Journal of Hydrology 674 (2026) 135490

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Out of bounds: the use of model ensembles to explore model structural uncertainty under extreme events

Margherita Evangelisti [a,*], Vincent Pons [b,c], Spyros Pritsis [b], Vittorio Di Federico [a], Franz Tscheikner-Gratl [b], Marco Maglionico [a]

[a] Department of Civil, Chemical, Environmental and Materials Engineering, University of Bologna, Viale del Risorgimento 2, 40136, Italy
[b] Department of Civil and Environmental Engineering, Norwegian University of Science and Technology (NTNU), Trondheim N-7491, Norway
[c] Department of Civil, Environmental, and Natural Resources Engineering, Luleå University of Technology, 97187 Luleå, Sweden

ARTICLE INFO

This manuscript was handled by Ashok Mishra, Editor-in-Chief, with the assistance of Emad Hasan, Associate Editor

Keywords:
Structural uncertainty
Boundary condition
Model diagnosis
Urban drainage modelling

ABSTRACT

The response of an urban drainage model to extreme events cannot be investigated independently from the abstraction process adopted in model development. The work explores structural model uncertainty, and particularly the impact of boundary conditions, using an Italian catchment that has been the subject of several modelling studies over the years. While previous research has focused on parameter optimization, limited attention has been devoted to model structure and boundary conditions. However, during extreme events, not only model parameters but also boundary conditions significantly influence system behavior.

Model structural uncertainty is investigated through the development of four model configurations organized into two model structure classes: the first class reflects the original system structure, while the second incorporates downstream boundary condition based on a re-analysis of the system. Within each class, two models were developed and calibrated by including or excluding extreme events. Each configuration is represented by an ensemble of models, which were then subjected to synthetic extreme storms. Disagreement among their responses is used to diagnose structural uncertainty.

Moreover, the study analyzed the risks associated with using a model calibrated on extreme events without properly accounting for system structure, underlining the potential consequences of such an approach. Given the expected increase in frequency and intensity of extreme events under future climate conditions, the work emphasizes the central importance of contextual knowledge and boundary conditions in model development.

A thorough understanding of system structure and context is a necessary prerequisite for effective calibration of model parameters. Parameter calibration cannot be separated from knowledge of model structure: rather than being a purely automatic process, it must be guided by the modeler's conscious choices.

1. Introduction

Model uncertainty has three dimensions: location or source, type, and nature (Tscheikner-Gratl et al., 2017; Walker et al., 2003). Location is where the uncertainty manifests in the model complex. The type is distinguished by the knowledge about the possible outcomes of a model and the probability of the occurrence of these outcomes ranging from the unattainable ideal of determinism to deep uncertainty. The nature of uncertainty can be defined by the knowledge paradigm: epistemic uncertainty describes uncertainty due to lack of knowledge, aleatory uncertainty represents the inherent variability of the examined system, and ambiguity the simultaneous presence of multiple equally valid frames of knowledge (Van Der Keur et al., 2008).

According to Deletic et al.(2012), sources of uncertainty in urban drainage models can be divided into three groups: model input, calibration, and model structure. Model input uncertainties can be present in model parameters and input data. Calibration uncertainties arise from the chosen approach for model calibration and the calibration data used.

Abbreviations: AMSL, Above Mean Sea Level; BMA, Bayesian Model Averaging; CDF, Cumulative distribution function; GLUE, Generalised likelihood uncertainty estimation; KS, Kolmogorov-Smirnov distance; MRC, Multiplicative Random Cascade model for downscaling; NSE, Nash-Sutcliffe Efficiency; SA, Sensitivity Analysis; SWMM, StormWater Management Model; UDS, Urban drainage system.

* Corresponding author.
E-mail address: margherita.evangelisti@unibo.it (M. Evangelisti).

https://doi.org/10.1016/j.jhydrol.2026.135490
Received 8 August 2025; Received in revised form 3 March 2026; Accepted 9 April 2026
Available online 12 April 2026

Model structure uncertainties depend on how well the model simulation represents the system and processes. They are also influenced by the modeler conscious choices in guiding the modelling development. The process of structuring a model can be divided in five formal steps: conceptual physical structure, conceptual process structure, spatial variability structure, equation structure, and computational structure (Gupta et al., 2012). The choices made in these steps will inadvertently introduce structural uncertainty to the model. More specifically, the first two steps (i.e., model conceptualization) relate to how well the model reflects our understanding of the system's structure, boundaries, and behaviour. Simplifications or incorrect assumptions made in these steps can lead to conceptual uncertainty. Spatial variability and equation structure relate to how the modelled processes are mathematically represented. Ill-posed or inadequate representation of the processes can also lead to structural uncertainty. Finally, the choice of computational structure can lead to uncertainty stemming from the numerical methods chosen to resolve the mathematical equations.

Despite a growing body of research on urban drainage modelling, model structure uncertainty has received little attention (Broekhuizen et al., 2019; Del Giudice et al., 2015). Significant efforts have been spent in development several tools to mitigate parameter uncertainty, however limited attention has been given to model structure and boundary conditions, although Rieckermann (2016) explicitly mentioned that in the uncertainty quantification process, model structure deficits need to be always considered. This imbalance may be due to the lack of a general methodology for assessing the effects of model structure uncertainty (Refsgaard et al., 2006), or to the difficulties in distinguishing it from other sources of uncertainty.

