 | Discharge coefficient | - | 0.36 | 0.72 |
| Manhole m1 | Discharge coefficient | - | 0.48 | 0.72 |
| Manhole m3 | Discharge coefficient | - | 0.42 | 0.78 |

### *2.5.2 Calibration algorithm*

Calibration was conducted with the Shuffled Complex Evolution - University of Arizona (SCE-UA) algorithm (Duan et al., 1993), a global search algorithm implemented in the SPOTPY Python package (Houska et al., 2018). The SCE-UA algorithm begins with points sampled randomly in the parameter space and divided into groups (complexes). Optimization is conducted in cycles, wherein the complexes are incrementally optimized in parallel (complex competitive evolution). At the end of each evolution cycle, if stopping criteria are not met, then information is shared (shuffled) between all complexes and a new cycle begins. The competitive evolution gives SCE-UA efficiency and the complex shuffling exploits the information contained in the initial population. When the stopping criteria are met, the parameter combination that gave the best performance during the whole calibration procedure is retained. With the 10-fold repetition of each calibration exercise, a qualitative understanding of uncertainty can be gained. SCE-UA does not have parameters that need to be "tuned" (unlike algorithms such as simulated annealing), which is favorable for conducting automatic calibration on different combinations of data.

### *2.5.3 Objective function for proxy data*

The original Spearman rank-order correlation coefficient (Spearman, 1904) measures the degree of association between two synchronous signals and was previously used for the study presented by van Meerveld et al. (2017)

to assimilate categorical proxy measurements for the water level of rivers. The Spearman rank-order correlation coefficient (or Spearman correlation) is suited for comparisons between signals of differing orders of magnitude, even with non-linear relationships. The coefficient is computed by first determining the relative rank of each signal value and computing the Pearson correlation coefficient between the ranks. For signals with duplicate values, ranks can be tied and the Spearman correlation $\rho$ is then given by

$$\rho = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}},$$

where $x_i$ and $y_i$ are the ranks of the two signals for time step $i$, and where $\bar{x}$ and $\bar{y}$ are the mean ranks of the SOFI and water level signals, respectively. A Spearman correlation of 1 indicates a perfectly monotonic increasing relationship between two signals, whereas a correlation of 0 indicates absence of correlation (positive or negative) between two signals. In the present study, the implementation of the Spearman correlation of the pandas Python library (McKinney, 2010) was used.

The Spearman correlation cannot be computed if all data points in one of the series has the same rank, as this leads to a division by zero. This was problematic for flood events 2 and 4 of this study, for which no flooding occurs in the basement, leading to all-zero values in the water level. In literature, the issue of undefined objective functions has been solved by modifying the objective function to remove singularities (Haupt et al., 2009). This solution is not satisfactory for the current problem, because although it would be possible to add terms to avoid singularity, the resulting function would have a constant value in situations without flooding. Thus, for flood events 2 and 4, calibration with proxy data from the basement would not be possible.

The solution found was to use the intersection over union (IoU) to evaluate the agreement between the proxy and sensor data when either signal has only zero values:

$$IoU = \frac{N_{s0\ \&\ t0}}{N_{s0\ |\ t0}}$$

where $N_{s0\ \&\ t0}$ is the number of time steps for which both the sensor data and proxy measurements have values of zero, and $N_{s0\ |\ t0}$ is the number of time steps for which either the sensor data or the proxy data have values of

zero. The IoU also varies between 0 (no agreement) and 1 (perfect agreement). In this study, a hybrid Spearman correlation that combines the IoU and Spearman correlation was used to evaluate model simulations with proxy data:

$$\rho_{IoU} = \begin{cases} IoU, & if\ any\ signal\ has\ only\ zero\ values \\ \rho, & if\ neither\ signal\ has\ only\ zero\ values \end{cases}$$

Again, this metric reaches a value of 1 for "perfect" correspondence between the two signals. Another possible solution to the problem of signals with only zero values would have been to append identical non-zero values to both signals.

### *2.5.4 Aggregated objective function*

The SCE-UA algorithm is a single-objective search algorithm, so when calibrating with multiple data sources the objective functions need to be aggregated. Aggregation of multiple objective functions is common in hydrological modeling. Madsen (2003) was among the first to demonstrate this approach for a catchment with multiple measurement locations and measurement types. More recently, Garcia et al. (2017) combined two objective functions to fit both high and low flow regimes. Vis et al. (2015) also used function aggregation, with the particularity of combining up to four objective functions of different types, the Spearman correlation being among them. In all cases, the aggregate objective functions lead to satisfactory model calibration.

The aggregated objective function used in this study is a weighted sum of the root mean square of errors (RMSE) for available sensor data and the hybrid Spearman correlation ($\rho_{IoU}$) for proxy data. Since the SCE-UA algorithm performs minimization of the objective function, the hybrid Spearman correlation was inverted so the best attainable value is zero. Additionally, the hybrid Spearman correlations were given a weight of 0.5 because the expected range of the Spearman correlation is about twice that of RMSE values:

$$OF = \sum_i RMSE_i + 0.5 * \sum_j 1 - \rho_{IoU_j}$$

Where $i$ represents locations where sensor data is available, and $j$ represents locations where proxy data are available. The weight of 0.5 was validated by comparing the distributions of the combined RMSE terms ($\mu$=0.27,

263 $\sigma$=0.14) and the weighted Spearman terms ($\mu$=0.36, $\sigma$=0.11) for flood event 1, with all measurement locations and
264 uniformly sampled parameters.

### 265 *2.5.5 Convergence criteria and computation cost*

266 The SCE-UA algorithm was run with seven simultaneous complex evolutions, mirroring the number of calibration
267 parameters. Calibration was considered successful if the value of the objective function did not improve more than
268 0.5% over five consecutive evolution cycles. The algorithm was capped by an upper limit of 2000 model runs, but
269 this limit was rarely reached. A typical model calibration took approximately 1500 seconds on a computer with 8
270 GB of RAM and an Intel® Core™ i7-3770K 3.5GHz processor.

### 271 *2.5.6 Evaluation of model performance with benchmarks*

272 Model prediction error was evaluated with the sum of RMSEs computed by comparing the simulated water level
273 with the sensor data at the pond, basement, and outlet shaft.

274 $$\sum RMSE = RMSE_{pond} + RMSE_{basement} + RMSE_{outlet}$$

275 While in practice the different errors are usually weighed based on the desired application of the model, the
276 present model did not have a clearly defined application and differentiated weighting would be arbitrary.

