Exploring the Performance of Ensemble Smoothers to Calibrate Urban Drainage Models

Huang, Yuan; Zhang, Jiangjiang ; Zheng, Feifei; Jia, Yueyi; Kapelan, Zoran; Savić, Dragan

DOI
10.1029/2022WR032440

Publication date
2022

Document Version
Final published version

Published in
Water Resources Research

Citation (APA)
Huang, Y., Zhang, J., Zheng, F., Jia, Y., Kapelan, Z., & Savić, D. (2022). Exploring the Performance of Ensemble Smoothers to Calibrate Urban Drainage Models. *Water Resources Research*, *58*(10), Article e2022WR032440. https://doi.org/10.1029/2022WR032440

AGU ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2022WR032440

Exploring the Performance of Ensemble Smoothers to Calibrate Urban Drainage Models

Yuan Huang[1], Jiangjiang Zhang[2], Feifei Zheng[3], Yueyi Jia[3], Zoran Kapelan[4], and Dragan Savic[5,6,7]

[1]College of Water Conservancy & Hydropower Engineering, Hohai University, Nanjing, China, [2]Yangtze Institute for Conservation and Development, Hohai University, Nanjing, China, [3]College of Civil Engineering and Architecture, Zhejiang University, Hangzhou, China, [4]Department of Water Management, Delft University of Technology, Delft, The Netherlands, [5]KWR Water Research Institute, Nieuwegein, The Netherlands, [6]Centre for Water Systems, University of Exeter, Exeter, UK, [7]Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, Bangi, Malaysia

Key Points:

- Ensemble smoothers are promisingly effective and robust methods to calibrate urban drainage models
- Ubiquitous parameter equifinality hinders unique parameter identification, which is a concern for practical applications
- Ensemble smoothers are validated on a real-world case, providing insights into how improved model performance can be achieved

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

F. Zheng,
feifeizheng@zju.edu.cn

Citation:

Huang, Y., Zhang, J., Zheng, F., Jia, Y., Kapelan, Z., & Savic, D. (2022). Exploring the performance of ensemble smoothers to calibrate urban drainage models. *Water Resources Research*, *58*, e2022WR032440. https://doi.org/10.1029/2022WR032440

Received 26 MAR 2022
Accepted 9 OCT 2022

Author Contributions:

Conceptualization: Yuan Huang, Jiangjiang Zhang, Feifei Zheng
Data curation: Yueyi Jia
Formal analysis: Yuan Huang, Feifei Zheng, Zoran Kapelan, Dragan Savic
Investigation: Yuan Huang
Methodology: Yuan Huang, Jiangjiang Zhang, Feifei Zheng, Dragan Savic
Resources: Jiangjiang Zhang, Zoran Kapelan, Dragan Savic
Software: Yuan Huang
Supervision: Feifei Zheng, Zoran Kapelan
Validation: Yuan Huang, Jiangjiang Zhang, Feifei Zheng, Yueyi Jia, Zoran Kapelan, Dragan Savic
Writing – original draft: Yuan Huang
Writing – review & editing: Yuan Huang, Jiangjiang Zhang, Feifei Zheng, Yueyi Jia, Zoran Kapelan, Dragan Savic

Abstract Urban drainage models (UDMs) are often used to manage urban flooding. However, these models generally involve many parameters to represent the underlying complex hydrodynamic processes. This results in significant challenges to achieving effective and robust model calibration especially with frequently limited observations, leading to unreliable model predictions. This paper makes the first attempt at UDM calibration using the Bayesian-based Ensemble Smoother (ES) method. Three ES variants are considered, that is, the primary ES, the versions with multiple data assimilation (ES-MDA) and iterative local update (ES-ILU). Two synthetic cases and one real-world application with up to 5,236 calibration parameters are tested. Results obtained show that: (a) both ES-MDA and ES-ILU can produce effective model calibration with ES-ILU outperforming ES-MDA in terms of both accuracy and uncertainty while ES exhibits limited performance; (b) for the real-world case, both the ES-MDA and ES-ILU methods provide better calibration results than the best-known solution manually obtained, (c) a minimum number of observations are required to enable an overall accurate model calibration (e.g., four and ten more monitoring sites are needed in the two synthetic cases); and (d) the model calibrated using an intense rainfall event is generally robust to make reliable predictions across different rainfall events while the model calibrated using less intense rainfall event does not perform well for more intense rainfall events. It was also found that ubiquitous parameter equifinality significantly hinders unique parameter identification even when overall accurate state estimates are obtained. This should be clearly understood in practical applications.

Plain Language Summary Urban floods have been a serious disaster worldwide. Urban drainage models (UDMs) have been widely used to facilitate flooding prevention and mitigation. However, a challenge associated with the UDMs is that a large number of parameters need to be specified and calibrated. While some optimization-based or manual calibration methods are available, their efficiency or accuracy are often unsatisfactory. This can lead to unreliable predictions of the rainfall-runoff process. To this end, this paper proposes the ensemble smoother (ES) methods to calibrate the UDMs. Benefits of these ES methods include the great efficiency, high accuracy and the identification of the uncertainty when calibrating model parameters. These conclusions are based on results of two synthetic cases and one real-world UDM.

1. Introduction

As a result of rapid urbanization and changing climate, urban catchments are experiencing a population growth in impervious areas and more frequent extreme rainfall events (Annus et al., 2021; Zheng et al., 2015). Consequently, urban floods, caused by storms and overloading of drainage systems in urban areas, have become a serious disaster worldwide, posing significant threats to the economy, urban water environment and public safety (Lin et al., 2020). For example, Henan Province in China was hit by historically rare heavy rainstorms through 17–23 July 2021. It has caused 398 deaths or missings and direct economic losses of 120 billion RMB (Xinhua News Agency, 2022). To mitigate the impact of urban floods, different solutions have been proposed, including (a) gray infrastructure solutions that use conventional drainage facilities, such as pumps, deep tunnels and large pipes, to increase the capacity of urban drainage systems (Berggren et al., 2012); (b) green or nature-based solutions that retrofit the existing drainage system by local semi-engineered structures, such as permeable pavements,

retention ponds, and rainwater gardens, to reduce the surface runoff as well as delay the flood peak (Fiori & Volpi, 2020; Lian et al., 2020).

To optimize flood prevention and mitigation, urban drainage models (UDMs) have been widely used, for example, the Storm Water Management Model (SWMM) (Zheng et al., 2018). UDMs can simulate complex hydrodynamic processes associated with urban floods, including precipitation, evaporation, infiltration, surface interception, overland flow, drainage network flow, water retention, etc. (Niazi et al., 2017). As a result, a large number of parameters usually need to be specified and calibrated, especially when many sub-catchments are involved (Wang et al., 2012). This poses huge challenges for model calibration, especially under limited observations.

Calibration methods for UDMs have been studied extensively (Clemens, 2001; Niazi et al., 2017). Manual or "trail-and-error" calibration is commonly used in practice (Wu et al., 2013). However, its utility relies heavily on the practitioner's experience (Duan & Gao, 2019), and this process is generally labor-intensive and time-consuming (Barco et al., 2008). Conversely, automatic calibration can improve calibration efficiency (Di Pierro et al., 2005), thus has received increasing attention in recent years (Alamdari et al., 2017). Most of the relevant methods developed so far use an optimization algorithm, such as gradient-based, heuristic or combinatorial methods, to search for an optimal parameter set that minimizes the difference between observed and modeling states (Behrouz et al., 2020).

While these automatic calibration methods are generally effective in finding an optimum, they are easily trapped in local minima or computationally inefficient. Those characteristics are particularly limiting the use of models for real-time prediction and control of urban floods (Niazi et al., 2017). Some studies have shown that the performance of optimization-based calibrations is case-dependent, influenced by rainfall event (Swathi et al., 2019), objective function (Reed et al., 2013) and parameter type (Kanso et al., 2003). In addition, literature shows that the reduction of the number of parameters via sensitivity analysis, parameter grouping, principal component analysis, is commonly used to deal with the "curse-of-dimensionality" (Salvadore et al., 2015). However, Niazi et al. (2017) reviewed 34 studies using sensitivity analysis for SWMM calibration and found that the identified sensitive parameters are not always the same and even quite different from case to case. That finding means that one cannot simply determine a generic set of reduced parameters for the calibration of different models. Consequently, there is still an urgent need to develop more robust methods for the UDM calibration.

Besides the above-mentioned approaches, there is also a need to explore the complexity and uncertainty underlying the calibration problem, aimed at improving the effectiveness of model calibration. Numerous studies have recognized the significant "parameter equifinality" problem. This refers to the situation when different parameter sets lead to an equally good fit between model predictions and observations (Her & Chaubey, 2015; Kelleher et al., 2017). Equifinality for UDMs mainly results from the insufficient observed data available and/or the observed data that is not well distributed, spatially and/or temporally, relative to calibration parameters (Vonach et al., 2018). It is anticipated that increasing the number of monitoring sites might mitigate this problem. This has led to the calibration practice moving away from the early studies that used only the observations at the outlet of the system (i.e., end-point observations) to multiple observation sites, thus improving the accuracy of model calibration (Guo et al., 2018; Vonach et al., 2018). However, investigations on how observations affect model calibration are still lacking. In addition, due to the equifinality issue, the calibration for observed states does not necessarily guarantee a good fit of model states other than those at observation sites (Moy de Vitry et al., 2017). Although this issue has been recognized in some previous studies (Moussa et al., 2007), surprisingly few efforts have been made to explicitly address this issue and formulate a rigorous calibration procedure.

This paper aims to address the above issues by introducing ensemble smoother (ES) methods to the UDM calibration. As a class of Bayesian inversion methods for data assimilation, the ES methods perform well in terms of efficiency and accuracy for parameter estimation and uncertainty quantification of nonlinear problems (Hutton et al., 2014; Kapelan et al., 2007; Stuart & Zygalakis, 2015). They have received increased attention in areas of oceanic, geophysical and hydrological sciences (Aanonsen et al., 2009; Xue & Zhang, 2014). In particular, an iterative application of ES with multiple data assimilation (ES-MDA), proposed by Emerick and Reynolds (2013), can be used for strongly nonlinear problems. Recently, Zhang et al. (2018) further integrated a local update strategy into the ES-MDA method to better deal with complex nonlinear models with multimodal distributions of parameters. The method is named the ES with iterative local update (ES-ILU) in this paper.

This paper presents the primary and two enhanced ES methods (i.e., ES, ES-MDA, and ES-ILU) for the UDM calibration. Two synthetic cases with investigations on key apsects of the calibration problem and a real-world case are tested. The novelty/contributions of this paper include: (a) providing an alternative calibration approach based on the ES methods to obtain effective and robust UDM calibration; and (b) building knowledges on how observations and rainfall characteristics affect the model calibration performance as well as the parameter identification issue for the calibration problem. This paper is organized as follows. Section 2 describes the proposed calibration approach using the three ES methods and the key aspects of the problem to be considered for model calibration. Section 3 introduces the three case studies for verification and investigation of the developed calibration method. Corresponding case results are given in Section 4, and conclusions are presented in Section 5.

2. Methodology

The methodology starts with the formulation of the UDM calibration problem based on the Bayesian framework. This is followed by the description of ES, ES-MDA, and ES-ILU. Then key aspects of the problem that affect the effectiveness and robustness of model calibration are considered. Note that the model calibration considered in this paper is event-based, that is, the UDM is forced by a single rainfall event. The event-based rainfall modeling can provide system responses at a finer time scale (hourly or minutely) than the continuous rainfall modeling that uses several rainfall events in a long time span (i.e., days) (Behrouz et al., 2020; Swathi et al., 2019).

2.1. Bayesian Parameter Estimation for Model Calibration

A UDM model predicts a number of outputs defining the system state over time (water levels and conduit/surface flows) for given inputs (e.g., system configuration data, rainfall data, etc.). A number of parameters describing the relevant process are also presented in these models as inputs (e.g., surface slope, Manning's coefficient, etc.). Let $\mathbf{x}$ denote a vector of all UDM inputs; $\mathbf{m}$, a $N_m \times 1$ vector of parameters; and $\mathbf{y}$, a vector of model outputs, the UDM can then be defined as $\mathbf{y} = f(\mathbf{x}; \mathbf{m})$ where $f(\cdot)$ represents the mapping function. To enable effective modeling of urban floods, the parameters $\mathbf{m}$ usually need to be calibrated based on the observed (i.e., measured) output states resulted from rainfall events. Given a $N_d \times 1$ vector of observed output states $\mathbf{d}$ and a $N_d \times 1$ vector of the observation errors $\boldsymbol{\eta}$, the relationship between model parameters and observed outputs can be formed as

$$\mathbf{d} = f(\mathbf{m}) + \boldsymbol{\eta} \tag{1}$$

The above equation represents the fundamental formulation of the UDM calibration problem. From the Bayesian perspective, it can be presented in the following way (Stuart & Zygalakis, 2015):

$$p(\mathbf{m}|\mathbf{d}) \propto p(\mathbf{d}|\mathbf{m})p(\mathbf{m}) \tag{2}$$

where $p(\mathbf{m})$ and $p(\mathbf{m}|\mathbf{d})$ are the prior and posterior distributions of model parameters, respectively; $p(\mathbf{d}|\mathbf{m})$ is the likelihood function. Equation 2 indicates that the prior and the likelihood can be combined to determine the posterior distribution of parameters. If we assume that model parameter distributions and observation errors are both Gaussian and $f(\cdot)$ is linear, Equation 2 can be formulated as a minimization problem (Stuart & Zygalakis, 2015):

$$\arg\min_{\mathbf{m}} \left\{ \frac{1}{2}\left(\mathbf{m} - \overline{\mathbf{m}}\right)^T \mathbf{C}_\mathbf{M}^{-1} \left(\mathbf{m} - \overline{\mathbf{m}}\right) + \frac{1}{2}[\mathbf{d} - f(\boldsymbol{m})]^T \mathbf{C}_\mathbf{D}^{-1}[\mathbf{d} - f(\boldsymbol{m})] \right\} \tag{3}$$

where $\overline{\mathbf{m}}$ is the mean of model parameters, $\mathbf{C}_\mathbf{M}$ is the $N_m \times N_m$ autocovariance matrix of model parameters, $\mathbf{C}_\mathbf{D}$ is the $N_d \times N_d$ covariance matrix of observation errors.