Some attempts at systematically assessing structural uncertainty in urban drainage models can be found in literature. Refsgaard et al. (2006) proposed a framework to deal with model conceptual uncertainty, based on the development and calibration of a set of plausible conceptual models that adequately represent the ranges of possible system behaviors. Predictions from the model set are then compared with field data: models with low predictive capabilities are then discarded and deeply investigated to gain insights into the origins of structural uncertainty. Pedersen et al. (2022b) provides real situations where model structure uncertainty can be encountered and provides a framework for diagnosis model structure errors by looking at hydraulic signatures. Broekhuizen et al. (2019) examined the differences in model structure of three models (SWMM, MIKE, MOUSE) used for urban drainage and their impact on simulation outcomes: the authors emphasize that model structure should be considered as an influential source of uncertainty in urban drainage modelling. Freni et al. (2024) proposed a Bayesian model averaging (BMA) approach to address the often unrecognized influence of this source of uncertainty. BMA is a statistical framework designed to address model structural uncertainty by combining predictions from multiple models into a single probabilistic forecast. Integrating information from multiple models, the method potentially reduces bias and provides more robust uncertainty estimates. However, BMA focuses on predictive aggregation rather than diagnosing the influence of individual model structures, and it requires sufficiently long datasets to reliably estimate model weights (Höge et al., 2019). Recently, Pritsis et al.(2024) investigated the effects of structural uncertainty in the network designing practice, which arose when engineering students were asked to design a network using a Chicago hyetograph. Additionally, Chrysochoidis et al. (2025) suggested the use of multi-modelling approach (detailed hydrodynamic, conceptual, hybrid machine-learning, and empirical modelling approach) when dealing with particulate pollutant in CSO. A similar outcome emerged in the work of Kreikenbaum et al. (2004), where the authors reflected on the errors in the modelling of complex systems resulting from simplification and incomplete knowledge of the underlying processes. They argue that the only way to estimate the magnitude of structural error is to compare concurrent models and computer programs.

Neglecting model structure uncertainty and focussing only on parameter optimization – i.e. the process of adjustment model parameters to improve the goodness of fit to measured data – might not be enough to develop a useful model. By doing so, there is a risk of overlooking the iterative process needed to ensure the model aligns with the intended objective/its intended use (Jakeman et al., 2006), and of addressing only superficially the assumptions that constrain the model structure and boundary conditions (Refsgaard et al., 2006). In short, if the structure of the model is not suitable for representing the phenomena of interest, under the range of expected conditions, fine-tuning the parameters does not guarantee a useful model.

In this work, model conceptual structural uncertainty is investigated considering the sewer network that drains an experimental catchment in Bologna (Italy). Under normal flow conditions, this network can be considered disconnected from the rest of the system, with no upstream inputs and assumed independence from downstream flow conditions. The analysis of the model subjected to extreme storm events (i.e. those that have an impact on the sewer infrastructure and fall outside the general range for which predictions are considered reliable) raises doubts about the independence of the subsystem and provides the opportunity to investigate an aspect of model structure uncertainty. The impact of boundary conditions choice is analyzed by evaluating disagreements among model ensembles calibrated for different storms and boundary conditions.

Focusing on model structure uncertainty and its relative impacts becomes even more important in future conditions. Projected changes in frequency and magnitude of extreme events in the Mediterranean area (Caillaud et al., 2021) may challenge the validity of assuming an independent subsystem. Model ensembles were stressed using different future storm conditions to estimate the potential increase in structural uncertainty in future conditions.

In this context, the current study investigates the impact of changing conditions on structural uncertainty. In particular, based on our case of inadequate boundary conditions under extreme event, we suggest and test a methodological framework designed to detect model structural inconsistencies by analyzing disagreements among model ensembles. Specifically, based on the analysis of our case study, we i) develop models with different structures to investigate their effects on the parametrization, ii) stress those models with extreme events under current conditions to investigate model disagreement, and iii) compare model disagreement under current condition with results obtained under future conditions characterized by extreme events.

2. Materials and methods

Exploring the independence of the subsystems led to analyzing the model structure and its response to weather conditions. To support this investigation, the methodology depicted in Fig. 1 was developed, and includes the following steps:

- Model structures, step 1: Two different model structures based on different underlying assumptions are considered: the original model structure, consistent with that used in previous research, and a modified model structure, including downstream boundary conditions.
- Model configurations, step 2: Both model structures are associated with two different sets of events that are expected to lead to different impact of structural uncertainty. This results in four different model configurations. As shown in the step 2 of Fig. 1, a distinction is made between a first subset of major rainfall events and a second subset of "extreme" events. The latter subset is expected to reveal the impact of structural uncertainty on the sewer system.
- Model calibration, step 3 and 4: a Monte Carlo filtering approach is used to explore the parameter's behavioural regions for each model configuration. For each configuration, an Ensemble of Models is generated. A Pareto front is used for model selection when multiple objective functions are considered.