277 In order to evaluate model performance despite flooding events of different magnitudes, a model performance
278 index was used to normalize each of the five flood experiments separately. Doing so makes it possible to aggregate
279 the performance of multiple events without having one event with consistently large errors dominate the results.
280 The performance index chosen, described by Seibert et al. (2018), uses an upper benchmark that represents the
281 best attainable model performance with the available monitoring data, and a lower benchmark that represents the
282 expected model performance attainable without monitoring data.

283 In this study, the upper benchmark was computed using sensor data for water level at all three measurement
284 locations, and the lower benchmark was computed as the median validation performance of 100 uncalibrated
285 model realizations, where parameter values were sampled with Latin Hypercube Sampling (LHS). The normalized

performance $p_{i,j}$ of a model calibrated with a combination $i$ of data sources (see Tab. 5 for the list of possible combinations) and a flood event j can then be written as:

$$p_{i,j} = \frac{\sum RMSE_{lower,j} - \sum RMSE_{i,j}}{\sum RMSE_{lower,j} - \sum RMSE_{upper,j}}$$

This normalized model performance $p_{i,j}$ takes values between 0 and 1, with 0 corresponding to the performance of an uncalibrated model, and 1 corresponding to the performance of a model calibrated with all available data for that experiment. The normalized model performance is always evaluated on events that were not used for model calibration.

Using the normalized performance, two aggregate performance indicators can be defined. First, the median model performance ($MP_i$) is the median normalized model performance of one or multiple combinations $i$ of data sources, and all flood events.

$$MP_i = median(\{p_{i,j} : j \in (1, \dots, 5)\})$$

Second, the marginal performance increase ($MPI_{i,k}$) is the median change in performance $MP$ when adding an additional data source $k$ (either sensor data or proxy data) to a given combination $i$ of data sources. The MPI provides a way to quantify the benefit of additional sensor or proxy data to a system in which some monitoring data is already available.

## 2.6 Calibration experiments

### *2.6.1 Calibration with ideal proxy data*

In these experiments, the value of calibrating with ideal (noise-free) water level proxy data was compared to the value of calibrating with actual water level sensor data, holding all other experimental variables like sensor location and measurement errors constant. To perform this comparison, the SWMM model was calibrated with different combinations of either sensor data or the corresponding proxy data at the three monitoring locations (pond, basement, or outlet shaft). For example, proxy data from the pond and sensor data from the outlet shaft are one possible combination, but it was not allowed to calibrate with both proxy data and sensor data from the

309 pond at the same time. The full list of 26 possible combinations can be found in Tab. 5. For each combination, the
310 model was calibrated 10 times on each of the five flood events. Each calibrated model was validated with the four
311 other flood events.

**Table 5: List of different data source combinations that were used to calibrate the flood model.**

| | |
|---|---|
| pond (proxy) | pond (sensor), outlet (sensor) |
| pond (sensor) | basement (proxy), outlet (proxy) |
| basement (proxy) | basement (proxy), outlet (sensor) |
| basement (sensor) | basement (sensor), outlet (proxy) |
| outlet (proxy) | basement (sensor), outlet (sensor) |
| outlet (sensor) | pond (proxy), basement (proxy), outlet (proxy) |
| pond (proxy), basement (proxy) | pond (proxy), basement (proxy), outlet (sensor) |
| pond (proxy), basement (sensor) | pond (proxy), basement (sensor), outlet (proxy) |
| pond (sensor), basement (proxy) | pond (proxy), basement (sensor), outlet (sensor) |
| pond (sensor), basement (sensor) | pond (sensor), basement (proxy), outlet (proxy) |
| pond (proxy), outlet (proxy) | pond (sensor), basement (proxy), outlet (sensor) |
| pond (proxy), outlet (sensor) | pond (sensor), basement (sensor), outlet (proxy) |
| pond (sensor), outlet (proxy) | pond (sensor), basement (sensor), outlet (sensor) |

### *2.6.2 Calibration with noisy proxy data*

The SWMM model was also calibrated with the synthetic proxy data containing complex correlated noise and Gaussian noise. The objective was to gain insight into how the structure and magnitude of noise in proxy measurements affects the utility of proxy measurements for calibration. The synthetic proxy data allow an isolation of the effects of noise structure and magnitude (e.g., Figs. 3g-j share the same noise structure). First, the situation was considered where only noisy proxy data from the pond are available. Second, the situation was considered where noisy proxy data from the pond were complemented by ideal proxy data from the basement. For each situation and each quality of proxy data, the SWMM model was calibrated 10 times on each of the five flood events. Each calibrated model was then validated with the four other flood events.

# 3 Results

## 3.1 Model calibration with ideal proxy data

To illustrate typical model calibration results, hydrographs for event 5 are plotted for an uncalibrated model (Fig. 5a), a model calibrated with sensor data at all three possible measurement locations (Fig. 5b), and a model calibrated with ideal proxy data at the same locations (Fig. 5c). While the level of agreement is very high for the calibrated models, certain features of the curves do not perfectly fit the measured data. In the pond, for example, simulated flooding ends more abruptly than the measured data. These differences can be due to simplifying assumptions used when setting up the model, or small inconsistencies between where the water level was measured and where it is reported from the model.

Comparing the models calibrated with sensor data (Fig. 5b) vs. with IDEAL-100 data (Fig. 5c), the proxy-based calibration gives priority to matching the end of the flooding in the basement and at the outlet shaft, as expected. It is surprising that although only proxy data are used for the model shown in Fig. 5c, the absolute errors are still relatively small.

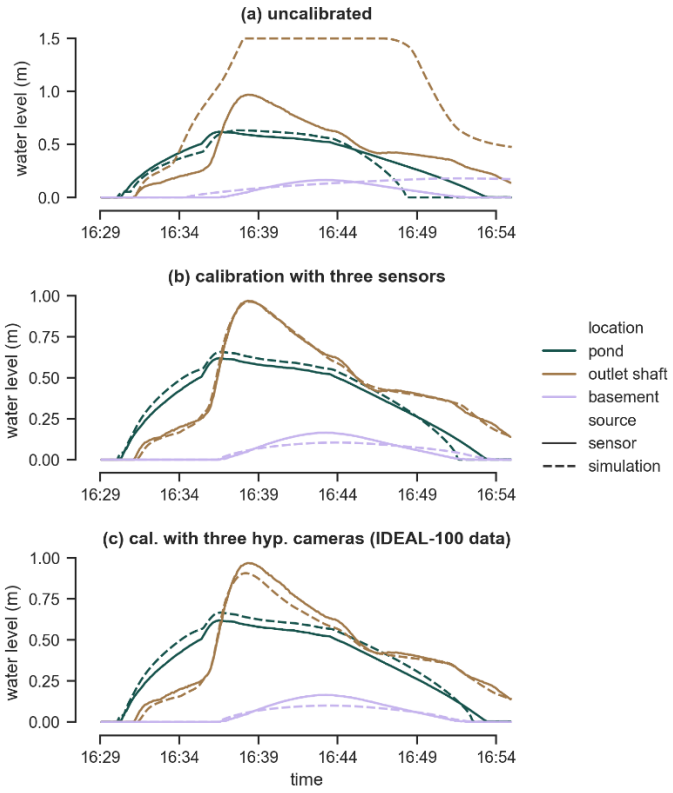


**Figure 5: Water level hydrographs for flood event 5, with sensor data and simulations of an uncalibrated model (a) and of models calibrated with sensor data (b) and with ideal proxy data (IDEAL-100). Note the slight mismatches that remain between the calibrated models and the monitoring data.**

Calibration experiments that failed to converge were filtered out of the dataset. Such failures occurred almost exclusively when calibrating with trend-like data in the basement from event 3. This issue, which appears to be caused by a local minimum in the objective function, is documented in the supporting information (Section S3).