Although the underlying minimization principle in Equation 3 is derived based on the linear Gaussian assumption, it can be naturally generalized to non-Gaussian and nonlinear problems, such as the calibration problem considered in this paper. This can be done by approximating the Gaussian distributions by the Monte Carlo method (Stuart & Zygalakis, 2015). A family of methods has been developed following this approximation strategy, such as ensemble smoother described below (Aanonsen et al., 2009; Xue & Zhang, 2014).

2.2. Ensemble Smoother Methods

2.2.1. The Primary ES Method

The primary ES method employs an ensemble of realizations to obtain Monte Carlo approximations of the mean and covariance of the parameter vectors and computes a global update for model parameters with all observed data considered (Stuart & Zygalakis, 2015). More specifically, ES updates model parameters from the prior ensemble $\mathbf{M}^f = \left[\mathbf{m}_1^f, \cdots, \mathbf{m}_{N_e}^f\right]$ to the posterior ensemble $\mathbf{M}^a = \left[\mathbf{m}_1^a, \cdots, \mathbf{m}_{N_e}^a\right]$ (N_e is the number of realizations) in the following way:

$$\mathbf{m}_j^a = \mathbf{m}_j^f + \mathbf{C}_{\mathbf{MD}}^f\left(\mathbf{C}_{\mathbf{DD}}^f + \mathbf{C}_{\mathbf{D}}\right)^{-1}\left[\mathbf{d}_j - f\left(\mathbf{m}_j^f\right)\right] \tag{4}$$

where $j = 1, \cdots, N_e$, $\mathbf{C}_{\mathbf{MD}}^f$ is the $N_m \times N_d$ cross-covariance matrix between $\mathbf{M}^f$ and $\mathbf{D}^f = \left[f\left(\mathbf{m}_1^f\right), \cdots, f\left(\mathbf{m}_{N_e}^f\right)\right]$, $\mathbf{C}_{\mathbf{DD}}^f$ is the $N_d \times N_d$ autocovariance matrix of $\mathbf{D}^f$, $\mathbf{d}_j = \mathbf{d} + \boldsymbol{\eta}_j$ is the jth realization of the observations, and $\boldsymbol{\eta}_j$ represents the jth random observation error.

2.2.2. ES-MDA

As ES performs only a single update, it may not be suitable for strongly nonlinear problems such as the UDM calibration, leading to the development of iterative forms of ES. Emerick and Reynolds (2013) proposed the ES-MDA method to assimilate the same observation data multiple times by inflating the covariance matrix of observation errors as $\alpha_i\mathbf{C}_{\mathbf{D}}$, in which α_i is the inflation coefficient, satisfying $\sum_{i=1}^{N_{\text{iter}}}(1/\alpha_i) = 1$. A simple choice for α_i is $\alpha_i = N_{\text{iter}}$ for all iterations. In addition, the way to compute realizations of the observations in Equation 4 should be changed accordingly:

$$\mathbf{d}_j = \mathbf{d} + \sqrt{\alpha_i}\mathbf{C}_{\mathbf{D}}^{1/2}\boldsymbol{z}_d \tag{5}$$

where $\boldsymbol{z}_d \sim N\left(0, \mathbf{I}_{N_d}\right)$.

With enough iterations, ES-MDA can generally better match observation data than the primary ES method. However, according to Equation 3, both ES and ES-MDA rely on the first two statistical moments, thus they are most suitable for problems with unimodal Gaussian distributions. For complex UDMs, the distribution of model parameters may have multiple peaks, hence there is still a need to further employ a method to cope with the possible multimodal distributions.

2.2.3. ES-ILU

Zhang et al. (2018) proposed the ES-ILU method for the estimation of model parameters with possible multimodal distributions by introducing a local update scheme into ES-MDA. The rationale is that the multimodal distributions are locally unimodal. The ES-ILU method identifies and updates each realization with a local ensemble rather than the global ensemble. For a realization $\mathbf{m}_j^f$, its local ensemble is identified by the following equation:

$$J(\mathbf{m}) = J_1(\mathbf{m})/J_1^{\max} + J_2(\mathbf{m})/J_2^{\max} \tag{6}$$

where $J_1(\mathbf{m}) = [\mathbf{d} - f(\mathbf{m})]^T\mathbf{C}_{\mathbf{D}}^{-1}[\mathbf{d} - f(\mathbf{m})]$ is the distance between the model responses $f(\mathbf{m})$ and the observations $\mathbf{d}$; $J_2(\mathbf{m}) = \left(\mathbf{m} - \mathbf{m}_j^f\right)^T\mathbf{C}_{\mathbf{MM}}^{-1}\left(\mathbf{m} - \mathbf{m}_j^f\right)$ is the distance between model parameters $\mathbf{m}$ and the realization $\mathbf{m}_j^f$, in which $\mathbf{C}_{\mathbf{MM}}$ is the $N_m \times N_m$ autocovariance matrix of model parameters; $J_1^{\max}$ and $J_2^{\max}$ are the maximum values of $J_1(\mathbf{m})$ and $J_2(\mathbf{m})$ respectively and used as scaling factors to scale the two parts within the same range of [0,1].

Then the local ensemble for $\mathbf{m}_j^f$ is selected as the realizations with $N_L = \beta N_e$ smallest J values, that is, $\mathbf{M}_{j,L}^f = \left[\mathbf{m}_{j,1}^f, \cdots, \mathbf{m}_{j,N_L}^f\right]$, where $\beta \in [0, 1]$ is the local factor. Using the local update scheme, ES-ILU updates the local ensemble $\mathbf{M}_{j,L}^f$ to $\mathbf{M}_{j,L}^a = \left[\mathbf{m}_{j,1}^a, \cdots, \mathbf{m}_{j,N_L}^a\right]$ by

$$\mathbf{m}_{j,l}^a = \mathbf{m}_{j,l}^f + \mathbf{C}_{\mathbf{MD}}^{L,f}\left(\mathbf{C}_{\mathbf{DD}}^{L,f} + \mathbf{C}_{\mathbf{D}}\right)^{-1}\left[\mathbf{d}_j - f\left(\mathbf{m}_{j,l}^f\right)\right] \tag{7}$$

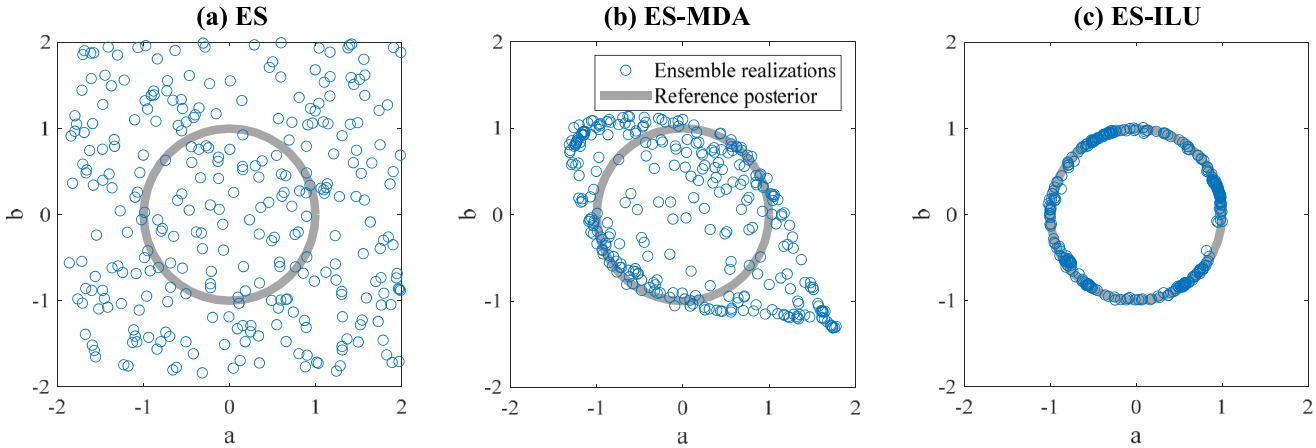

Figure 1. Results of parameter estimations for the illustrative example.

where $l = 1, \cdots, N_L$, $\mathbf{C}_{\mathbf{MD}}^{L,f}$ is the $N_m \times N_d$ cross-covariance matrix between $\mathbf{M}_{j,L}^f$ and $\mathbf{D}_{j,L}^f = \left[f\left(\mathbf{m}_{j,1}^f\right), \cdots, f\left(\mathbf{m}_{j,N_L}^f\right)\right]$, $\mathbf{C}_{\mathbf{DD}}^{L,f}$ is the $N_d \times N_d$ autocovariance matrix of $\mathbf{D}_{j,L}^f$. Then the updated realization $\mathbf{m}_j^a$ is randomly sampled from the updated local ensemble $\mathbf{M}_{j,L}^a$. In this way, the possible multimodal distributions of model parameters are well explored.

Complete schemes of the three methods are given in Algorithms S1, S2 and S3 in Supporting Information S1. Nonlinear but simple function $y = a^2 x_1 + b^2 x_2$ is used as an illustrative example to demonstrate the performance of the three ES methods. In this case, both parameters a and b follow a uniform prior distribution $U(-2, 2)$ and the observed output is $d = 1$ with observation error following $N\left(0, 0.01^2\right)$ corresponding to the model input $x_1 = x_2 = 1$. It can be easily deduced that the posterior distribution of the two parameters approximates a circle of radius 1, which means that there is an infinite number of modes in this problem. By using the setting $N_e = 300$, $N_{\text{iter}} = 6$ and $\alpha_i = 0.1$ based on a few trial tests, the parameter estimates are obtained, as shown in Figure 1. It can be observed that the ES results (Figure 1a) are not good as the ensemble realizations are distributed far from matching the reference posterior distribution (i.e., a circle). ES-MDA (Figure 1b) obtained an improved parameter estimation since the posterior distribution is more consistent to a circle; and a much better result can be further obtained by ES-ILU as shown in Figure 1c. These results indicate that ES-ILU can better handle multimodal problems involving equifinality issues.

2.3. Key Aspects of the Calibration Problem

The ideal calibration approach should be effective, robust and computationally efficient. By model robustness we mean the ability of a calibrated model to preform well across a range of rainfall events with varying characteristics. As this paper investigates different ES methods for the UDM calibration, it is necessary to investigate the key aspects of the problem that might impact the calibration effectiveness and robustness. By revisiting relevant literature (e.g., Reed et al., 2013; Swathi et al., 2019), this paper focuses on the following two key aspects.

1. **Spatial density of monitoring sites.** The information provided by monitoring sites is crucial for model calibration (Freni et al., 2009). This paper considers various monitoring schemes starting from only one monitoring site being available at the outlet of the system to increasing numbers of sites spatially distributed within the system. The aim is to investigate how the number of monitoring sites affects the results of model calibration and how many observations are required to obtain accurate calibration results for the entire model domain. It is acknowledged that the layouts of monitoring sites would also affect the performance of model calibration, which is not considered in this paper.
2. **Temporal variability of rainfall events.** Various studies on optimization-based calibration approaches have found that a calibration based on a single rainfall event cannot be normally generalized (Clemens, 2001). In other words, the calibrated model cannot guarantee good performance for another rainfall event, especially