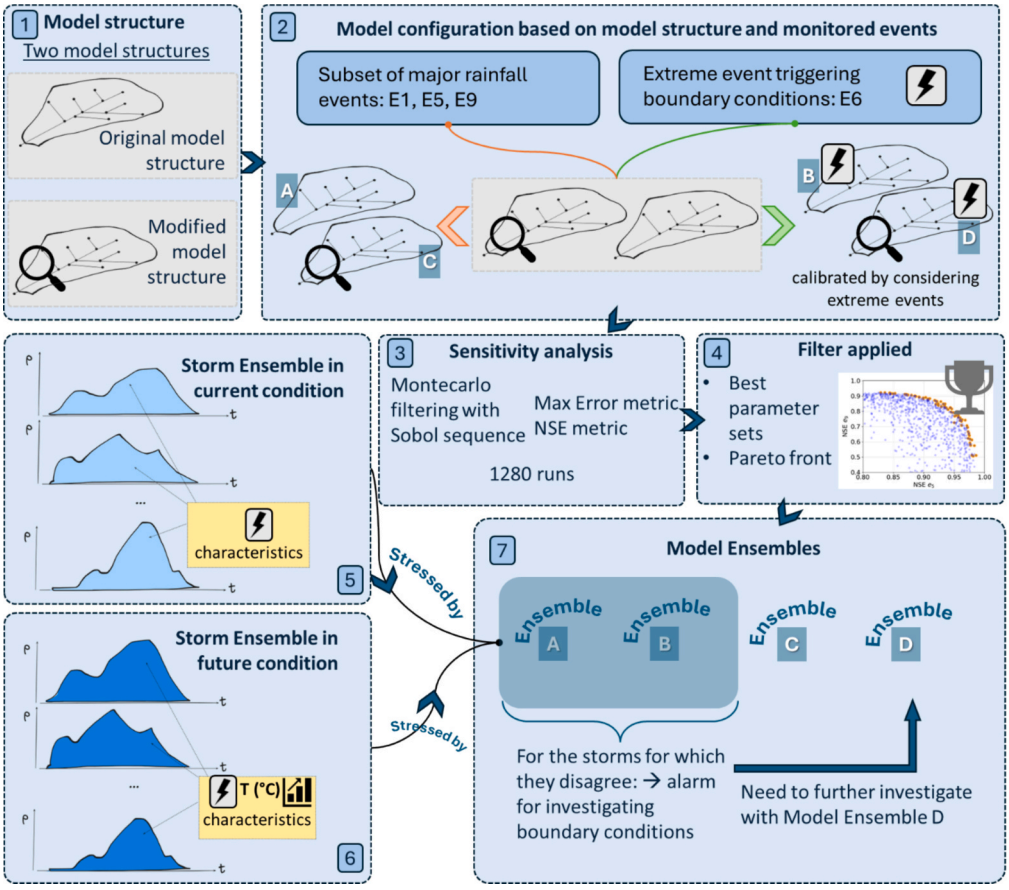

Fig. 1. Workflow adopted for the analysis of structural uncertainty.

- Model Ensembles associated with the original model structure (A and B) are stressed by a synthetic Storm Ensemble generated with the same characteristics of the observed extreme event (step 5). A second synthetic Storm Ensemble is also generated to represent future conditions (step 6).
- The final step (step 7) includes the evaluation of the disagreement between outputs of Model Ensembles A and B, which serves as an indicator that boundary conditions have been triggered (i.e., structural uncertainty). Events associated with the highest model disagreement are then used to stress the model ensemble based on the modified model structure, which include downstream boundary conditions.

2.1. *Case study: Fossolo catchment boundaries conditions within Bologna sewer system*

Fossolo catchment, a 41 ha neighbourhood of Bologna, is drained by a combined sewer network, considered independent from the surrounding sewer networks, as it does not receive inputs from upstream. The catchment has a gentle topographical gradient oriented from southwest to northeast, with an average slope of 0.3% and elevation ranging from 67.90 to 60.60 AMSL. The catchment was extensively monitored during the 90′s, for both water quantity and quality (Artina et al., 1997; Marinelli et al., 1997). Monitoring relied on a single measurement point located on the terminal pipe of the catchment network (see Fig. 2), which has a polycentric cross-section with a maximum height of 1.44 m and a maximum width of 1.88 m.

Measurements from this experimental activity have been exploited during the past 20 years for multiple calibration exercises. For example, Mannina and Viviani (2010) used the observed data to calibrate their useful tool to simulate water quantity and quality processes in urban drainage and to estimate model uncertainty by means of the GLUE methodology, whereas Freni and Mannina (2010) applied a Bayesian uncertainty estimation approach; De Paola et al. (2018) tested an innovative optimization procedure based on the harmony calibration algorithm.

The model adopted in these studies had the configuration presented in Fig. 2: disconnections from the global sewer network were represented by "outfall" objects (Out_1, Out_2 and Out_3). This set-up is certainly suitable given the geometry of the network: points Out_2 and Out_3 show a difference in the elevation of about 0.80 m between the Fossolo network and the main one. Outfall Out_1 does not coincide with the connection to the global sewer network, but it is placed slightly upstream to avoid any interference. The final pipe draining the analysed catchment discharges into the global sewer network – which has a polycentric section 2200 x 1760 mm – at approximately 0.80 m above the invert.