## 3.2 Benchmarks

Figure 6 shows the distribution of absolute modelling error achieved for uncalibrated models (blue) and for models calibrated with all available sensor data (red). The calibrated models have such consistent performance that variability is not visible at the scale of the figure. The diamonds indicate the median model error of each case, which are used for the upper and lower benchmarks described in Section 2.5.6.

Flood event 5 stands out from the other events with higher error rates for both the upper and lower benchmarks. This difference can probably be explained by the fact that flood event 5 involved roughly two times the water volume compared to the other events (see Tab. 1), resulting in higher water levels and therefore higher possible errors.

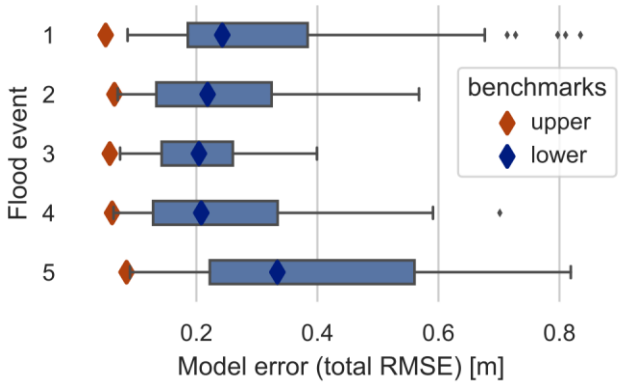


**Figure 6: Absolute error of model runs for uncalibrated models (blue) and models calibrated with all available sensor data (red). The diamonds indicate the median values of model performance, which are used as benchmarks for each flood event. Uncalibrated models have a broad performance distribution represented by boxplots.**

## 3.3 Ideal proxy data would be competition for sensor data

This section presents the results of the first calibration experiment described in Section 2.6, which explores the calibration value of ideal proxy data (IDEAL-100), which represent a hypothetical situation where surveillance footage analysis has been perfected and no moving obstructions affect the measurement. In Fig. 7, the 26 data source combinations have been summarized by the number of proxy and sensor data sources used. The model performance distributions are depicted, with boxplots showing the median, 25$^{th}$, and 75$^{th}$ percentiles. The "whiskers" show the full extent of each distribution, but extend at most 1.5 times the inter-quartile range (the distance between the 25$^{th}$ and 75$^{th}$ percentiles). Any data beyond that limit are considered outliers and plotted as individual points.

The situation where neither sensor nor proxy data are used (left-most boxplot in Fig. 7) groups together all uncalibrated model runs, for which median model performance (MP) is zero by definition of the lower benchmark in Section 2.5.6. As the number of data sources used for calibration increases, performance tends to increase as

well. As might be expected, models calibrated with sensor data perform better than models calibrated with the same number of proxy data sources. Even models calibrated with just two sensors (MP=97%) tend to perform better than models calibrated with three proxy data sources (MP=90.2%).

Models calibrated with just one data source have high variability in performance, both for sensor and proxy data. Different factors contribute to the variability in performance. One factor is that model performance depends not only on the type (proxy or sensor) of the data used for calibration but also on their location (pond, basement, or outlet shaft). For example, upstream measurement locations are less informative than downstream measurement locations. In certain cases, some of the parameters remain poorly defined after calibration. Another factor is linked to the discrepancies between model and measurement, which are visible in the calibrated models plotted in Figs. 5b and 5c. These discrepancies represent a risk of overfitting if only certain locations are available for calibration, and overfitting leads to higher errors during validation (which is always performed with all measurement locations). An example and analysis of overfitting is provided in the supporting information (Section S4).

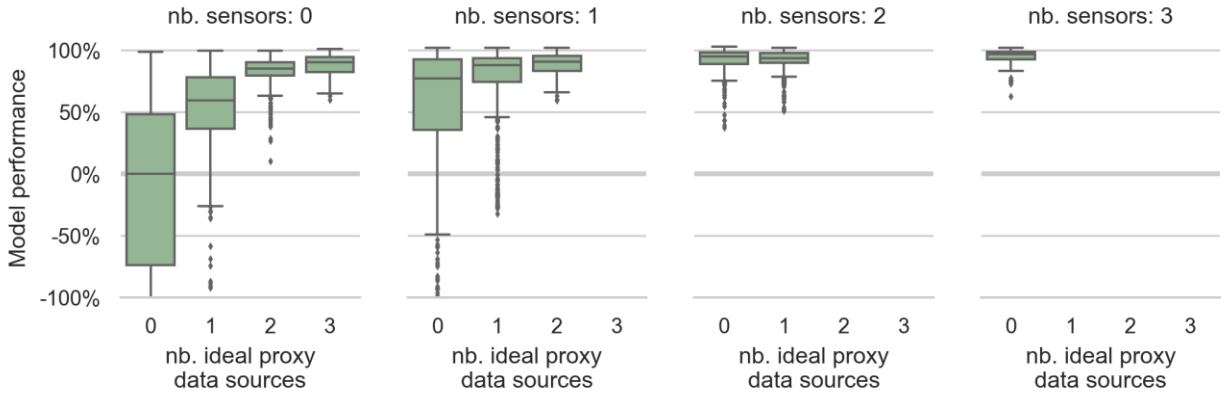


**Figure 7: Model performance (validation) for different combinations of data sources, grouped by the number of sensor or ideal proxy data sources used for calibration. When three sensors are used (far right), there are no locations left for proxy data to be collected, hence the single boxplot.**

Diving deeper into the results, Fig. 8 presents the marginal performance increase (MPI) when adding a data source (either a sensor or an ideal proxy measurement) at a given location, for various initial data source combinations. Initial data source combinations are defined by the number of sensors and/or proxy measurements used for calibration, as indicated on the heatmap axes. For example, the +94% of the heat map in the upper right tells us

that, compared to the median uncalibrated model (zero initial sensors and zero initial proxy measurements), adding a sensor to the outlet shaft will result in a median increase of 94 percentage points (median value). Since there are multiple variations for each situation (for instance different flood events to calibrate on), the MPI is computed as the median increase of MP across all possible combinations.