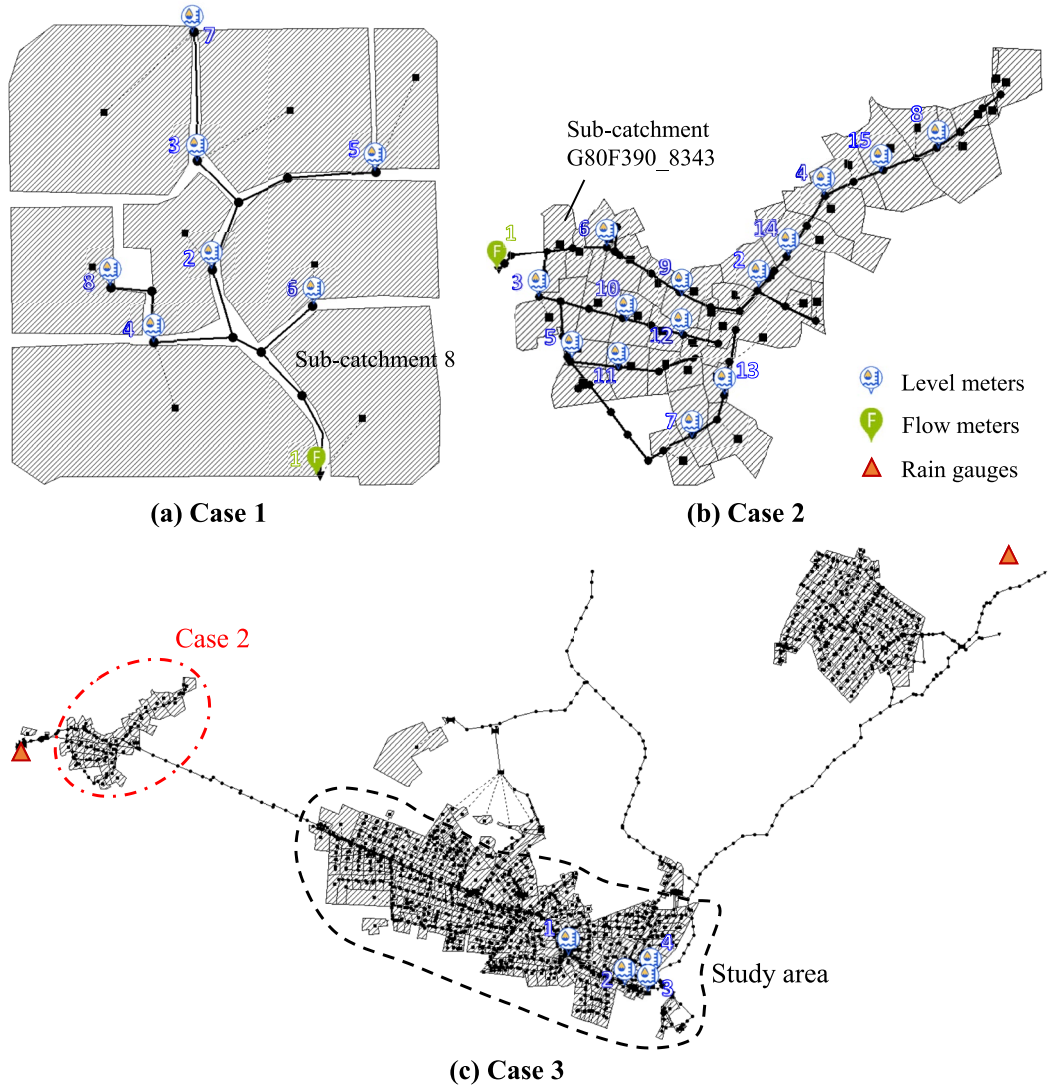

Figure 2. Urban drainage model layouts of the three cases.

when the temporal characteristic of events are quite different (Barco et al., 2008; Swathi et al., 2019). To investigate the robustness of the ES methods on rainfall events, this paper considers multiple rainfall events with varying temporal characteristics to enable model calibration and validation.

3. Case Studies

3.1. Description of the Three Case Studies

Three cases (Cases 1–3 as shown in Figure 2) of UDMs built using the SWMM model (EPA, 2020) are tested in this paper. Case 1 (Figure 2a) is taken from the first illustrative example in the SWMM software manual that consists of 8 sub-catchments, 14 nodes and 13 linkes as shown in Table 1. Both Cases 2 and 3 originate from an open-access SWMM model of the real urban drainage system in Bellinge, Denmark (Figure 2c, Pedersen et al., 2021), which is termed as the Bellinge model in this paper. Case 2 (Figure 2b) is taken from a part of the Bellinge model (i.e., the part circled by the dot dash line in Figure 2c), where a free outfall has been added to this model as the terminal.

Table 1
System Information of the Three Cases

Case	Sub-catchments	Nodes	Links	Study area (km^2)	Number of calibration parameters
1	8	14	13	0.287	$8 \times 11 = 88$
2	38	54	53	0.211	$38 \times 11 = 418$
3	476	555	558	1.539	$476 \times 11 = 5{,}236$

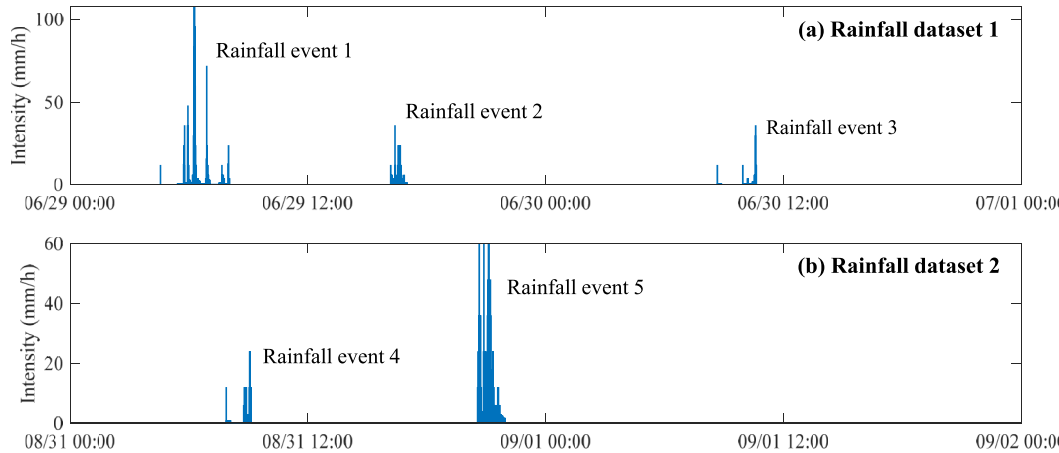

Figure 3. Five rainfall events included in two rainfall data sets for the Bellinge model.

Case 2 consists of 38 sub-catchments, 54 nodes and 53 links (Table 1). Case 3 is the original Bellinge model (Figure 3c) and field observations are used for the analysis. Therefore, Cases 1 and 2 are two synthetic systems exhibiting different levels of complexity to demonstrate the utility of the three ES methods, Case 3 is a real-world system used to further validate the effectiveness of the proposed methods in practical applications. System information of the three cases is given in Table 1.

In this study, five real rainfall events provided in Pedersen et al. (2021) are used for model inputs. Details of these rainfall events are presented in Figure 3 and Table 2. As shown in Figure 3, these five events occurred between 29 June 2012, and 30 June 2012, and between 31 August 2015, and 1 September 2015. Rain gauge locations are presented in Figure 2c. As shown in Table 2, the mean intensities of the five rainfall events range from 1.6 to 15.7 mm/hr with various peak intensities, total depths and durations. The flooding volumes of Cases 1 and 2 corresponding to the five rainfall events are obtained by model simulations given the true parameter values ranging from 0 to 589 m^3. Therefore, the analyzed five rainfall events represent a wide range of rainfall characteristics and hence can be used for effective model calibration and validation done in this paper.

A total of 8 and 15 in-situ monitoring sites, distributed across the systems analyzed, is assumed in Cases 1 and 2, respectively. They are flow meters at the outlet recording the terminal discharge flow and water level meters at manholes recording the water levels, as shown in Figures 2a and 2b. The locations of these sites are artificially selected and used to provide various monitoring schemes (see Section 3.2 for detail). Synthetic state data at these monitoring sites are generated by adding white noise to the corresponding simulation model predictions based on true parameter values for the analyzed rainfall events. Here, Gaussian white noise $\boldsymbol{\eta} \sim N(0, 0.033\mathbf{d})$ is considered in all cases to approximately represent a typical 10% relative error of the monitoring devices according to the three-sigma rule (i.e., the observation errors are within $\pm 0.099\mathbf{d}$ with 99.7% confidence). For real-world Case 3, there are four water level meters installed in the system, as shown in Figure 2c. The observed data recorded at these meters during analyzed rainfall events are used in Case 3 analyses. A part of the Bellinge model where the four observation sites are located (i.e., the study area as shown in Figure 2c) is analyzed in detail. The relevant

Table 2
Key Characteristics of the Five Real Rainfall Events

Rainfall event	Intensity (mm/h)		Total (mm)	Duration (min)	Time		Flooding volume (m^3)	
	Mean	Peak			Start	End	Case 1	Case 2
1	5.5	108	19.4	210	2012/6/29 04:33	2012/6/29 08:03	383	135
2	7.7	36	6.4	50	2012/6/29 16:10	2012/6/29 17:00	13	4
3	1.6	36	3.2	119	2012/6/30 08:39	2012/6/30 10:38	0	1
4	3.3	24	4.2	76	2015/8/31 07:52	2015/8/31 09:08	0	1
5	15.7	60	22.0	84	2015/8/31 20:33	2015/8/31 21:57	589	118

Table 3
Hydrological Parameters to Be Calibrated for the Three Cases

Parameters	Units	Descriptions	Intervals
Width	m	Width of overland flow path	$(0.5m_0, 1.5m_0)$
%Slope	%	Average surface slope	(0.01, 10)
%Imperv	%	Percent of impervious area	(0, 100)
N-Imperv	–	Manning coefficient for Impervious area	(0.01, 0.04)
N-Perv	–	Manning coefficient for pervious area	(0.1, 0.8)
Dstore-Imperv	mm	Depth of depression storage on impervious area	(0.2, 5)
Dstore-Perv	mm	Depth of depression storage on pervious area	(2, 10)
%Zero-Imperv	%	Percent of impervious area with no depression storage	(0, 100)
MaxRate	mm/h	Maximum rate on Horton infiltration curve	(20, 80)
MinRate	mm/h	Minimum rate on Horton infiltration curve	(0, 10)
DecayCoeff	1/h	Decay constant for the Horton infiltration curve	(2, 7)

components are listed in Table 1. Detailed information of the monitoring sites for the three cases is listed in Table S1 in Supporting Information S1. To clarify synthetic and real data used in the case studies, model simulations at the monitoring sites conditioned on true model parameters with noises added for Cases 1 and 2 are defined as synthetic data, while the real field data used in Case 3 are termed as observations. Note that dealing with fault data, anomalies and missing data are out of the scope of this study.

The parameters to be calibrated in the three cases are hydrological parameters associated with sub-catchments in the SWMM models. This is mainly due to: (a) sub-catchments are conceptual components with many parameters that are difficult or impossible to be measured directly; and (b) hydrological parameters are usually time-dependent, closely related to the changing land cover in urban areas. In contrast, the properties of the drainage network are relatively stable and much easier to identify. To simplify the implementation of the calibration method, no parameter reduction strategies are considered in this paper and a total of 11 hydrological parameters for each sub-catchment are calibrated. The types of calibration parameters and the corresponding ranges used for calibration are listed in Table 3 and follow the literature recommendations (Rossman, 2015; Swathi et al., 2019; Zhao et al., 2013). Note that m_0 for the parameter "width" in the table indicates its assumed true value or specified value by modeler for each sub-catchment, and the range of "width" is given as ±50% of m_0 rather than a fixed range as this parameter is empirical and different for each sub-catchment. In addition, for the two synthetic cases (i.e., Cases 1 and 2), the true values of the same type of parameter in different sub-catchments are intentionally modified to be more diverse enabling better generalization of calibration results. Modified parameter values of different sub-catchments are given in Data Sets S1–S4 in Supporting Information S1. As a result, there are 88, 418, and 5,236 parameters associated with 8, 38, and 476 sub-catchments to be calibrated for Cases 1, 2, and 3, respectively, as shown in Table 1.

Table 4
Monitoring Schemes for Cases 1 and 2

Cases	Schemes	Sites
Case 1	M1	1
	M2	1/2
	M4	1/2/3/4
	M6	1/2/3/4/5/6
	M8	1/2/3/4/5/6/7/8
Case 2	M1	1
	M3	1/2/3
	M5	1/2/3/4/5
	M10	1/2/3/4/5/6/7/8/9/10
	M15	1/2/3/4/5/6/7/8/9/10/11/12/13/14/15

3.2. Numerical Experiments for Model Calibration

A series of numerical experiments are designed and conducted with the ES methods for Cases 1 and 2, with the consideration of two key aspects of the calibration problem:

1. **Spatial density of monitoring sites**. Five monitoring schemes with different distribution densities are applied to Cases 1 and 2 respectively, as shown in Table 4, to investigate the impact of the spatial density of monitoring sites on model calibration. Note that the scheme M1 with an end-point site for the two cases is the typical approach that is widely used in practice; the scheme M8 with eight sites for Case 1 is an ideal scheme as the outlets of eight sub-catchments (i.e., the related manholes) are all observed.

2. **Temporal variability of rainfall events.** The five real rainfall events with different temporal characteristics, as shown in Figure 3 and Table 2, are used for Cases 1 and 2 to investigate the robustness of the calibration method for different rainfall events. That is, the UDMs are calibrated based on each of the five rainfall events and validated for predicting model states of other rainfall events.

3.3. Setup of the ES Methods for Model Calibration

The settings of the three ES methods for model calibration were optimally determined based on a few trial-error test runs. For Cases 1 and 2, ensembles of 300 and 1,000 realizations with 10 iterations are adopted, respectively. The inflation coefficient for ES-MDA and ES-ILU is set as $\alpha_i = N_{\text{iter}}$ for all iterations and the local factor for ES-ILU is set as $\beta = 0.2$. In terms of the settings of the ES methods for Case 3, a large number of ensemble realizations are generally required due to the high dimension of the calibration problem (i.e., 5,236 calibration parameters) for this complex real-world case. However, a large ensemble size would lead to significant computational inefficiency that has hindered the application of many optimization methods in literature (Niazi et al., 2017). To cope with this issue, we managed a solution that uses a small ensemble of 1,000 realizations with 10 iterations and optimal settings of the inflation coefficient and the local factor to obtain similar calibration and validation results to that using large ensembles. That is, the inflation coefficients were set in a decreasing order of $\alpha_i = \{10000, 5000, 1000, 500, 100, 10, 8, 7, 5, 2.3875\}$ for the 10 iterations (Emerick & Reynolds, 2013) and the local factors were unevenly set as $\beta_i = \{1.0, 0.8, 0.6, 0.4, 0.2, 0.1, 0.1, 0.1, 0.1, 0.1\}$ for the 10 iterations. Note that the optimal method setting in Case 3 reveals the potential performance improvements of the ES-MDA and ES-ILU methods. However, it is not the focus of this paper, thus the relevant results and analysis are not given herein.