However, model redevelopment and recalibration in a different context raises doubts about the independence of the subsystem. In particular, we found that the outfall Out_1 is influenced by the flow in the downstream collector of the sewer system in extreme rainfall conditions. Preliminary investigations into the behavior of the Fossolo sewer network can be found in Supplementary Information (S1).

2.2. *Development of models with concurrent hypothesis*

Structural sources of uncertainty have been addressed by developing two classes of models based on different concepts and level of complexity (Reichert, 2012): the first class reflects the **original assumed structure of the system**, as described above, with outfalls Out_1, Out_2 and Out_3 of type "normal", while the second proposes a **new configuration,** based on a re-analysis of the system. A "normal" outfall

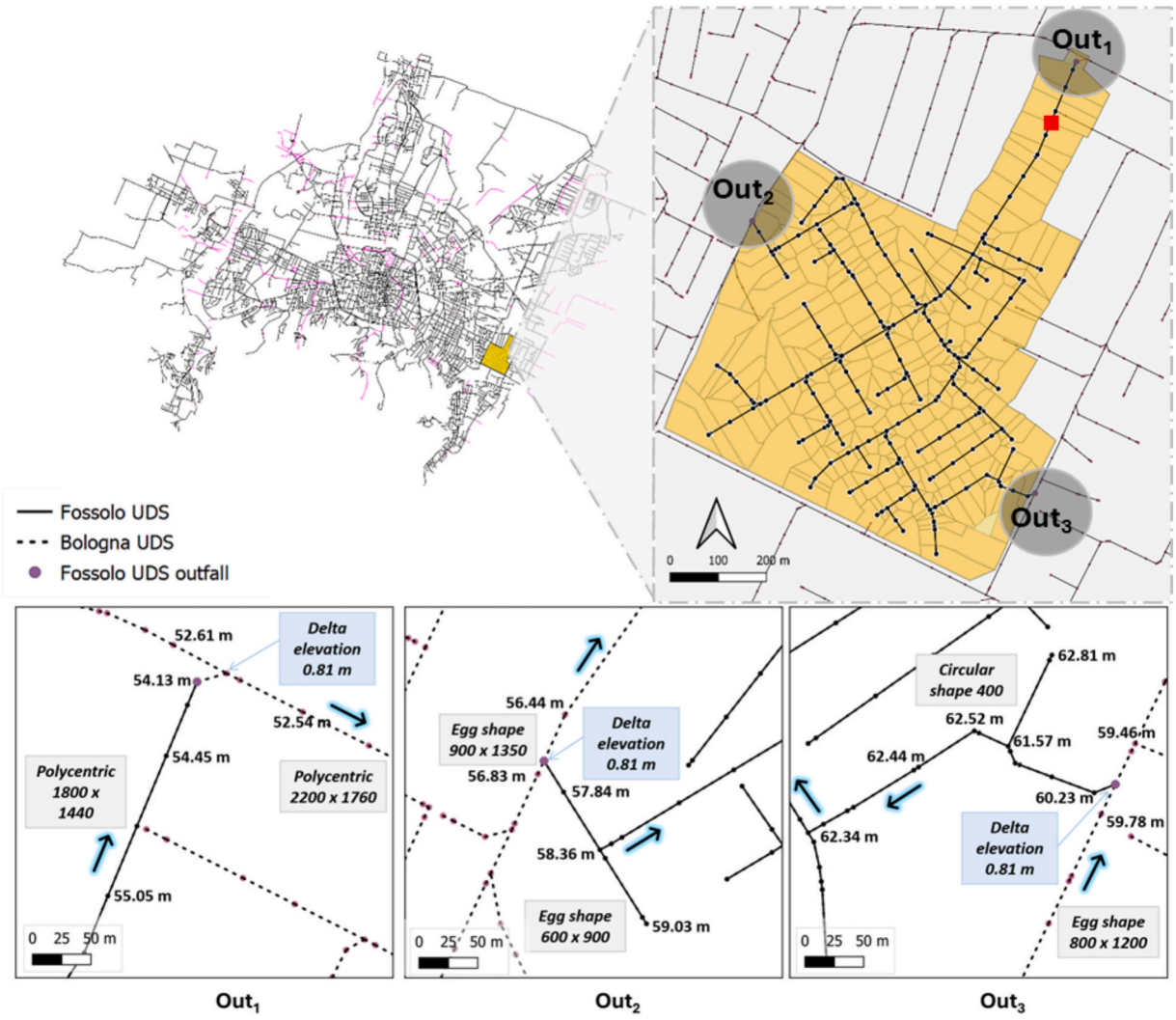

Fig. 2. Bologna sewer network with a focus over Fossolo catchment: the red square indicates the location of water level sensor during past monitoring activity. The panels show detailed views of the three outfalls (Out_1, Out_2, and Out_3) of the Fossolo sewer network. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

condition is based on normal flow conditions in the connected conduit. The new configuration, specifically, assumes that at point Out_1, the system is influenced by downstream conditions. In this model, Out_1 is treated as an outfall of type "timeseries": the full-city-scale sewer model is used to input water level timeseries as a boundary condition at the outlet Out_1. Even though the timeseries is affected by uncertainty, – since, due to the size and complexity of the network, a procedure of automatic calibration has not been applied to the full-city-scale model – this approach helps refine the understanding of system behaviour driven by interactions with the downstream conditions. The performance of the full-city-scale model was recently reassessed using water level measurements provided by the local water stakeholder at strategic locations within the network, yielding acceptable results (Nash-Sutcliffe Efficiency around 0.57 – 0.77).