As might be expected, the highest MPI is achieved when adding a sensor or proxy measurement to an initially ungauged system (lower left of each heatmap). As the number of initial measurement locations rises and the system is saturated with data, the MPI decreases. Thus, MPI is under 5% for all situations where two sensors are already present in the system (uppermost cell in each heatmap).

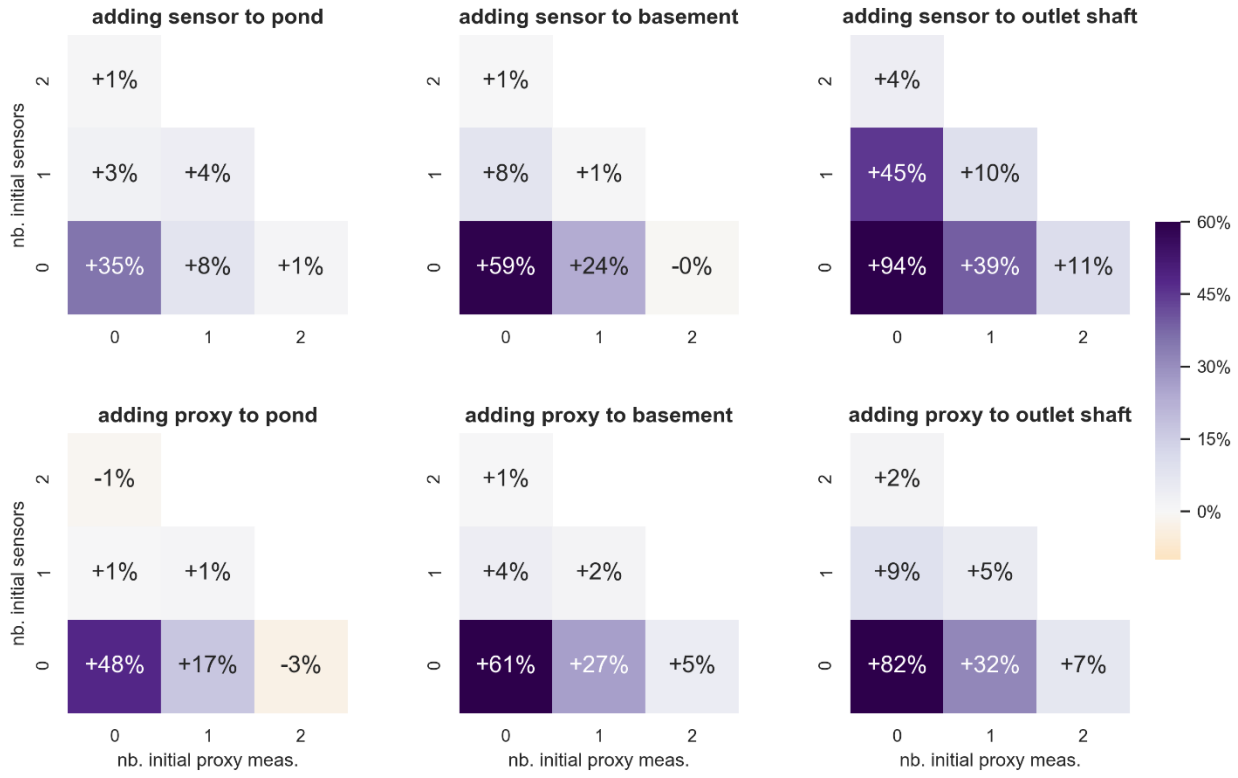


**Figure 8: Marginal performance increase (MPI) when adding either a sensor (top) or an ideal proxy measurement (bottom) for calibration, relative to an initial calibration situation defined by number of sensors and proxy measurements already present in the system. The values represent the median increase of median performance (MP) across all possible variations of a situation (e.g., different events used for calibration)**

There are systematic differences between the three measurement locations. Adding a data source to the pond, which is furthest upstream in the catchment, has the lowest impact whereas a measurement in the outlet shaft, which is furthest downstream, increases the performance significantly more.

Although sensor data delivers a higher MPI than proxy data at the outlet shaft, proxy data generally provides a high improvement at the pond and in the basement. Again, the surprising advantage of proxy data is linked to overfitting, as described in the supporting information (Section S4).

On average over all situations, the median MPI from using additional proxy data is 5.8%, whereas the median MPI from using additional sensor data is 10.1%. By comparing proxy and sensor data on a case-by-case basis, with each case representing an initial configuration of sensors and data and given calibration and validation events, the MPI of proxy data is just over 70% that of sensor data from the same location (median value).

## 3.4 Costs and benefits of noise

In this section, the impact of noise on the calibration value of proxy data is assessed. Both complex correlated noise and simple Gaussian noise are assessed with the help of the synthetic proxy data presented in Section 2.3. This analysis has an illustrative purpose and limits itself to two situations that have measurements at the pond and/or in the basement. These locations were selected because measurements at the outlet shaft are too redundant with the other two locations.

### *3.4.1 Situation 1: Calibration with proxy data from pond only*

In the first situation, a single measurement point at the pond is available. At the pond, the following possible data are considered: none (uncalibrated), sensor data, ideal proxy data, proxy data with Gaussian noise (Fig. 9a), and proxy data with complex correlated noise (Fig. 9b). The situation with no data (Data series at pond = “none”) corresponds to the uncalibrated model simulations from which the lower benchmark was computed.

Compared to the uncalibrated models, models calibrated with ideal proxy data (IDEAL-100) or with sensor data have higher performance and lower variability. Nevertheless, variability is still high, with a significant portion of models performing worse than the median uncalibrated model. This variability has to do with parameter

uncertainty and overfitting (as discussed in relation to Fig. 7). Overfitting is also the reason why models calibrated with IDEAL-100 have a higher MP (48%) than those calibrated with sensor data (35%), as explained in the supporting information (Section S4).