To assess the accuracy and robustness of the calibrated models, two specific metrics, Nash-Sutcliffe Efficiency (NSE) and relative bias (RB; Zheng et al., 2018), are adopted to evaluate the accuracy of output states and estimated parameters, respectively. Here RB is calculated as

$$\text{RB} = |m - m_0| / (m_2 - m_1) \times 100\% \tag{8}$$

in which the absolute deviation from the true parameter, that is, $|m - m_0|$, is normalized by the parameter interval $m_2 - m_1$, to enable objective comparisons among different parameters. Note that NSE can range from $-\infty$ to 1 with NSE = 1 being the optimal and the RB can range from 0 to 1 with RB = 0 being the optimal.

4. Results and Discussions

4.1. Performances of the Three ES Methods

To demonstrate the performances of the three ES methods in the UDM calibration, a general case with one end-point monitoring site (i.e., the scheme M1 for Cases 1 and 2) was first considered as an example. The rainfall events 1 and 2 were selected for model calibration and validation, respectively. Accordingly, the synthetic data of discharge flows at site 1 as shown in Figures 2a and 2b were used to enable the model calibration and validation in this case.

4.1.1. State Estimation

Figure 4 presents flows at system outlets in Cases 1 and 2 for the three ES methods. The ensemble mean and the NSE mean value indicate the accuracy of state estimate, the 95% confidence interval (i.e., 95% CI) and the NSE standard deviation value (i.e., NSE std) represent the spread of the ensemble that indicate the uncertainty associated with the state estimate.

It can be generally observed from Figure 4 that the estimation accuracy improves in the order of ES, ES-MDA, and ES-ILU as the ensemble means match the state observation states in an increasing order. Meanwhile, the estimation uncertainties (i.e., the spreads of ensembles) decrease sequentially. This trend can be visualized more clearly by the NSE mean and std values in Figure 4, where a larger value of NSE mean and a smaller value of NSE std represent a better calibration/validation. More specifically, ES-ILU generally outperforms ES-MDA in terms of the estimation accuracy and the associated uncertainty (see the NSE mean and std values) while both the two

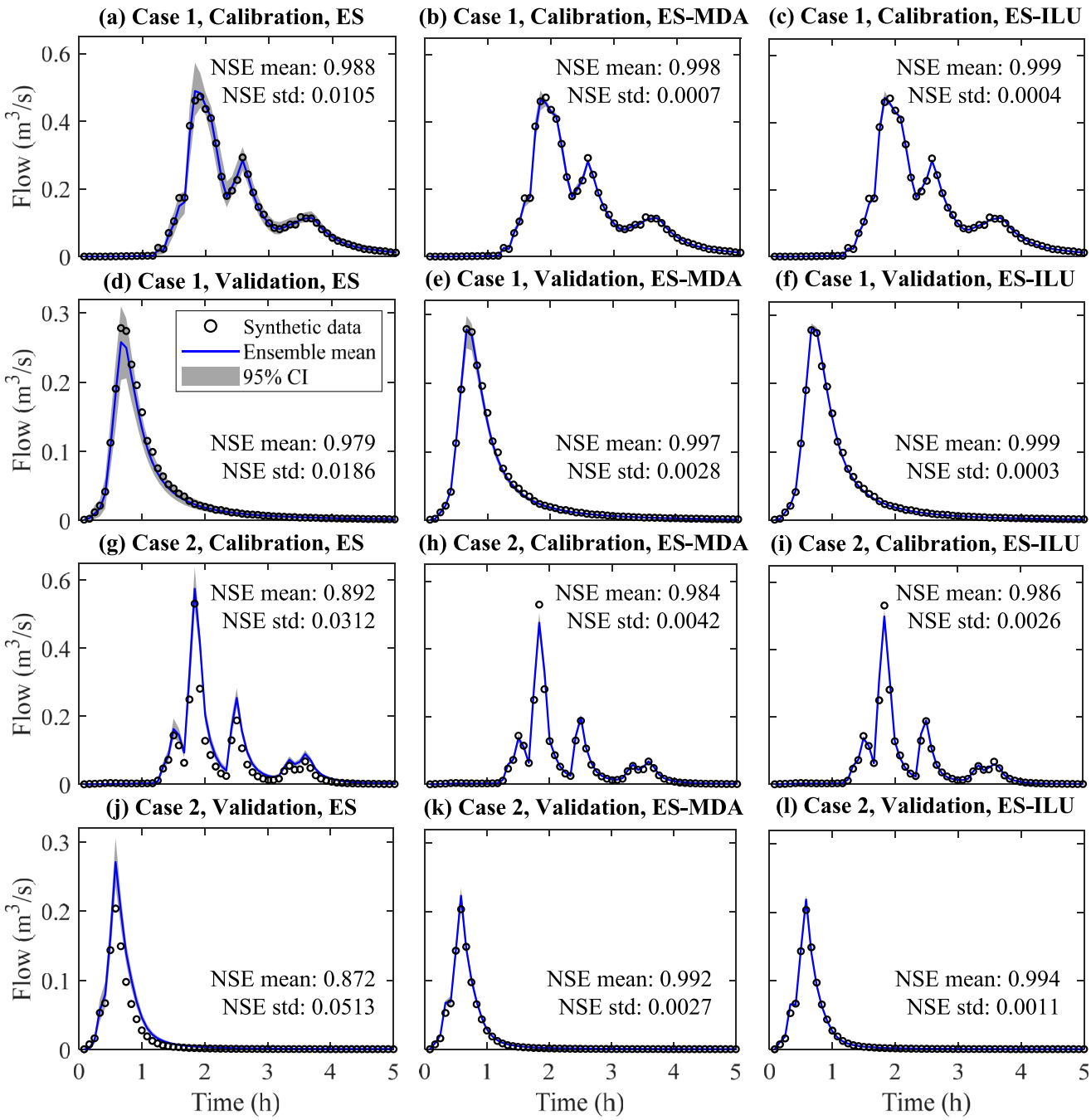

Figure 4. Calibration and validation results of output state at the end-point monitoring sites for Cases 1 and 2 by the three ensemble smoother methods.

methods can achieve highly accurate estimates for the observed states (all the NSE mean values above 0.98). The standard ES method is less effective than the former two methods (e.g., the NSE mean values are below 0.90 for Case 2 as shown in Figures 4g and 4j).

A rigorous calibration should ensure that the state estimates of the entire model match the true states, not just for the observation states. Thus, we further investigated the accuracy of calibration and validation for water levels at all nodes for Cases 1 and 2, which is represented by boxplots of NSE mean and std values in Figure 5. As it can be seen from this figure, the state estimates for the entire model are not as accurate as that for the observation

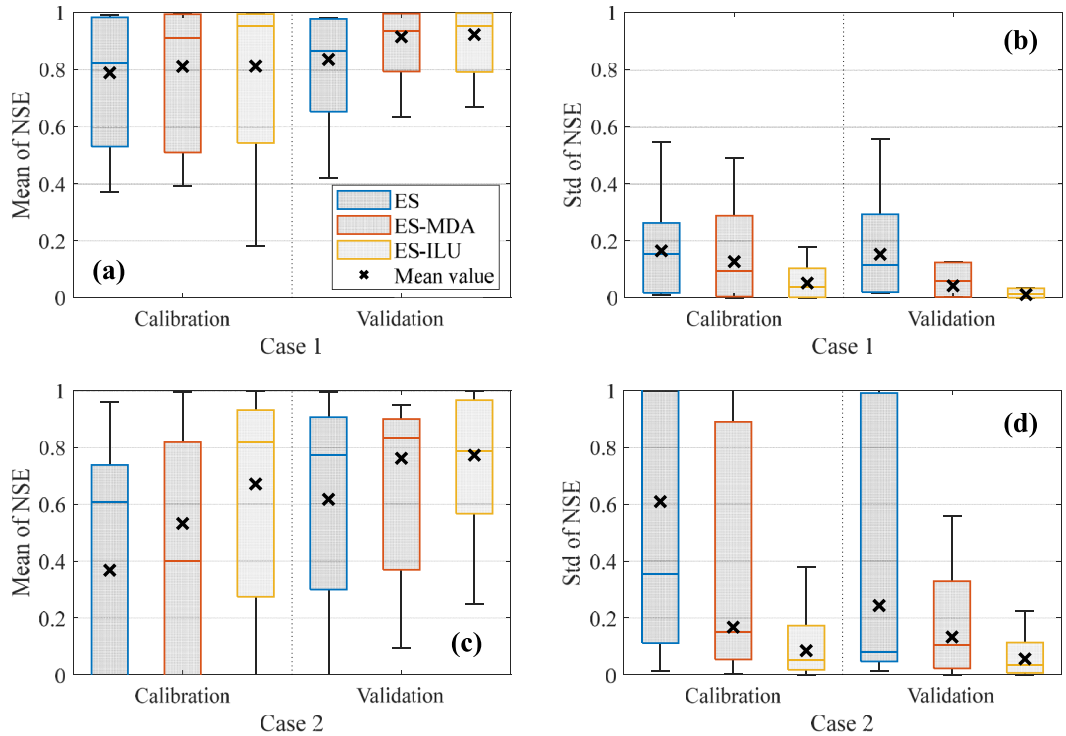

Figure 5. Boxplots of calibration and validation results (Nash-Sutcliffe Efficiency mean and std values for water levels at all nodes) for Cases 1 and 2.

state due to the distinct variation range of NSE mean and std values (i.e., box height), especially for Case 2 as shown in Figures 5c and 5d. For example, the NSE mean values range between −3.14 (not shown in the figure) and 0.96 with the mean value of 0.37 for the calibration results by ES for Case 2. These results reveal that the model calibration based on the single end-point observation cannot guarantee highly accurate state estimates for the entire model domain. This is easily overlooked in practical applications of model calibration.

Regarding the performance comparison of the three methods, it can be generally observed from Figure 5 that ES-ILU can achieve overall better state estimates than ES-MDA as the boxes of NSE mean values obtained by ES-ILU are generally higher (approaching one) than those by ES-MDA and the boxes of NSE std values by ES-ILU are lower (approaching zero) than those by ES-MDA. Similar comparison results can be observed for ES-MDA and ES, indicating that ES-ILU outperforms ES. This finding can be further verified by the mean values of boxes in Figure 5, where a larger value of the box mean for NSE mean values and a smaller value of the box mean for NSE std values represent a better calibration/validation for the entire model state.

4.1.2. Parameter Estimation

To explicitly demonstrate the parameter estimation results, the posterior distributions of the calibration parameters estimated by the three methods are compared in Figure 6. Considering that it is difficult to display a total of 88 and 418 calibration parameters for Cases 1 and 2, we only selected the parameters of two sub-catchments for demonstration purpose. It can be observed that only a few estimated parameter values are distributed close to the true values (e.g., the parameter "%Imperv" for both cases) while most of them deviate significantly. This finding suggests that despite the high accuracy of the single observation state, the parameter estimates by all three methods are inaccurate. In addition, by inspecting shapes of parameter distributions, we can identify that ES-ILU tends to get the tightest estimation results (i.e., the highest peaks) among the three methods. This is followed by ES-MDA and finally ES. This result indicates that the ability of the three methods to explore the parameter space increases in the order of ES, ES-MDA, and ES-ILU, as noted in literature (Emerick & Reynolds, 2013; Zhang et al., 2018).