Both model classes were built with EPA's Storm Water Management Model (SWMM), version 5.2, an open-source software widely used for planning, analysis and design of urban drainage systems (Rossman, 2010). SWMM models were run with the Dynamic Wave solver, which applies the full Saint-Venant equations with calculation step variable between 1 s during wet periods to 1 min during dry periods. Simulation outputs are reported at a 1-minute resolution. Model simulations, sensitivity analysis and scenarios development were performed in Python using the *swmm-api* package (Pichler, 2022).

Two model classes led to the creation of **four models**, each with different assumptions which represent distinct real-world scenarios.

- **Original structure**: the model structure assumes the original configuration, assuming the Fossolo sewer network to be entirely independent of the main system. Point Out_1 is type "Normal" outfall.

Model A: The model is characterized by a parameter set derived by not considering extreme events in the calibration process. This model reflects situations where a calibrated model is needed (e.g. for planning or management) but data on extreme events are unavailable and the detailed structure of the network is not fully known, or it is intentionally simplified for modelling purposes.

Model B: The model is characterized by a parameter set derived by considering only extreme events in the calibration process. This approach is particularly relevant in cases where municipalities or stakeholders prioritize the evaluation of system performance under rare, high-intensity conditions, often due to their critical impact on infrastructure and public safety. However, it also reflects situations where the detailed structure of the network is not fully known, or it is intentionally simplified for modelling purposes.

- **New configuration structure:** the models assume an interaction between the subsystems in correspondence of point Out_1.

Model C: The model is characterized by a parameter set derived by not considering extreme events.

Model D: The model is characterized by a parameter set derived by considering only extreme events.

Fig. 3 provides a summary of the four models developed, highlighting their key differences. The models differ respect to two primary aspects: the assumed independence between the Fossolo sewer network and the main drainage system, and the adopted parameter sets, which were calibrated either by considering extreme events or by excluding them.

2.3. Assessment of structural uncertainty impact under current conditions

2.3.1. Set of events

A set of four events was selected, with their main characteristics presented in Table 1. The four models described above were subjected to an event-based simulation approach. Model A and Model C were subjected to Event 1, 5, and 9, whereas Model B and D were stressed by Event 6. The simulated water levels were compared with observed values at the final section, whose main hydraulics characteristics are reported in Supplementary Material (S1). For Models C and D, the new inputs assigned to outfall Out_1 were derived from water levels simulated using the full-city-scale sewer network model, by assuming a spatial uniform distribution of rainfall over the Bologna catchment. This assumption can be considered a limitation for the representation of convective extremes; however, for Event 6, which corresponds to the extreme rainfall event, comparable rainfall intensities were recorded at the regional rain gauge located in the centre of the Bologna catchment, at a distance of approximately 4 km. Rainfall characteristics of Event 6 are provided in Supplementary information (S2). For the following considerations, in order to reduce computational costs and obtain a more reliable estimation of the fit for the initial portion of the event, only this section — with a duration of 3 h and a total depth of 38.32 mm — is considered, as it corresponds to the period when the highest rainfall intensity was recorded. During the remaining part of Event 6, rainfall intensities were modest and the system returned to reduced flow conditions; therefore, including this portion would not provide additional information relevant to the analysis. Observed rainfall data were recorded at 1-minute resolution using two rain gauges located within the catchment, as described by (Artina et al., 1997).

The simulations were conducted in an event-based configuration: no formal spin-up/warm-up period was applied due to the limited dimensions and high imperviousness of the drained catchment. Initial hydraulic conditions were set assuming empty pipes and nodes: to ensure hydraulic stabilization prior to rainfall onset, a 20-minute dry-weather period was included at the beginning of each event simulation. This duration was verified to be sufficient for the system to reach steady dry-weather flow conditions before the storm.

Table 1
Main characteristics of the rainfall events used for the analysis. *For Event 6, the values in parentheses refer to the portion of the event used in the analysis (first 3 h), corresponding to the period with the highest rainfall intensity.