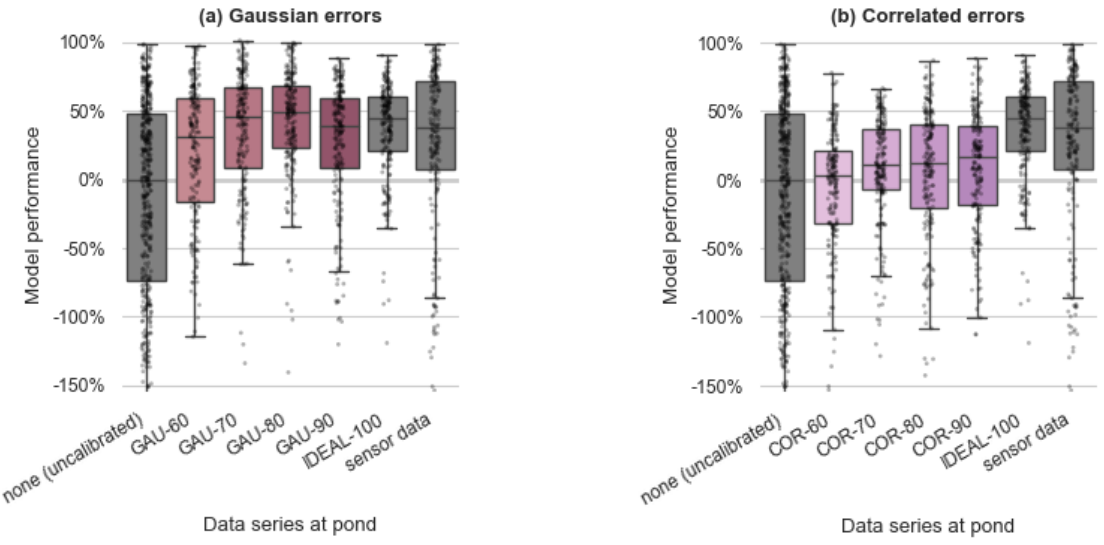


**Figure 9: Performance of models calibrated with data from pond of different qualities. In (a), the proxy data have Gaussian distributed noise, whereas in (b) the proxy data have correlated noise with the same structure as the proxy measurements obtained by automatic analysis of surveillance camera footage.**

The presence of Gaussian noise does not seem to have a strong negative impact on the MP, which hovers below 50% for all quality levels above GAU-60 (Fig. 10a). For GAU-60, overall performance is lower, closer to that of the uncalibrated models. Surprisingly, it appears that overall performance is highest for GAU-80, which could be linked to a reduction of the overfitting mentioned in the previous paragraph.

Correlated errors (Fig. 9b) have a visible impact on model performance, with the MP for COR-60 near 0%. Nevertheless, the overall performance is improved by a reduction in variance as compared to the uncalibrated models. As the quality of the proxy data increases, the calibration performance increases as well, although the change is slight and even COR-90 has a distinctly lower performance than IDEAL-100.

Further differences appear when distinguishing performance based on the event used for calibration. A discussion of the differences is provided in the supporting information (Section S5).

### 3.4.2 Situation 2: Calibration with proxy data from pond and basement

The second situation is equivalent to the first except that the noisy proxy data from the pond are supplemented with ideal proxy data (IDEAL-100) from the basement (Fig. 10). The base case (Data series at pond = none) now corresponds to models calibrated only with ideal proxy data from the basement, hence the improved performance relative to Fig. 9. Considering that the performance with sensor data at the pond is now clearly better than the performance with ideal proxy data at the pond, it now appears that the problem with overfitting, which occurred in absence of other data sources (Fig. 9), is resolved.

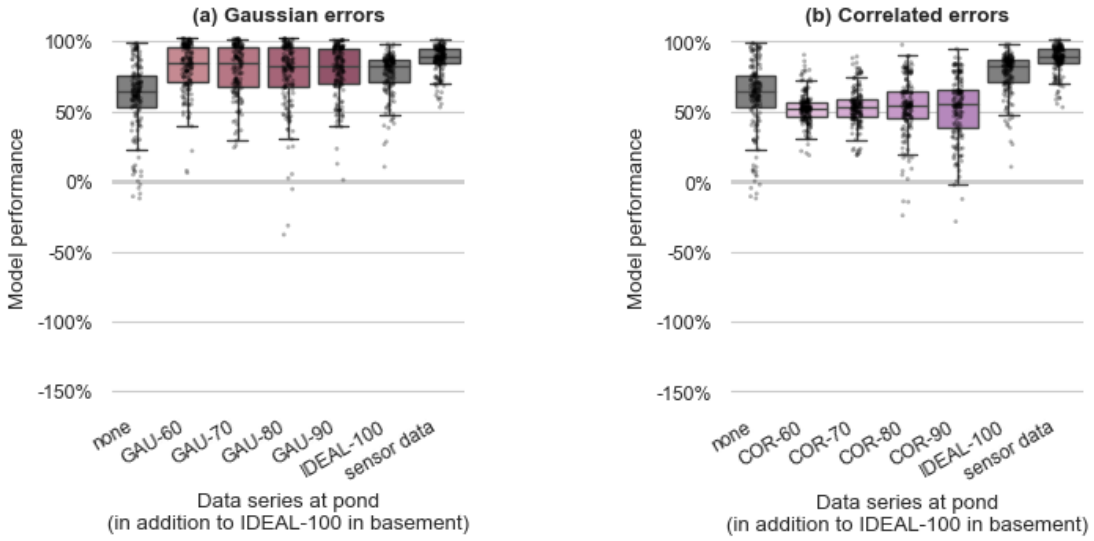


**Figure 10: Performance of models calibrated with ideal proxy data from basement and data from pond of different qualities. In (a), proxy data from pond have Gaussian distributed noise, whereas in (b) the proxy data have correlated noise with the same structure as the proxy measurements obtained by automatic analysis of surveillance camera footage.**

Again, the proxy data with Gaussian noise are not systematically worse than the proxy data with no noise (Fig. 10a), although the variability is higher. For all cases, the MP is higher than if only calibrating with ideal trend-like data from the basement (the "none" base case). Even GAU-60, for which the noise previously appeared to negatively affect the MP, performs no worse than the other proxy data series with Gaussian noise.

In contrast, calibrating with trend-like data with complex correlated noise decreases the MP as compared to the base case (Fig. 10b), even when the magnitude of noise is small (COR-90). Performance does seem to correlate

457 positively with data quality, but the relationship is weak. The poorly performing models and increased variability
458 for COR-80 and COR-90 are again linked to overfitting that occurs specifically when calibrating with events 2 and 4,
459 which are the only events to lack flooding in the basement.

460 Further differences appear when distinguishing performance based on the event used for calibration. A discussion
461 of the differences is provided in the supporting information (Section S5).