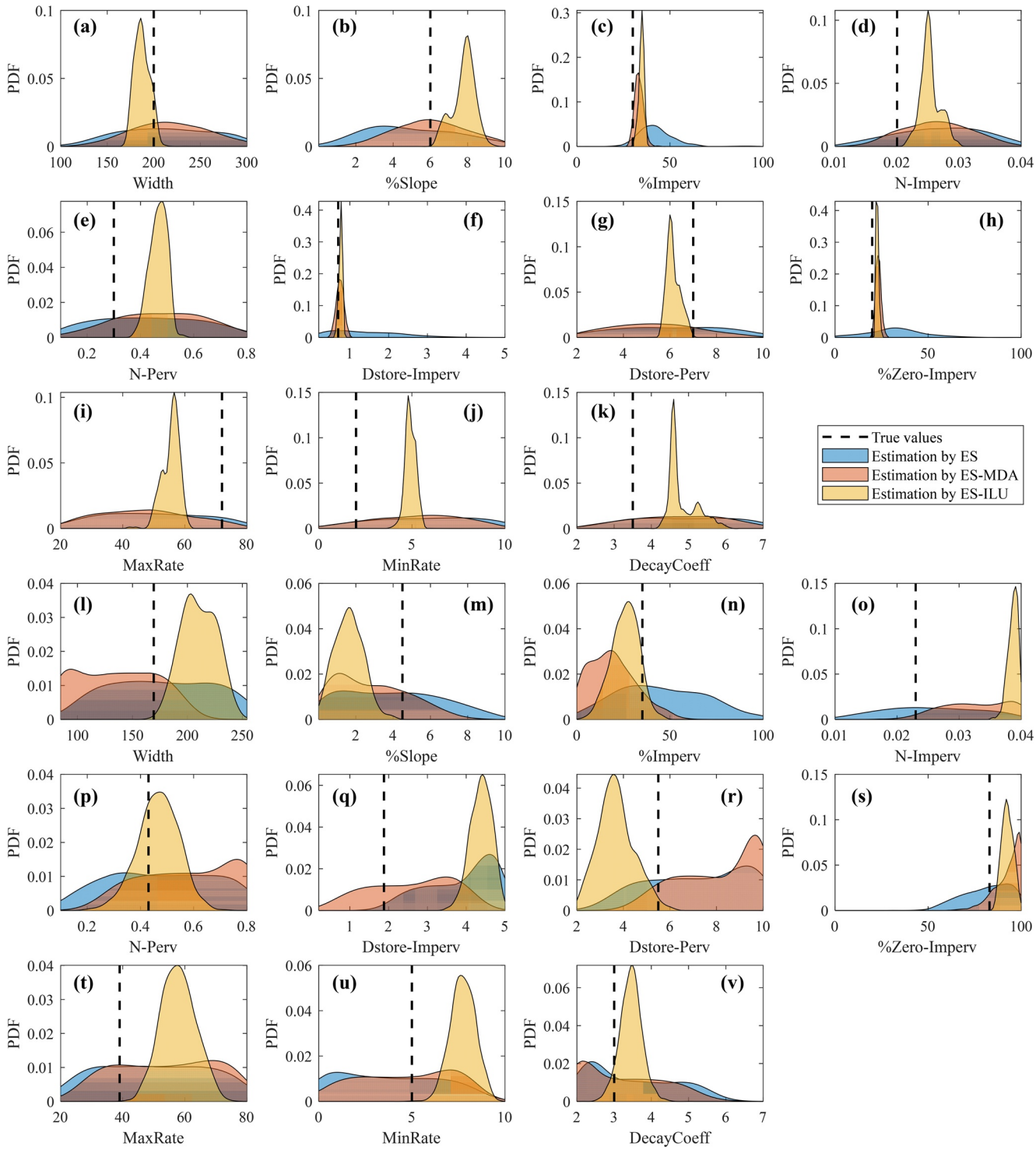

Figure 6. Posterior distributions of calibration parameters for sub-catchment 8 in Case 1 (a–k) and sub-catchment G80F390_8343 in Case 2 (l–v).

Overall, we can tentatively conclude that for the general case of model calibration and validation based on the end-point monitoring site: (a) all the three methods can produce accurate estimation for the observed states with their performances increasing in the order of ES, ES-MDA, and ES-ILU; (b) however, the accuracies of state and parameter estimation for the entire model domain are much less satisfactory, especially for parameter estimates wherein most values largely deviate from true values. The possible reason for the lack of estimation accuracy of

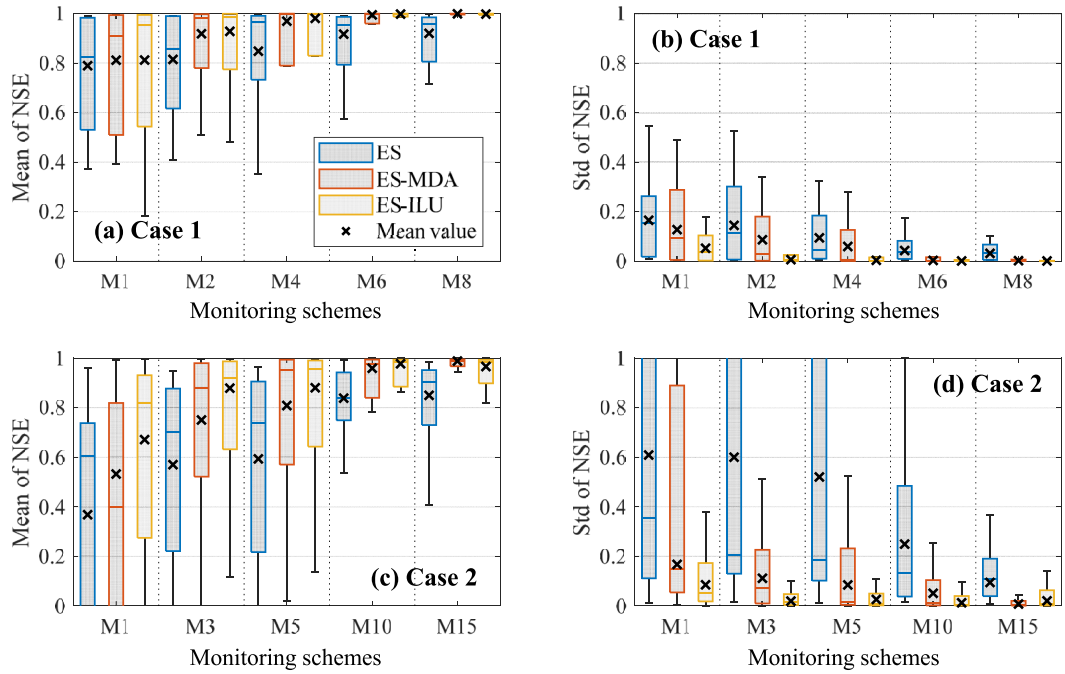

Figure 7. Boxplots of calibration results (Nash-Sutcliffe Efficiency mean and std values for water levels at all nodes) for Cases 1 and 2 utilizing different monitoring schemes.

the entire model could be attributed to the lack of observations in this case, which is further investigated in the following section.

4.2. Investigation of Key Aspects of the Calibration Problem

4.2.1. Spatial Density of Monitoring Sites

The impact of the spatial density of monitoring sites on model calibration was investigated by considering five different monitoring schemes as listed in Table 3 for Cases 1 and 2, respectively. The rainfall event 1 was used for this investigation. The calibration results for the observed states and the entire model states (i.e., water levels at all nodes) for Cases 1 and 2 are given as boxplots in Figures S1 in Supporting Information S1 and Figure 7, respectively. For the estimation of observed states (see Figure S1 in Supporting Information S1), all the three methods produce highly accurate data-fits across different monitoring schemes for both cases. Particular good results are obtained by ES-ILU and ES-MDA, where the box mean of the NSE mean values are all above 0.98 and the box mean of the NSE std values are all below 0.01. This outcome reveals that the accuracy of the fit to the observed states are good no matter how many monitoring sites are used.

However, this is not the case for the estimation of the entire model states as shown in Figure 7. For each method, both the accuracy and associated uncertainty of the model calibration for all model states improve gradually as the number of monitoring sites increases (i.e., the boxes of NSE mean and std values approach one and zero, respectively). This trend can be visualized more clearly for the increase of the box mean values in Figures 7a and 7c and the decrease of the box mean values in Figures 7b and 7d with the increasing number of sites. Such consistent trends in both cases imply that increasing the number of monitoring sites can improve the overall performance of model calibration. Specifically, when four and ten monitoring sites are deployed in Cases 1 and 2 respectively (i.e., M4 for Case 1 and M10 for Case 2), the NSE mean values for the entire model state obtained by ES-MDA and ES-ILU are overwhelmingly greater than 0.80, indicating that these two monitoring schemes can guarantee the overall high accuracy of model calibration for the two cases.

In addition, for the performance comparison among the three methods, it can be generally observed from Figure 7 that the boxes of NSE mean values (and the box mean values) gradually increase to approach the optimal value of 1 and the boxes of NSE std values (and the box mean values) decrease to the optimal value of 0 in the order

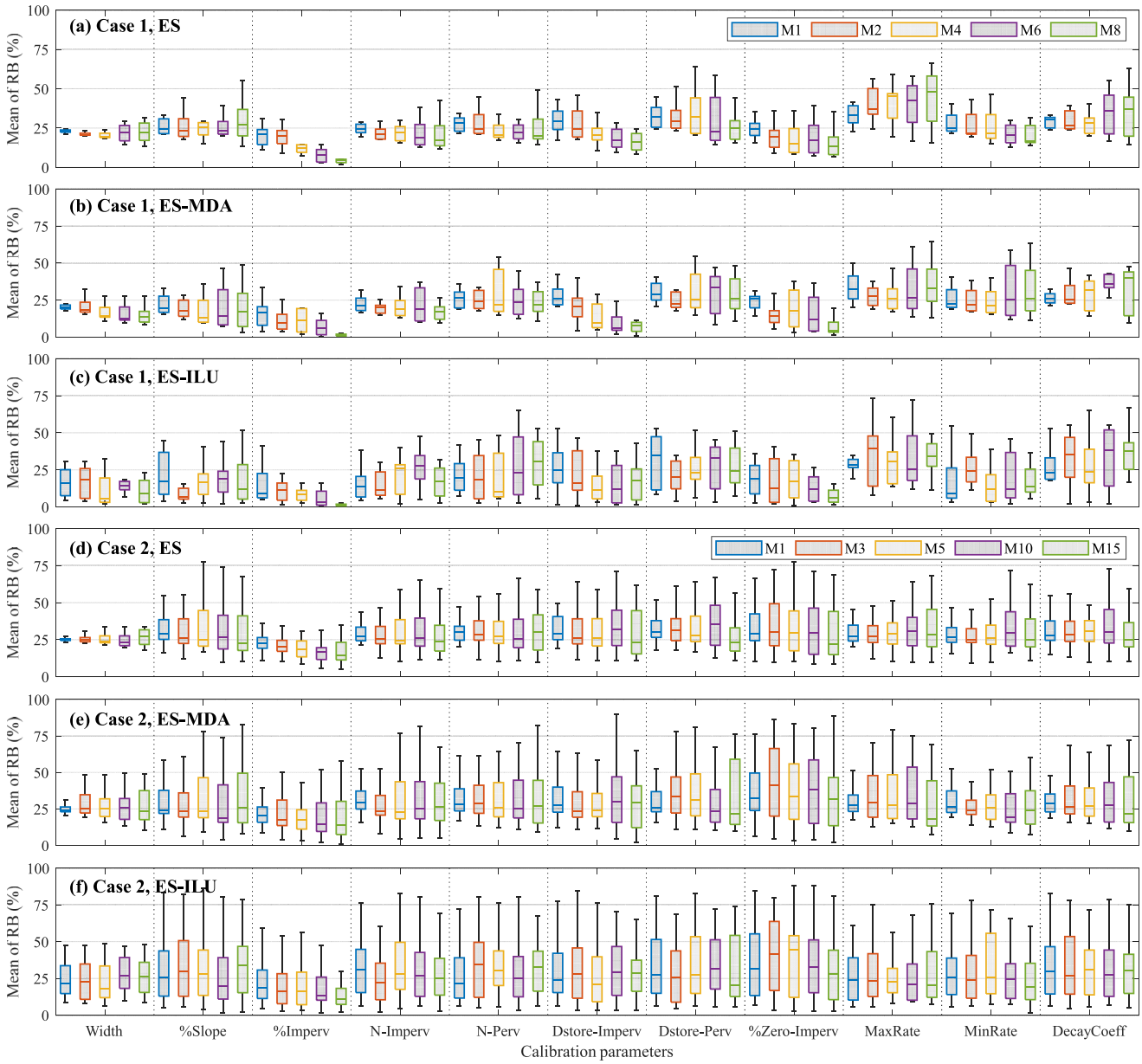

Figure 8. Boxplots of calibration results (relative bias mean values) for the 11 parameters of all sub-catchments in Cases 1 and 2 by the three ensemble smoother methods utilizing different monitoring schemes.

of ES, ES-MDA, and ES-ILU for all monitoring schemes. This finding confirms again that ES-ILU has the best performance for model calibration, followed by ES-MDA and finally ES.