Event	Date	Duration (min)	Precipitation Depth (mm)	Maximum Rainfall Intensity (mm/h)
1	25/04/94	77	7.82	26.10
5	28/10/94	289	23.06	60.00
6	23/06/95	538 (180*)	72.72 (38.32*)	147.97
9	13/11/95	921	42.65	60.00

2.3.2. Sensitivity analysis

The four models were subjected to a sensitivity analysis (SA) to evaluate whether the best parameter set is common to all models and to identify the most influential parameters (Saltelli et al., 2010). Parameter sampling was performed using Sobol sequences, designed to reduce variance and improve the uniformity of the sampling point distribution. Model parameters selected for the sensitivity analysis and relative boundaries are reported in Table 2. Baseline baseflow values were derived from an analysis of yearly water consumption data for 2004, provided by the local water utility: these values were imposed as constant inflows during the simulations. Baseline imperviousness values were obtained through a preliminary GIS-based spatial analysis of the study area. A total of 1280 simulations per model were carried out, and a Monte Carlo filtering approach was applied for the subsequent analysis based on two criteria, as detailed in Section 2.3.3. The objective function adopted for evaluating the agreement between hydrographs is the Maximum Error; this metric captures the worst-case error between simulated values y_i and observed values $\hat{y}_i$:

$$MaxError(y, \hat{y}) = \max(|y_i - \hat{y}_i|)$$

The choice of the Maximum Error metric was motivated by the need to provide a non-cumulative, absolute measure of divergence, in contrast to commonly used cumulative metrics such as RMSE or NSE. This choice allows to specifically capture peak divergences, which are of primary interest in the study.

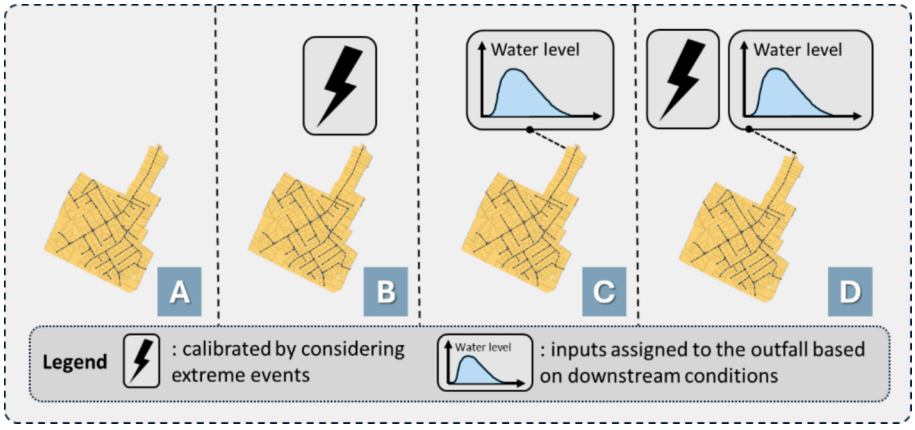

Fig. 3. Development of four models based on different assumptions: the models differ in the assumed independence between the Fossolo and the global sewer network, and in the parameter set derived from a calibration exercise that either include or exclude extreme events.

Table 2
SWMM parameters selected for calibration and their selected boundaries.

	Parameter	Unit	Lower Bound	Upper bound
Horton – Infiltration model	Max Infiltration rate	mm/hr	50	150
	Min Infiltration rate	mm/hr	10	40
	Decay	1/hr	2	8
Hydrological losses	Depression storage in pervious zones	mm	0.8	5
	Depression storage in impervious zones	mm	0.1	1
Roughness	Pipe roughness	$s/m^{1/3}$	0.010	0.020
	Impervious surface roughness	$s/m^{1/3}$	0.01	0.035
	Pervious surface roughness	$s/m^{1/3}$	0.1	2
Base flow	Base flow in dry period (*variation)	m^3/s	0.85	1.15
Imperviousness	Percentage of impervious surfaces (*variation)	%	0.85	1.15

2.3.3. *Synthetic storm events based on Event 6*

An ensemble of 1000 synthetic storms was generated using a multiplicative random cascade model (MRC) (Pons et al., 2022) trained on historical rainfall time-series from Bologna (Supplementary Material – S6). The approach is similar to the one described in Pritsis et al. (2024): total precipitation depth, storm duration and temperature have been set considering the characteristics of Event 6 (storm duration 3 h, total precipitation depth 38.32 mm). Temperature used for generating the ensemble was sampled from the distribution of temperature observed in the summer months in the period from 2006 to 2016. The synthetic events were designed to stress-test Model A and Model B, with the aim of identifying whether a tipping point exists – i.e., a specific storm configuration for which the models begin to behave differently – and, if so, to determine what that tipping point is. The size of the ensemble (1000 storms) was made to keep the total computational load relatively low, following the practice of Pritsis et al. (2024).

However, since Model A and Model B consist of multiple parameters sets derived from prior sensitivity analysis, with considerable redundancy observed among them, a filtering process was applied before proceeding to the stress-test phase.

The Pareto front was applied to filter the parameter sets in model A (Leimgruber et al., 2018). It consists of a set of non-dominated solutions that simultaneously satisfy three objective functions: maximizing NSE_e1, NSE_e5, and NSE_e9. In this step, Nash-Sutcliffe Efficiency was preferred due to the resulting shape of the Pareto front. Because this is a three-objective optimization problem, the Pareto front is inherently three-dimensional. In Fig. 4, it is represented through three two-dimensional projections, each showing the relationship between a pair of objectives. The resulting Pareto front contains 74 solutions (orange points in Fig. 4). From these, a subset of six parameter sets was extracted to ensure a good distribution along the front. The selection logic was as follows: we included the points corresponding to the maximum and minimum values of each individual objective function (4 magenta points), and two additional points corresponding to the median distances between the Pareto front solutions and the point (1,1). Table 3 outlines the models included in Ensemble A as previously described. Notably, some of the selected models (a_1 and a_2) contextually perform well with respect to one objective function but poorly with respect to another.