# 4 Discussion

## 4.1 Significance of this work

### *4.1.1 Importance and novelty*

No conventional sensor solution exists for measuring small-scale surface flooding in urban areas, even though such measurements are needed to ensure the reliability of models and thus the effectiveness of risk mitigation measures. Thanks to a unique case study, this work sheds light on a new opportunity for urban pluvial flood model calibration with alternative data sources, namely with proxy data automatically obtained from surveillance footage. The aspect of automatic image analysis is particularly important for privacy reasons, which are one of the main risks associated to smart urban water solutions (Moy de Vitry et al., 2019b). While many studies have looked into the collection of data from alternative sources including social media and drones, none has yet investigated the actual benefits that such data provides. This work tackles that question specifically, while at the same time differentiating between sensor placement and different data qualities.

### *4.1.2 Why overfitting was tolerated*

Discrepancies between model and data, which sometimes led to overfitting, were observed in this study. These discrepancies were tolerated even though they complicated the analysis and could have been resolved, e.g., by optimizing model structure. The presence of such discrepancies can be attributed to the realism of the case study, which gives significant practical value to the findings. In particular, this rather unique approach reduces the risk of overestimating the value of trend-like data, as might occur when using simulated data (Viero, 2018).

## 4.2 Could cameras replace sensors?

### *4.2.1 Error-free proxy data are almost as useful as sensor data*

In this study, ideal error-free proxy data were used to calibrate a simple flash flood model and were found to improve validation performance of the model in a consistent manner. This finding suggests that proxy measurements have the potential to satisfy the lack of monitoring data for urban pluvial flood model calibration.

Other studies have come to comparable conclusions for other alternative data sources, for example regarding the use of citizen science to calibrate hydrologic river models (van Meerveld et al., 2017) or the use of binary sensors to calibrate an urban drainage model (Wani et al., 2017).

This study also found that the difference in value between sensor data and error-free proxy data is smaller than might be expected: in more than half of all the calibration situations analyzed, proxy data provided at least 70% of the benefit of sensor data. To take a more specific example, calibrating with proxy data at all available locations provided a median performance that was only 4.5% poorer than when calibrating with sensor data at the same locations.

In situations where overfitting is a risk, such as when calibrating with only one upstream measurement in a catchment, the results point to the surprising conclusion that trend data provides more robust models than sensor data. This was explained by the fact that proxy-based calibration is less opinionated as compared to calibration with sensor data and objective functions that favor perfect fitting of the model to measurement values.

### *4.2.2 Measurement noise is usually detrimental*

The proxy data that were automatically extracted from surveillance footage of the case study contained noise, and although there are ways to reduce it, it is unlikely that such noise can be completely resolved. The errors are due to classification problems and changes in the flooding scene (Moy de Vitry et al., 2019a), and they have a complex autocorrelation structure.

In the present study, the effect of complex correlated noise was usually detrimental. In the worst cases, even the proxy data with the lowest level of noise (COR-90) would give significantly lower performance than the ideal proxy data (IDEAL-100). Gaussian noise (random noise with low autocorrelation, in contrast to complex correlated noise) had a vastly different effect on performance. When calibrating with proxy data from the pond alone, then only the highest noise level (GAU-60) provoked a clear reduction in performance compared to the ideal proxy data. In the other cases, performance was similar or even higher than the ideal proxy data. Compared to previous research that suggests that random noise in calibration data do not have a strong negative effect during calibration (Dotto

et al., 2014), we go a step further and suggest that some noise could even be beneficial in situations where model overfitting is a risk.

It follows that until complex correlated noise is kept below an acceptable threshold (which remains to be defined), proxy data obtained from surveillance footage should be used for calibration only with extreme care. Possible approaches for reducing errors have been outlined by Moy de Vitry et al. (2019a). Until then, proxy-based calibration should be limited to the assimilation of proxy data with uncorrelated errors, such as proxy measurements from uncalibrated conventional sensors (data from uncalibrated sensors do have systematic errors, but these do not affect the trend information contained in the data).

## 4.3 Outlook and future research

### *4.3.1 Suitability of the Spearman correlation for calibrating proxy data*

The hybrid Spearman correlation used in this study was effective in fitting the SWMM model to the shape of the proxy data, even when a large amount of noise was present. The use of the IoU in cases where the proxy data were flat (e.g., when no flooding occurred in the basement) proved to be effective for calibration, although future research could investigate whether there are better alternatives to the proposed solution.

There are situations in which calibration with the Spearman correlation is difficult. For flood event 3, which involves a very short and shallow flood in the basement, calibrations were sometimes unsuccessful because they would stop at a local minimum. It appears that such issues could be avoided by making the calibration stopping criteria more strict.

A strong downside of the Spearman correlation for describing data quality is that it is not very telling of the noise structure. As seen in the results, two proxy measurements with the same Spearman correlation can give very different results during calibration. Better characterization of the noise structure would possibly also allow for correlated errors in the measurement data to be accounted for.

### 4.3.2 Further experiments are necessary

The floodX data sets are of unique value thanks to their provision of multiple flood events on which flood models can be calibrated and validated. Nevertheless, they are limited due to the small size of the facility. In this study, the small catchment size meant that with three sensors, redundancy was higher than what is expected in a real catchment with perhaps a thousand times more components. This led to interdependence between measurement locations, meaning that it was not completely possible to isolate the effect of measurement location from the effect of other measurements. Thus, larger case studies should be considered in the future.

Measurement location was found to have a strong influence on the value of the monitoring data, which underlines the importance of research for identification of optimal sensor placement for monitoring campaigns (Vonach et al., 2018). Compared to reality, the flooding events studied in this paper lacked rainfall-runoff processes. In cases where these processes need to be modelled, they can introduce parameter uncertainty that might shift the need for monitoring data upstream. Therefore, it is also important that future case studies include such rainfall-runoff processes. It is expected that these processes increase modelling complexity and uncertainty, thereby increasing the need for data (especially surface flooding data) and the potential benefit of proxy measurements.

As this study focused only on proxy measurements for water level from cameras using the SOFI method, future studies should consider combining SOFI proxy measurements with other data sources. Examples include flow velocity data from surveillance cameras (Leitão et al., 2018), citizens (Le Boursicaud et al., 2016), and drones (Perks et al., 2016), or water level information obtained from social media (Chaudhary et al., 2019).

### 4.3.3 Implications for model calibration with alternative data sources

In this study, the identifiability of parameters was found to depend on measurement location, data type, and data quality. For example, proxy data from the pond was found to provide only limited information on parameters downstream. In certain cases, this led to the selection of extreme parameter values. In future studies, calibration parameters should be defined according to a sensitivity analysis based on the available measurement data. Alternatively, parameters that remain undefined after calibration should be assigned a value based on prior

knowledge. While a Bayesian approach may be applicable in certain cases, the definition of appropriate likelihood functions may be difficult, as in the case of proxy data.