To evaluate the impact of the spatial density of monitoring sites on parameter estimation, the mean and standard deviation values of the metric RB (i.e., RB mean and std values) for the 11 calibration parameters for all sub-catchments and different monitoring schemes are summarized as boxplots in Figure 8 and Figure S2 in Supporting Information S1, respectively. Similar to the meanings of NSE mean and std values, the RB mean and std values herein indicate the accuracy of parameter estimate and the associated uncertainty, respectively. It can be observed from Figure 7 that for all the three ES methods and most of the 11 parameter types, the accuracy of parameter estimates (i.e., the RB mean values) does not improve with the increasing number of monitoring sites, even when the overall state estimates are highly accurate, such as M8 for Case 1 and M15 for Case 2. A few exceptions are the parameters "%Imperv," "Dstore-Perv," and "%Zero-Imperv" in Case 1 (with a small number of

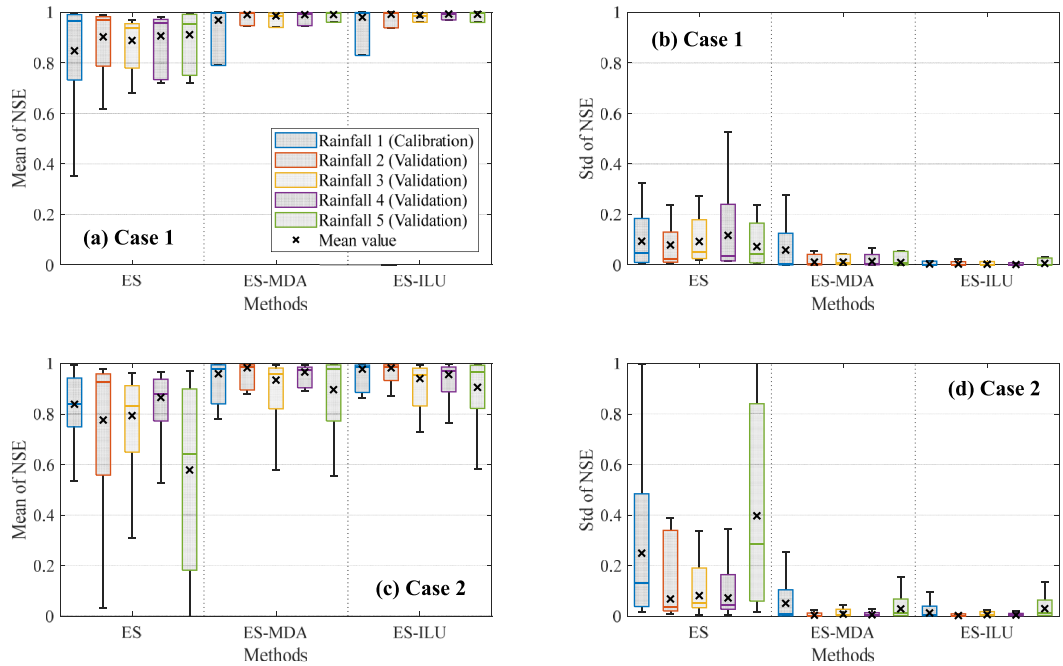

Figure 9. Boxplots of calibration and validation results (Nash-Sutcliffe Efficiency mean and std values for water levels at all nodes) for Cases 1 and 2 based on rainfall event 1.

calibration parameters) that have been widely identified as sensitive parameters in literature (Niazi et al., 2017). Such results demonstrate the complexity and challenges of parameter identifiability for UDM calibration, which is intrinsically due to the parameter equifinality effects. When equifinality situation arises, using RB cannot reasonably evaluate the performance of parameter estimation. In addition, the results in Figure S2 in Supporting Information S1 (i.e., the RB std values) show that the associated uncertainties of parameter estimates can be decreased by increasing the number of monitoring sites.

4.2.2. Temporal Variability of Rainfall Events

Five real rainfall events with different temporal characteristics as shown in Figure 3 and Table 2 were adopted to investigate the robustness of model calibration using the three ES methods. To enable a comprehensive investigation, the calibrated models based on each of the five rainfall events were validated for predicting model states of other rainfall events. The monitoring schemes M4 in Case 1 and M10 in Case 2 were utilized considering that these two schemes can guarantee the overall high performance of model calibration as analyzed earlier from Figure 7. The calibration and validation results for the five rainfall events were given in Figures 9, 10 and Figures S3–S5 in Supporting Information S1. The results for rainfall events 1 and 3 were selected for detailed analysis (see Figures 9 and 10) as these two events respectively represent intense rainfall cases that lead to serious urban floods and small rainfall cases without or with minor overflows refer to the key rainfall characteristics as listed in Table 2.

It can be generally observed from Figure 9 that for ES-MDA and ES-ILU the boxes of NSE mean and std values for rainfall events 2–5 are essentially at the same level (i.e., similar heights and variation ranges) with the boxes for rainfall event 1. More specifically, the corresponding box mean values for ES-MDA and ES-ILU are all above 0.90 for all the rainfall events. Such results indicate that the calibrated model by ES-MDA and ES-ILU based on rainfall event 1 can also achieve highly accurate model predictions for other four rainfall events. Similar results can be found for the calibration and validation results based on rainfall event 5 (see Figure S5 in Supporting Information S1).

However, this is not the general case for the validation results based on rainfall event 3 in Figure 10. For ES-MDA and ES-ILU in this figure, while the calibrated model based on rainfall event 3 can produce similarly and highly accurate model states for rainfall event 4, the predicted model states for rainfall events 1, 2, and 5 are less accurate. For example, the boxes of ES-MDA and ES-ILU in Figure 10c for rainfall events 1, 2, and 5 are much wider

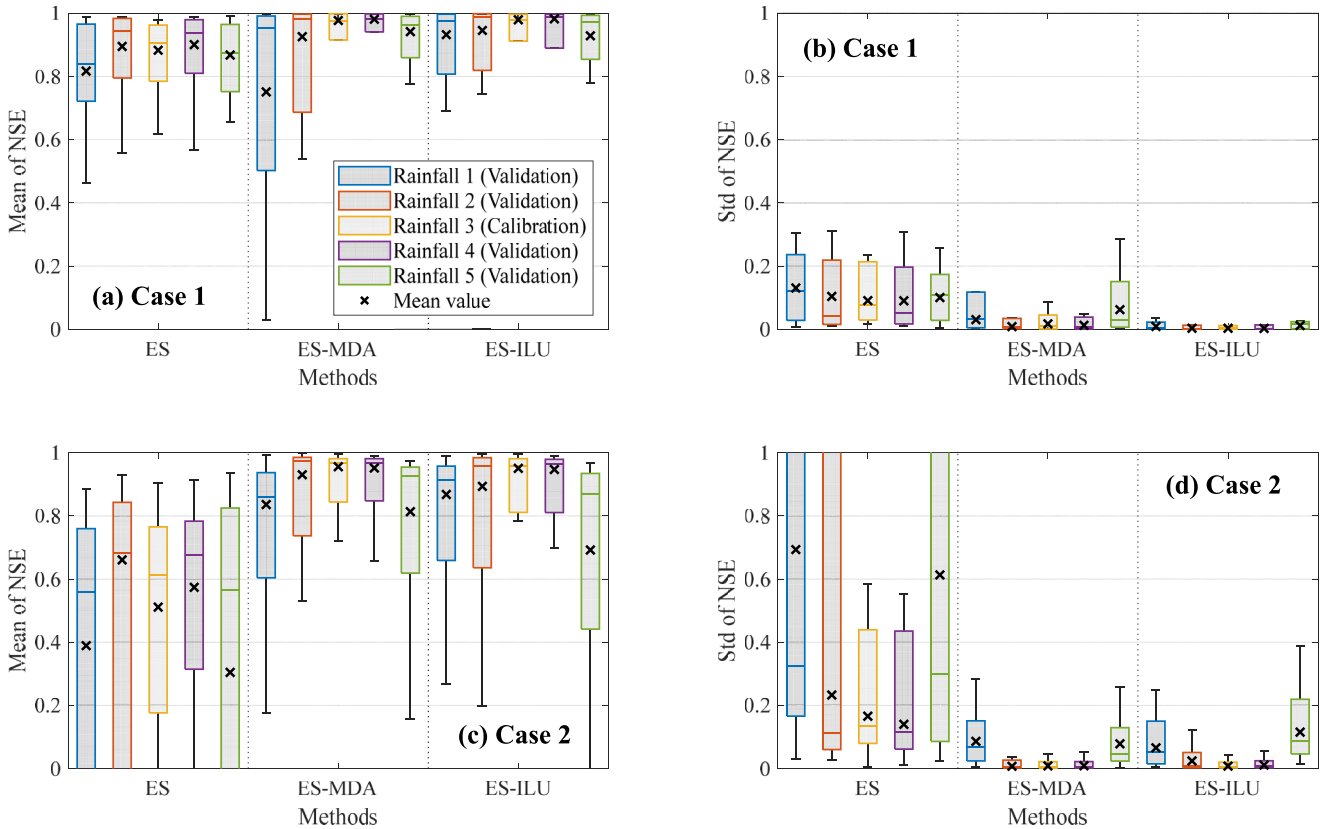

Figure 10. Boxplots of calibration and validation results (Nash-Sutcliffe Efficiency mean and std values for water levels at all nodes) for Cases 1 and 2 based on rainfall event 3.

than those for rainfall events 3 and 4, and the corresponding box mean values can be as low as 0.68. This result indicates that the calibrated model based on rainfall event 3 cannot provide reliable model performance for rainfall events with larger intensities. Similar findings can be discovered from the validation results of rainfall events 2 and 4 in Figures S3 and S4 in Supporting Information S1.

Overall, it can be generally concluded that ES-MDA and ES-ILU can produce robust UDM calibration based on intense rainfall events. The model calibration based on small rainfall events is less robust as the calibrated model cannot perform well for predicting model states for intense rainfall events. One possible reason could be that the system responses for intense rainfall events are more informative than those for small rainfall events. In addition, it can be also observed from Figures 9 and 10 that the ES method is less effective and robust than ES-MDA and ES-ILU, as previously observed in Figures 5 and 7. This can be due to the primary ES method performing only a single global update that is likely to be insufficient for the highly nonlinear problem of the UDM calibration in this paper. However, the ES-MDA and ES-ILU methods perform multiple smaller updates to the calibration problem, thus are more effective and robust than the ES method.

4.3. Further Discussion of Parameter Identifiability for UDM Calibration

From the above investigation and analysis, we can generally identify that: (a) increasing the spatial density of monitoring sites can improve the performance of estimating the entire model states; (b) however there is no clear indication of noteworthy improvement in the identification of true parameter values. Therefore, it can be deduced that significant parameter equifinality exists within the calibration process, as recognized in literature (Her & Chaubey, 2015; Kelleher et al., 2017). The issue is thus further investigated and discussed hereafter.

To explicitly demonstrate the equifinality issue, we searched out the ensemble realizations with the mean of NSE values for water levels at all nodes above 0.99 from all the numerical experiments conducted in Sections 4.2.1.

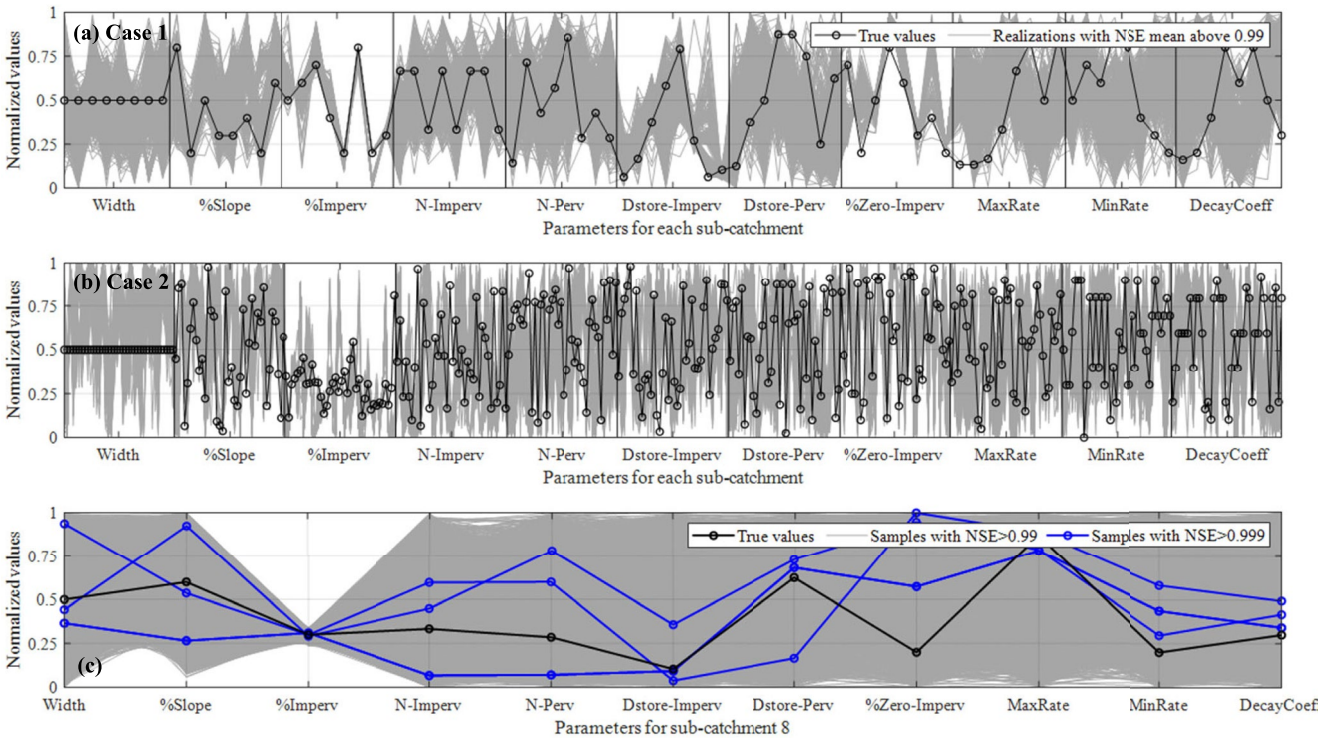

Figure 11. Parameter equifinality illustration: (a, b) parameter values of ensemble realizations with highly accurate estimation of the water levels at all nodes; and (c) parameter values of samples with highly accurate estimation of the runoff flow for sub-catchment eight in Case 1.

Then the corresponding estimated parameter values of these selected realizations were plotted in Figures 11a and 11b. Note that the values of the 11 parameters for all sub-catchments of the two cases are normalized within the 0–1 range. As shown in Figures 11a and 11b, the parameter equifinality across different sub-catchments can be clearly identified as the estimated values for each parameter exhibit a wide range within their intervals. They also differ from the true values while these realizations can all be potential solutions for practical applications (i.e., the NSE mean values above 0.99).