Regarding Model B, the use of a Pareto front was not applicable, as only a single objective function needed to be satisfied. Therefore, to obtain an ensemble comparable to that of Model A, the six best parameter sets were selected from Model B. The selection criterion coincides with the condition Maximum Error lower than 0.25 m: this threshold has an indirect physical interpretation, as the corresponding simulated water level timeseries represent flow conditions ranging from approximately 83% to 100% of full-pipe capacity.

The overall process resulted in six parameter sets for Model Ensemble A and six parameter sets for Model Ensemble B.

Once Model Ensembles A and B were created, they were stressed using the generated storms. Model agreement was evaluated by calculating the absolute maximum difference in terms of water level at the final section of the catchment:

$$Max\ Disagreement(m_1, m_2) = \max(|water\_level_{m_1} - water\_level_{m_2}|)$$

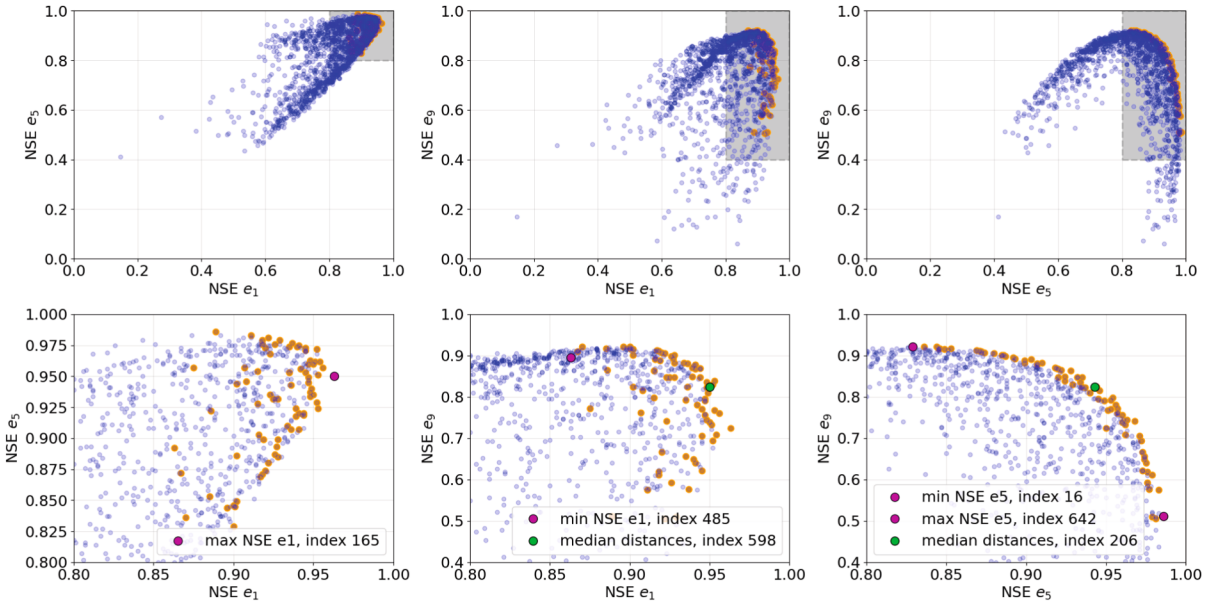

Fig. 4. Pareto-based filtering of Model A parameters sets. The first row shows the two-dimensional projections of the Pareto front, while the second row provides a zoomed view on the front. Orange points represent the parameter sets that simultaneously maximize NSE e1, NSE e5, NSE e9. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Characteristics of filtered models within Model Ensemble A.

Model	NSE E1	NSE E5	NSE E9
a_1	0.889	0.986	0.511
a_2	0.900	0.829	0.922
a_3	0.963	0.950	0.725
a_4	0.863	0.892	0.895
a_5	0.936	0.943	0.824
a_6	0.950	0.935	0.825

Legend	
	Max NSE for Event
	Min NSE for Event
	Median Distance

This metric was preferred over RMSE or NSE since it provides a non-cumulative, absolute measure of divergence.

2.4. *Impact of structural infrastructure under changing conditions: Climate and system modification*

The ensemble of storms generated with the current temperature conditions is used to estimate the likelihood of triggering the boundary conditions in the system under extreme events. Future climatic conditions, in particular in the Mediterranean region, are expected to lead to an increased frequency and magnitude of summer convective storms, according to recent advances in convection-permitting climate models (Caillaud et al., 2021). Such changes in weather patterns may affect the likelihood of triggering the system's boundary conditions.