The presence of complex correlated noise was found to be a critical factor influencing model calibration in this study. Since such noise might be expected in practice, a method to identify and screen erroneous data before calibration is needed. Accurate characterization of noise structure is also necessary for effective Bayesian parameter estimation (Wani et al., 2019). The Spearman correlation seems to be an insufficient indicator of quality in this regard. A similar call was issued concerning erroneous crowd-sourced data in hydrological modelling (Mazzoleni et al., 2017). Once the reliability of a data source has been assessed, it can be used or rejected based on a threshold. Alternatively, when calibrating with multiple sources of data, the reliability of each source can be used to weigh its objective function, so that information that is more reliable has precedence.

# 5 Conclusions

The lack of overland flood monitoring data is a recognized issue in urban pluvial flood modelling, but in most situations, conventional sensor solutions are very impractical if not impossible. Alternative data sources like social media and surveillance cameras appeal as a novel and cost-effective approach to the problem. This study assessed the potential of proxy measurements such as could be obtained from surveillance footage by automatically tracking the evolution of the visible flood extent with computer vision. Data from five flood events were used to calibrate and validate a one-dimensional dual-drainage EPA SWMM model with different combinations of proxy and sensor data. Overall, the results showed that ideal proxy data with no errors or with Gaussian noise are almost as good as sensor data. However, the complex correlated errors currently featured in camera-sourced proxy measurements are detrimental and can completely cancel any benefits of the proxy data. The use of multiple measurement locations for calibration is also important for model performance. The main contributions of this study are:

- An objective function, based on the Spearman rank correlation coefficient, was proposed and demonstrated for calibrating with water level proxy data. It worked in most cases but experienced difficulty with one very short and shallow flooding event.
- The results confirmed that proxy data have a strong potential to improve a urban pluvial flood model's predictive performance. For example, models calibrated with three ideal proxy data series had a median performance of 90% that of a perfectly calibrated model. However, complex correlated noise reduces the utility of proxy data and can even be counter-productive. Gaussian noise was generally unproblematic and was sometimes beneficial in making the model more robust.
- Inevitable discrepancies in the model and data, for example due to modelling assumptions, can be harmless if a sufficient number of measurements are available. Too few measurements can lead to overfitting, in which case performance was sometimes improved when data contained Gaussian noise.

- The Spearman correlation is not a sufficient indicator of the presence of complex correlated noise as opposed to random Gaussian noise, which is a critical factor for model calibration. In practice, alternative data sources such as surveillance cameras are not expected to provide perfect data, so the issue of data quality assessment and screening should be investigated further.

## Code and data availability

The code used in this work for creating, training, and evaluating the DCNN, as well as extracting the SOFI and plotting results can be found in the following repositories:

- https://github.com/mmmatthew/cq-analysis contains code to prepare and run the calibration experiments
- https://github.com/mmmatthew/swmm_calibration contains a framework to calibrate a SWMM model with sensor and/or proxy data

The results generated from the calibration experiments and the analysis of the results can be found in the following archive:

Moy de Vitry, Matthew. (2019). Trend-based calibration experiments with floodX data (Version 1.0) [Data set]. http://doi.org/10.25678/0001B6

## Acknowledgements

The authors acknowledge A. Scheidegger, O. Wani, and K. Villez for the useful discussions of the method and for the interpretation of results. The authors thank M. Y. Schneider and R. Dawson for reviewing the manuscript. This project was financed by the Swiss National Science Foundation under grant #169630. The funding body had no role in study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

## References


Assumpção, T.H., Popescu, I., Jonoski, A., Solomatine, D.P., 2018. Citizen observations contributing to flood modelling: Opportunities and challenges. Hydrol. Earth Syst. Sci. 22, 1473–1489. https://doi.org/10.5194/hess-22-1473-2018

Bhola, P.K., Nair, B.B., Leandro, J., Rao, S.N., Disse, M., 2019. Flood inundation forecasts using validation data generated with the assistance of computer vision. J. Hydroinformatics 21, 240–256. https://doi.org/10.2166/hydro.2018.044

Chaudhary, P., Aronco, S.D., Moy de Vitry, M., Leitao, J.P., Wegner, J.D., 2019. Flood-Water Level Estimation from Social Media Images, in: ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. pp. 5–12. https://doi.org/10.5194/isprs-annals-IV-2-W5-5-2019

Dotto, C.B.S., Kleidorfer, M., Deletic, A., Rauch, W., McCarthy, D.T., 2014. Impacts of measured data uncertainty on urban stormwater models. J. Hydrol. 508, 28–42. https://doi.org/10.1016/j.jhydrol.2013.10.025

Duan, Q.Y., Gupta, V.K., Sorooshian, S., Duan, Y., 1993. Shuffled complex evolution approach for effective and efficient global minimization. J. Optim. Theory Appl. 76, 501–521. https://doi.org/10.1007/BF00939380

Emerson, S.C., Owen, A.B., 2009. Calibration of the empirical likelihood method for a vector mean. Electron. J. Stat. 3, 1161–1192. https://doi.org/10.1214/09-EJS518

Field, C.B., Barros, V., Stocker, T.F., Dahe, Q., Jon Dokken, D., Ebi, K.L., Mastrandrea, M.D., Mach, K.J., Plattner, G.K., Allen, S.K., Tignor, M., Midgley, P.M., 2012. Managing the risks of extreme events and disasters to advance climate change adaptation: Special report of the intergovernmental panel on climate change, Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change. https://doi.org/10.1017/CBO9781139177245

Garcia, F., Folton, N., Oudin, L., 2017. Which objective function to calibrate rainfall–runoff models for low-flow index simulations? Hydrol. Sci. J. 62, 1149–1166. https://doi.org/10.1080/02626667.2017.1308511

Haupt, S.E., Pasini, A., Marzban, C., 2009. Artificial Intelligence Methods in the Environmental Sciences. Springer Netherlands, Dordrecht. https://doi.org/10.1007/978-1-4020-9119-3

Hénonin, J., Hongtao, M., Zheng-Yu, Y., Hartnack, J., Havnø, K., Gourbesville, P., Mark, O., 2015. Citywide multi-grid urban flood modelling: the July 2012 flood in Beijing. Urban Water J. 12, 52–66. https://doi.org/10.1080/1573062X.2013.851710

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

Jiang, J., Liu, J., Cheng, C., Huang, J., Xue, A., 2019. Automatic Estimation of Urban Waterlogging Depths from Video Images Based on Ubiquitous Reference Objects. Remote Sens. 11, 587. https://doi.org/10.3390/rs11050587

Jiang, Y., Zevenbergen, C., Ma, Y., 2018. Urban pluvial flooding and stormwater management: A contemporary review of China's challenges and "sponge cities" strategy. Environ. Sci. Policy 80, 132–143. https://doi.org/10.1016/j.envsci.2017.11.016