Based on the above findings, we further explored whether parameter equifinality exists within each sub-catchment. Here, we selected sub-catchment 8 in Case 1 as an example and conducted Monte Carlo simulations to find samples that can produce an acceptable estimation of runoff flows for the sub-catchment. As a result, the samples with NSE > 0.99 were selected and plotted in Figure 11c. Three samples with NSE > 0.999 (blue lines) and the true parameter values (black lines) were also given for comparison. It can be observed that a wide range of parameter sets with significantly different values from true parameter values can yield highly accurate simulation results (i.e., NSE > 0.99). Interestingly, parameter values of the three samples with NSE > 0.999 still differ significantly from true values. This finding, combined with that in Figures 11a and 11b, further reveal that parameter equifinality exists not only across different sub-catchments but also within each sub-catchment. Consequently, parameter equifinality underlying the calibration problem makes the unique parameter identification a significantly complex and difficult task.

Notwithstanding the difficulties with unique parameter identification, it should be noted that good state estimates can be achieved for the entire model while the parameter estimates are less accurate, as demonstrated in Section 4.2. Therefore, for practical applications where model estimation of system states are more concerned, the proposed ES-MDA and ES-ILU methods can be effective and robust for the UDM calibration. This seems to hold because these methods can identify a series of feasible parameter sets that can partially mitigate the equifinality issue.

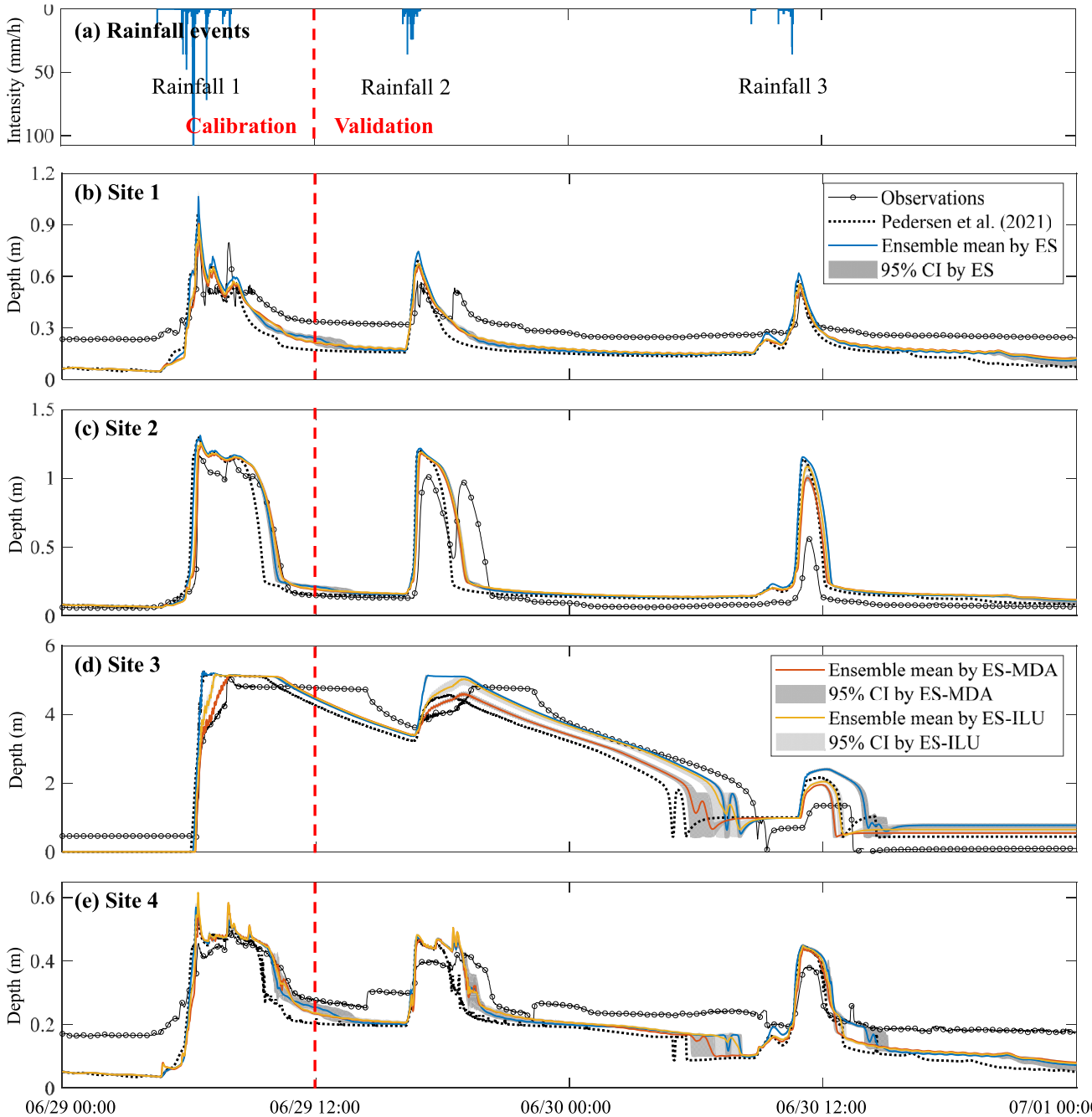

Figure 12. State trace results of calibration and validation at four observation sites for rainfall events 1–3 in Case 3.

4.4. Application to a Real-World Case

The proposed calibration methods were finally applied to a real-world Case 3 to demonstrate its utility in practical settings. As mentioned in Section 3.1, the original Bellinge model (Figure 2c) was used and a total of 5,236 hydrological parameters (Table 1) were calibrated against four in-situ water level observations (Figure 2c). The parameter ranges were given in Table 3. Two continuous rainfall data sets including five rainfall events (i.e., rainfall events 1–5 in Figure 3) were considered for model calibration and validation. More specifically, for the first data set, rainfall event 1 was used for model calibration and rainfall events 2 and 3 for model validation; for the second data set, rainfall event 4 for model calibration and rainfall event 5 for model validation, as shown in Figures 12a and 13a. Correspondingly, the observations responding to these rainfall events were splitted into two

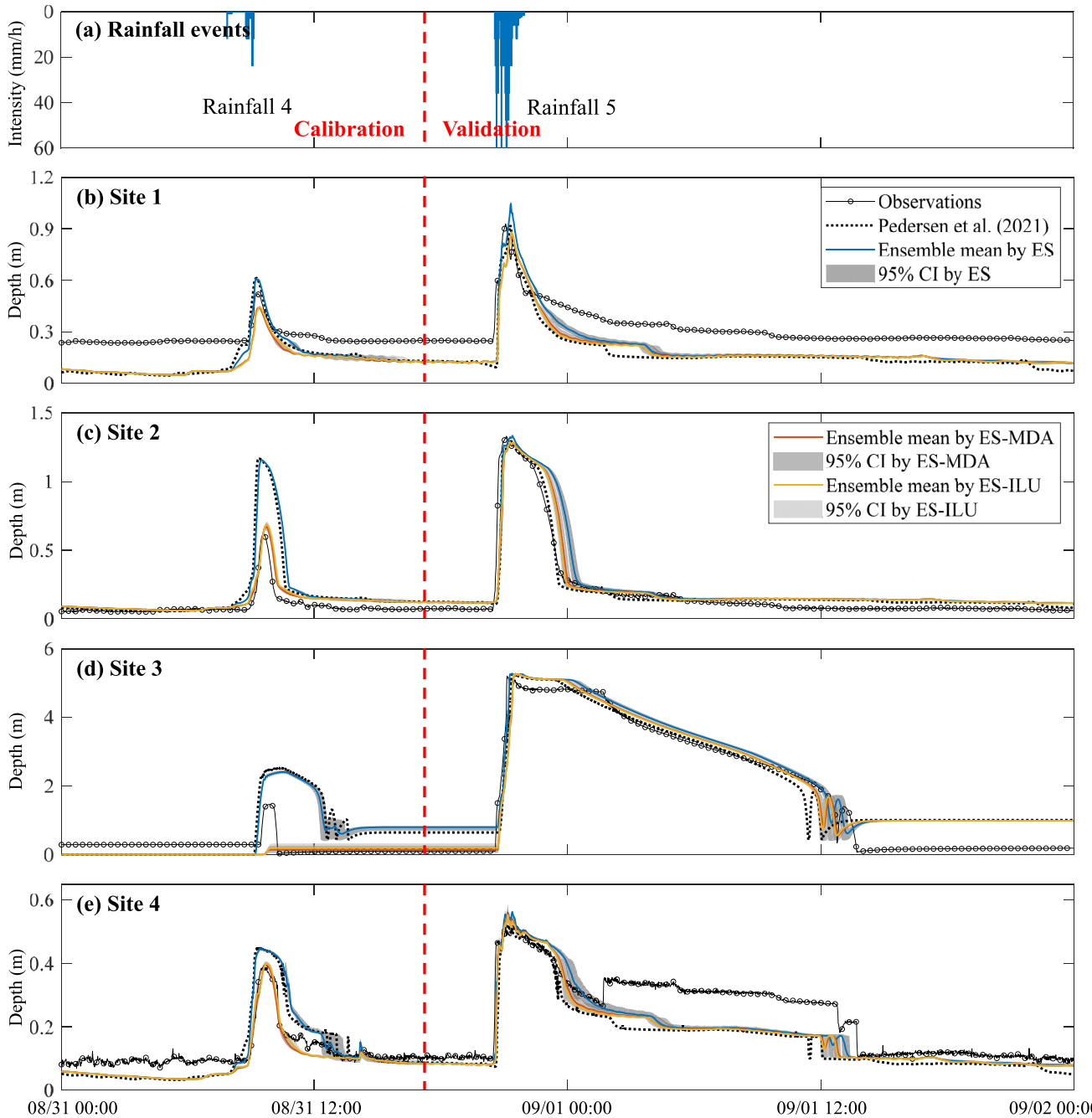

Figure 13. State trace results of calibration and validation at four observation sites for rainfall events 4 and 5 in Case 3.

parts for model calibration and validation, respectively. The results of model calibration and validation at the four observation sites are given in Figures 12b–12e and 13b–13e. The original simulation results manually validated by Pedersen et al. (2021) are also given for comparison. Statistical metrics including the NSE values for the hydrographs as well as the differences between estimated and observed water level peaks (time of occurrence and maximum value) are presented in Figure S6 in Supporting Information S1.

It can be observed from Figures 12 and 13 and Figure S6 in Supporting Information S1 that, for both calibration and validation, ES-MDA and ES-ILU produce overall similar observation-fitting results at the four sites. The results of the two methods improve upon those obtained by Pedersen et al. (2021), while ES exhibits the worst performance. For the estimation of peak states, which is often the key concern in practical applications, the peak

magnitude and timing estimates obtained by ES-MDA and ES-ILU, in most cases, have both smaller deviations than those obtained by Pedersen et al. (2021). For instance, at the calibration stage for rainfall event 4 as shown in Figure 13c, the deviations of the peak water depth at site 2 estimated by ES-MDA and ES-ILU (0.05 and 0.07 m respectively) are much lower than those obtained by Pedersen et al. (2021) (0.55 m). For the validation of the model predictions for rainfall event 3 at site 2 as shown in Figure 12c, the peak times estimated by ES-MDA and ES-ILU are 4 and 3 min earlier than the observations, respectively, while it is 15 min earlier for the estimation by Pedersen et al. (2021). This apparent advantage of ES-MDA and ES-ILU can be useful for state prediction and real-time control in practical applications. Therefore, the above results indicate that ES-MDA and ES-ILU can provide improved calibration of the complex Bellinge model. Considering that the proposed calibration methods can be fully automated (i.e., no parameter reduction strategies are involved), it is a very effective tool for practical applications.

4.5. Computational Efficiency

The calibration procedures in this paper were coded in MATLAB 2021a and run on a PC with a 10-core Intel Core i9-10900 (2.8 GHz) CPU. The times needed for the above ES, ES-MDA and ES-ILU procedures were about 0.2, 1.4, and 1.5 min, respectively, for Case 1, and 2.3, 16.7, and 26.8 min, respectively, for Case 2. For the real-world case (Case 3), the model calibration using ES, ES-MDA, and ES-ILU consumed about 1.4, 14.8, and 22.3 hr, respectively, for rainfall event 1, and 1.2, 11.2, and 14.2 hr, respectively, for rainfall event 4.

5. Summary and Conclusions

Urban drainage models (UDMs) are commonly used in urban flooding management. Due to complex and dynamic processes of urban flooding (especially the hydrological part), a typical UDM involves a large number of model parameters to represent the relevant hydrodynamic processes. Consequently, the application of such models is often limited by the difficulty of obtaining effective and robust calibration results. Limited observations, rainfall events with varying characteristics and the resultant parameter equifinality issue are the key reasons. To address these issues, this paper presented a Bayesian-based calibration method with three ensemble smoothers (ESs), that is, the primary one (ES), ES-MDA and ES-ILU. These ES methods can provide the estimates of parameter/state values and the associated uncertainty (i.e., related errors to the estimates) at the same time. Aspects that may affect the model calibration performance, that is, the spatial density of monitoring sites and the temporal variability of rainfall events, were also investigated.

The utility of the three ES methods was first demonstrated with two synthetic cases. All hydrological parameters necessary to run the stormwater management model (88 and 418 parameters for the two cases, respectively) were calibrated against various monitoring schemes and then validated for different real rainfall events. Finally, a real-world case of complex UDM with 5,236 uncertain parameters was calibrated and validated to verify the effectiveness of the proposed method in practical applications. The main results and findings can be summarized as follows:

1. When the same observation data and rainfall events are used for model calibration, the ES-ILU method generally outperforms the ES-MDA method in terms of both prediction accuracy and uncertainty while the primary ES method exhibits the worst performance. Specifically, for the real-world case, both the ES-MDA and ES-ILU methods provide better calibration results than the best-known solution manually obtained by Pedersen et al. (2021).
2. It is found that a good overall fit to the observed data does not necessarily guarantee the good accuracy of estimations for all model states. A minimum number of monitoring sites is required to enable an overall accurate model calibration. For instance, four more and ten more sites are needed to produce accurate estimates for all states in the two cases analyzed here.
3. The models calibrated using ES-MDA and ES-ILU with intense rainfall events are found to be robust when predicting model states across different rainfall events. Opposite of this, the model calibrated using the same methods but with less intense rainfall event predicts well only for less intense rainfall events. This finding suggests that the proposed ES-based calibration method is promisingly robust when the intense rainfall events are used for calibration.

4. Regarding the UDM parameter calibration, the estimated values generally deviate significantly from the true values, even when state estimates for the entire model domain are accurate. Assimilating more observations has no significant effect on improving the parameter fit, at least not in the cases analyzed here. A more detailed investigation has found that the parameter equifinality issue is ubiquitous across different sub-catchments as well as within each sub-catchment. As a result, the unique identification of UDM parameters remains an extremely complex and difficult task.

In conclusion, the proposed ES-based method is a promising alternative for the calibration of complex UDMs in terms of effectiveness and robustness. In addition, this paper provides knowledges on the impacts of observations and rainfall characteristics on model calibration performance and parameter equifinality issue, which can facilitate the implementation of UDMs in practical applications. It is noted that the ES-based calibration can output the estimated parameter/state values as well as the associated uncertainties, which is more informative than the classical deterministic approaches (e.g., optimization-based methods). However, it is also noted that the model calibration is still challenging in practical applications, for example, distinct deviations from fitting the observations still remain in the real-world case in this paper. Therefore, there is still a need to further improve the calibration performance for complex UDMs. For example, updating model structural errors together with model parameters (Evensen, 2019) or enhancing the capacity of ES with deep learning (Zhang et al., 2020) are potential solutions that can be further investigated.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The data and model files used in this paper are available at https://drive.matlab.com/sharing/0e530be0-592c-4eff-8928-4f857e553107.

Acknowledgments
This work is supported by the National Natural Science Foundation of China (Grant: 52000156; 51922096), the Excellent Youth Natural Science Foundation of Zhejiang Province, China (LR19E080003) and the Fundamental Research Funds for the Central Universities (Grant: B210201011, B210201048). Jiangjiang Zhang is supported by the Jiangsu Provincial Innovation and Entrepreneurship Doctor Program (JSSCBS20210260). The authors would acknowledge Agnethe Nedergaard Pedersen from DTU Environment, Denmark for providing the open data and models for community-wide urban drainage systems research.

References

Aanonsen, S. I., Nævdal, G., Oliver, D. S., Reynolds, A. C., & Valles, B. (2009). The ensemble Kalman filter in reservoir engineering—A review. *SPE Journal*, *14*(3), 393–412. https://doi.org/10.2118/117274-PA

Alamdari, N., Sample, D., Steinberg, P., Ross, A., & Easton, Z. (2017). Assessing the effects of climate change on water quantity and quality in an urban watershed using a calibrated stormwater model. *Water*, *9*(7), 464. https://doi.org/10.3390/w9070464

Annus, I., Vassiljev, A., Kändler, N., & Kaur, K. (2021). Automatic calibration module for an urban drainage system model. *Water*, *13*(10), 1419. https://doi.org/10.3390/w13101419

Barco, J., Wong, K. M., & Stenstrom, M. K. (2008). Automatic calibration of the U.S. EPA SWMM model for a large urban catchment. *Journal of Hydraulic Engineering*, *134*(4), 466–474. https://doi.org/10.1061/(asce)0733-9429(2008)134:4(466)

Behrouz, M. S., Zhu, Z., Matott, L. S., & Rabideau, A. J. (2020). A new tool for automatic calibration of the storm water management model (SWMM). *Journal of Hydrology*, *581*, 124436. https://doi.org/10.1016/j.jhydrol.2019.124436

Berggren, K., Olofsson, M., Viklander, M., Svensson, G., & Gustafsson, A. M. (2012). Hydraulic impacts on urban drainage systems due to changes in rainfall caused by climatic change. *Journal of Hydrologic Engineering*, *17*(1), 92–98. https://doi.org/10.1061/(asce)he.1943-5584.0000406

Clemens, F. H. L. R. (2001). *Hydrodynamic models in urban drainage: Application and calibration. PhD thesis*. Delft University of Technology.

Di Pierro, F., Djordjevic, S., Kapelan, Z., Khu, S. T., Savic, D., & Walters, G. A. (2005). Automatic calibration of urban drainage model using a novel multi-objective genetic algorithm. *Water Science and Technology*, *52*(5), 43–52. https://doi.org/10.2166/wst.2005.0105

Duan, H. F., & Gao, X. (2019). Flooding control and hydro-energy assessment for urban stormwater drainage systems under climate change: Framework development and case study. *Water Resources Management*, *33*(10), 3523–3545. https://doi.org/10.1007/s11269-019-02314-8

Emerick, A. A., & Reynolds, A. C. (2013). Ensemble smoother with multiple data assimilation. *Computers & Geosciences*, *55*, 3–15. https://doi.org/10.1016/j.cageo.2012.03.011

EPA. (2020). EPA SWMM. Retrieved from https://www.epa.gov/water-research/storm-water-management-model-swmm,%20lastaccess

Evensen, G. (2019). Accounting for model errors in iterative ensemble smoothers. *Computational Geosciences*, *23*(4), 761–775. https://doi.org/10.1007/s10596-019-9819-z

Fiori, A., & Volpi, E. (2020). On the effectiveness of LID infrastructures for the attenuation of urban flooding at the catchment scale. *Water Resources Research*, *56*(5), e2020WR027121. https://doi.org/10.1029/2020wr027121

Freni, G., Mannina, G., & Viviani, G. (2009). Assessment of data availability influence on integrated urban drainage modelling uncertainty. *Environmental Modelling & Software*, *24*(10), 1171–1181. https://doi.org/10.1016/j.envsoft.2009.03.007

Guo, X., Zhao, D., Du, P., & Li, M. (2018). Automatic setting of urban drainage pipe monitoring points based on scenario simulation and fuzzy clustering. *Urban Water Journal*, *15*(7), 700–712. https://doi.org/10.1080/1573062x.2018.1539504

Her, Y., & Chaubey, I. (2015). Impact of the numbers of observations and calibration parameters on equifinality, model performance, and output and parameter uncertainty. *Hydrological Processes*, *29*(19), 4220–4237. https://doi.org/10.1002/hyp.10487

Hutton, C. J., Kapelan, Z., Vamvakeridou-Lyroudia, L., & Savić, D. A. (2014). Dealing with uncertainty in water distribution system models: A framework for real-time modeling and data assimilation. *Journal of Water Resources Planning and Management*, *140*(2), 169–183. https://doi.org/10.1061/(asce)wr.1943-5452.0000325

Kanso, A., Gromaire, M. C., Gaume, E., Tassin, B., & Chebbo, G. (2003). Bayesian approach for the calibration of models: Application to an urban stormwater pollution model. *Water Science and Technology*, *47*(4), 77–84. https://doi.org/10.2166/wst.2003.0225

Kapelan, Z. S., Savic, D. A., & Walters, G. A. (2007). Calibration of water distribution hydraulic models using a Bayesian-type procedure. *Journal of Hydraulic Engineering*, *133*(8), 927–936. https://doi.org/10.1061/(asce)0733-9429(2007)133:8(927)

Kelleher, C., McGlynn, B., & Wagener, T. (2017). Characterizing and reducing equifinality by constraining a distributed catchment model with regional signatures, local observations, and process understanding. *Hydrology and Earth System Sciences*, *21*(7), 3325–3352. https://doi.org/10.5194/hess-21-3325-2017

Lian, H., Yen, H., Huang, J. C., Feng, Q., Qin, L., Bashir, M. A., et al. (2020). CN-China: Revised runoff curve number by using rainfall-runoff events data in China. *Water Research*, *177*, 115767. https://doi.org/10.1016/j.watres.2020.115767

Lin, R., Zheng, F., Savic, D., Zhang, Q., & Fang, X. (2020). Improving the effectiveness of multiobjective optimization design of urban drainage systems. *Water Resources Research*, *56*(7), e2019WR026656. https://doi.org/10.1029/2019wr026656

Moussa, R., Chahinian, N., & Bocquillon, C. (2007). Distributed hydrological modelling of a Mediterranean mountainous catchment—Model construction and multi-site validation. *Journal of Hydrology*, *337*(1–2), 35–51. https://doi.org/10.1016/j.jhydrol.2007.01.028

Moy de Vitry, M., Dicht, S., & Leitão, J. P. (2017). Floodx: Urban flash flood experiments monitored with conventional and alternative sensors. *Earth System Science Data*, *9*(2), 657–666. https://doi.org/10.5194/essd-9-657-2017

Niazi, M., Nietch, C., Maghrebi, M., Jackson, N., Bennett, B. R., Tryby, M., & Massoudieh, A. (2017). Storm water management model: Performance review and gap analysis. *Journal of Sustainable Water in the Built Environment*, *3*(2), 04017002. https://doi.org/10.1061/jswbay.0000817

Pedersen, A. N., Wied Pedersen, J., Vigueras-Rodriguez, A., Brink-Kjær, A., Borup, M., & Steen Mikkelsen, P. (2021). The Bellinge data set: Open data and models for community-wide urban drainage systems research. *Earth System Science Data*, *13*(10), 4779–4798. https://doi.org/10.5194/essd-13-4779-2021

Reed, P., Hadka, D., Herman, J., Kasprzyk, J., & Kollat, J. (2013). Evolutionary multiobjective optimization in water resources: The past, present, and future. *Advances in Water Resources*, *51*, 438–456. https://doi.org/10.1016/j.advwatres.2012.01.005

Rossman, L. A. (2015). *Storm water management model user's manual version 5.1* (p. 353). United States Environment Protection Agency.

Salvadore, E., Bronders, J., & Batelaan, O. (2015). Hydrological modelling of urbanized catchments: A review and future directions. *Journal of Hydrology*, *529*, 62–81. https://doi.org/10.1016/j.jhydrol.2015.06.028

Stuart, A., & Zygalakis, K. (2015). *Data assimilation: A mathematical introduction*. Springer.

Swathi, V., Srinivasa Raju, K., Varma, M. R. R., & Sai Veena, S. (2019). Automatic calibration of SWMM using NSGA-III and the effects of delineation scale on an urban catchment. *Journal of Hydroinformatics*, *21*(5), 781–797. https://doi.org/10.2166/hydro.2019.033

Vonach, T., Tscheikner-Gratl, F., Rauch, W., & Kleidorfer, M. (2018). A heuristic method for measurement site selection in sewer systems. *Water*, *10*(2), 122. https://doi.org/10.3390/w10020122

Wang, S., Zhang, Z., Sun, G., Strauss, P., Guo, J., Tang, Y., & Yao, A. (2012). Multi-site calibration, validation, and sensitivity analysis of the MIKE SHE Model for a large watershed in northern China. *Hydrology and Earth System Sciences*, *16*(12), 4621–4632. https://doi.org/10.5194/hess-16-4621-2012

Wu, J. Y., Thompson, J. R., Kolka, R. K., Franz, K. J., & Stewart, T. W. (2013). Using the storm water management model to predict urban headwater stream hydrological response to climate and land cover change. *Hydrology and Earth System Sciences*, *17*(12), 4743–4758. https://doi.org/10.5194/hess-17-4743-2013

Xinhua News Agency. (2022). Investigation report on "July 20" torrential rain disaster in Zhengzhou, Henan Province. Retrieved from http://www.gov.cn/xinwen/2022-01/21/content_5669723.htm

Xue, L., & Zhang, D. (2014). A multimodel data assimilation framework via the ensemble Kalman filter. *Water Resources Research*, *50*(5), 4197–4219. https://doi.org/10.1002/2013wr014525

Zhang, J., Lin, G., Li, W., Wu, L., & Zeng, L. (2018). An iterative local updating ensemble smoother for estimation and uncertainty assessment of hydrologic model parameters with multimodal distributions. *Water Resources Research*, *54*(3), 1716–1733. https://doi.org/10.1002/2017wr020906

Zhang, J., Zheng, Q., Wu, L., & Zeng, L. (2020). Using deep learning to improve ensemble smoother: Applications to subsurface characterization. *Water Resources Research*, *56*(12), e2020WR027399. https://doi.org/10.1029/2020wr027399

Zhao, D., Chen, J., Wang, H., & Tong, Q. (2013). Application of a sampling based on the combined objectives of parameter identification and uncertainty analysis of an urban rainfall-runoff model. *Journal of Irrigation and Drainage Engineering*, *139*(1), 66–74. https://doi.org/10.1061/(asce)ir.1943-4774.0000522

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

Zheng, F., Westra, S., & Leonard, M. (2015). Opposing local precipitation extremes. *Nature Climate Change*, *5*(5), 389–390. https://doi.org/10.1038/nclimate2579