Le Boursicaud, R., Pénard, L., Hauet, A., Thollet, F., Le Coz, J., 2016. Gauging extreme floods on YouTube: application of LSPIV to home movies for the post-event determination of stream discharges. Hydrol. Process. 30, 90–105. https://doi.org/10.1002/hyp.10532

Leitão, J.P., Peña-Haro, S., Lüthi, B., Scheidegger, A., Moy de Vitry, M., 2018. Urban overland runoff velocity measurement with consumer-grade surveillance cameras and surface structure image velocimetry. J. Hydrol. 565, 791–804. https://doi.org/10.1016/j.jhydrol.2018.09.001

Lv, Y., Gao, W., Yang, C., Wang, N., 2018. Inundated Areas Extraction Based on Raindrop Photometric Model ({RPM}) in Surveillance Video. Water 10, 1332. https://doi.org/10.3390/w10101332

Madsen, H., 2003. Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. Adv. Water Resour. 26, 205–216. https://doi.org/10.1016/S0309-

1708(02)00092-1

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? Hydrol. Earth Syst. Sci 21, 839–861. https://doi.org/10.5194/hess-21-839-2017

McKinney, W., 2010. Data Structures for Statistical Computing in Python, in: Proceedings of the 9th Python in Science Conference. Austin, TX, USA.

Moy de Vitry, M., Dicht, S., Leitão, J.P., 2017. floodX: urban flash flood experiments monitored with conventional and alternative sensors. Earth Syst. Sci. Data 9, 657–666. https://doi.org/10.5194/essd-9-657-2017

Moy de Vitry, M., Kramer, S., Wegner, J.D., Leitão, J.P., 2019a. Scalable flood level trend monitoring with surveillance cameras using a deep convolutional neural network. Hydrol. Earth Syst. Sci. 23, 4621–4634. https://doi.org/10.5194/hess-23-4621-2019

Moy de Vitry, M., Schneider, M.Y., Wani, O., Manny, L., Leitão, J.P., Eggimann, S., 2019b. Smart urban water systems: what could possibly go wrong? Environ. Res. Lett. https://doi.org/10.1088/1748-9326/ab3761

Ochoa-Rodriguez, S., Onof, C., Maksimovic, C., Wang, L., Willems, P., Assel, J., Gires, A., Ichiba, A., Bruni, G., Veldhuis, T., 2015. Urban pluvial flood modelling: current theory and practice. Review document related to Work Package 3 – Action 13 2013, 1–45.

Perks, M.T., Russell, A.J., Large, A.R.G., 2016. Technical Note: Advances in flash flood monitoring using unmanned aerial vehicles (UAVs). Hydrol. Earth Syst. Sci. 20, 4005–4015. https://doi.org/10.5194/hess-20-4005-2016

Rossman, L.A., 2010. Storm water management model user's manual, version 5.0. National Risk Management Research Laboratory, Office of Research and Development, US Environmental Protection Agency.

Seibert, J., Vis, M.J.P., Lewis, E., van Meerveld, H.J., 2018. Upper and lower benchmarks in hydrological modelling. Hydrol. Process. 32, 1120–1125. https://doi.org/10.1002/hyp.11476

677 Skougaard Kaspersen, P., Høegh Ravn, N., Arnbjerg-Nielsen, K., Madsen, H., Drews, M., 2017. Comparison of the
678 impacts of urban development and climate change on exposing European cities to pluvial flooding. Hydrol.
679 Earth Syst. Sci. 21, 4131–4147. https://doi.org/10.5194/hess-21-4131-2017

680 Spearman, C., 1904. The Proof and Measurement of Association between Two Things. Am. J. Psychol. 15, 72.
681 https://doi.org/10.2307/1412159

682 ten Veldhuis, J.A.E., 2011. How the choice of flood damage metrics influences urban flood risk assessment. J. Flood
683 Risk Manag. 4, 281–287. https://doi.org/10.1111/j.1753-318X.2011.01112.x

684 Thorndahl, S., Beven, K.J., Jensen, J.B., Schaarup-Jensen, K., 2008. Event based uncertainty assessment in urban
685 drainage modelling, applying the GLUE methodology. J. Hydrol. 357, 421–437.
686 https://doi.org/10.1016/j.jhydrol.2008.05.027

687 Tscheikner-Gratl, F., Zeisl, P., Kinzel, C., Leimgruber, J., Ertl, T., Rauch, W., Kleidorfer, M., 2016. Lost in calibration:
688 why people still do not calibrate their models, and why they still should – a case study from urban drainage
689 modelling. Water Sci. Technol. 74, 2337–2348. https://doi.org/10.2166/wst.2016.395

690 van Meerveld, H.J.I., Vis, M.J.P., Seibert, J., 2017. Information content of stream level class data for hydrological
691 model calibration. Hydrol. Earth Syst. Sci. 21, 4895–4905. https://doi.org/10.5194/hess-21-4895-2017

692 Viero, D.P., 2018. Comment on “can assimilation of crowdsourced data in hydrological modelling improve flood
693 prediction?” by Mazzoleni et al. (2017). Hydrol. Earth Syst. Sci. 22, 171–177. https://doi.org/10.5194/hess-
694 22-171-2018

695 Vis, M., Knight, R., Pool, S., Wolfe, W., Seibert, J., 2015. Model Calibration Criteria for Estimating Ecological Flow
696 Characteristics. Water 7, 2358–2381. https://doi.org/10.3390/w7052358

697 Vonach, T., Tscheikner-Gratl, F., Rauch, W., Kleidorfer, M., 2018. A Heuristic Method for Measurement Site
698 Selection in Sewer Systems. Water 10, 122. https://doi.org/10.3390/w10020122

699 Wang, Y., Chen, A.S., Fu, G., Djordjević, S., Zhang, C., Savić, D.A., 2018. An integrated framework for high-resolution

urban flood modelling considering multiple information sources and urban features. Environ. Model. Softw. 107, 85–95. https://doi.org/10.1016/j.envsoft.2018.06.010

Wani, O., Scheidegger, A., Carbajal, J.P., Rieckermann, J., Blumensaat, F., 2017. Parameter estimation of hydrologic models using a likelihood function for censored and binary observations. Water Res. 121, 290–301. https://doi.org/10.1016/j.watres.2017.05.038

Wani, O., Scheidegger, A., Cecinati, F., Espadas, G., Rieckermann, J., 2019. Exploring a copula-based alternative to additive error models—for non-negative and autocorrelated time series in hydrology. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2019.06.006

Zoppou, C., 2001. Review of urban storm water models. Environ. Model. Softw. https://doi.org/10.1016/S1364-8152(00)00084-0