To investigate the effect of non-stationary climate conditions on structural uncertainty and boundary conditions, a second ensemble of storm events, with similar characteristics in terms of precipitation depth and duration but different temperature distribution, was generated. The MRC model generates hyetographs using temperature as an input which influences the final hyetograph. For generating the future ensemble, we considered a timeseries of daily summer temperature from 2060 to 2070 derived from a convection permitting climate model (Raffa et al., 2023):

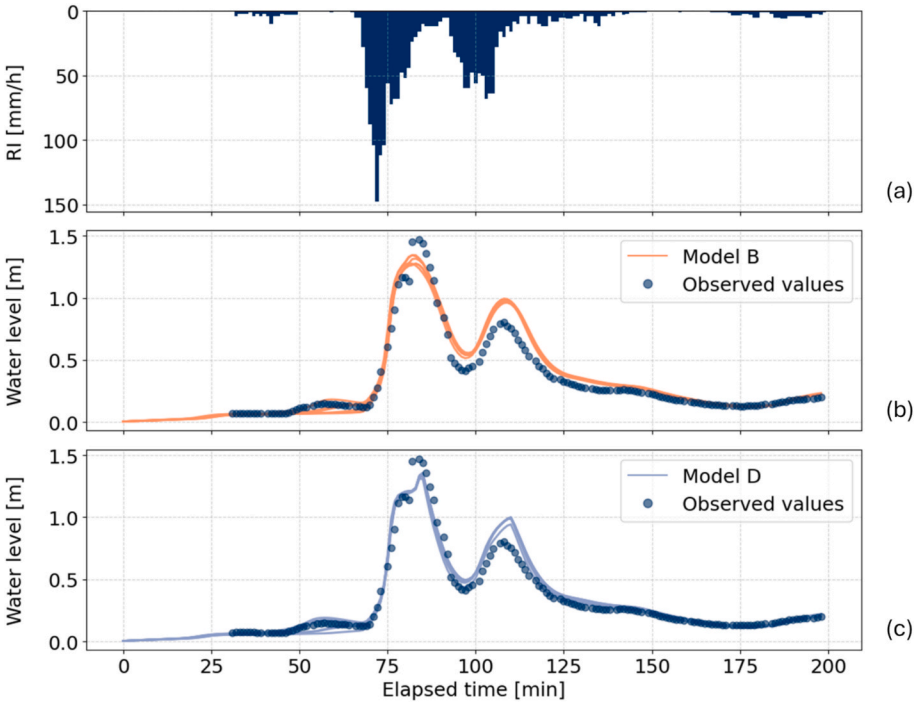

Fig. 5. Simulated and observed water level trends in the final section of Fossolo network for Event 6. (a) Rainfall intensity (mm/h) of Event 6. (b) Water level trend simulated by the original model (Model B) and observed values. (c) Water level trends simulated by the new configuration model (Model D) and observed values.

the projections are based on the COSMO-CLM model under the RCP 8.5 scenario, and are provided at hourly resolution. The resulting distribution of future daily summer temperatures was bias-corrected by estimating the bias between modelled and observed temperature distributions for the period 2006–2016 using empirical quantile mapping.

The proportion of extreme events triggering the boundary conditions is evaluated in a similar way to the current conditions in order to estimate the increase in structural uncertainty in future conditions due to inadequate boundary conditions at Out_1.

3. Results and discussion

3.1. Characterization of the structural source of uncertainty

3.1.1. Sets of events

Fig. 5 shows the simulated and observed water levels for Event 6, using the best parameter sets of Model B and D that satisfy the condition of a Maximum Error lower than 0.25 m: the number of parameter sets meeting the criterion is limited (six and seven respectively). The definition of the threshold is physically motivated, as the corresponding simulated water level timeseries represent flow conditions ranging from approximately 83% to 100% of full-pipe capacity. Indeed, observed water levels in the pipe reached a maximum of 1.44 m, indicating a situation of full pipe operation: the corresponding estimated flow, with the adoption of the rating curve, is around 4.1 m^3/s (Artina et al., 1997). While the original configuration (Model B, Fig. 5(b)) fails to reproduce the shape of the observed water level, the revised configuration (Model D, Fig. 5(c)), which includes downstream boundary conditions derived from the use of the full-city-scale model, affected by uncertainty, is consistent with the observed trend. The peak simulated by Model D was delayed by four minutes compared to Model B, and appears to reproduce the condition of backflow in the pipe. In contrast, parameter sets of the original model, lacking boundary conditions downstream, compensate by a high roughness to match the peak, which results in a smoothed signal. This behavior does not represent a physical process: it is a case of overfitting a model subject to structural uncertainty (Kirchner, 2006; Pedersen et al., 2022a). Although both configurations achieve good adherence to observed data (the curves correspond to a Maximum Error metric lower than 0.25 m), Model B appears to perform well but possibly for the wrong reasons: it achieves a good fit, but only by compensating for a structural deficiency, which may undermine the model's reliability under different conditions (Gupta et al., 2012). A sensitivity checks around the initial threshold, restricted to flow conditions between approximately 83% and 100% of full-pipe capacity, yielded consistent results.

To further assess the overfitting of Model B to observed data, the best parameter sets from both the model's configuration (Model B and Model D) was subjected to the untested events (Event 1, Event 5, Event 9). In all cases, Model B sets show a slight overestimation of the peak water levels compared to both observations and Model D results. Additional details are provided in the Supplementary information (S3).

As mentioned in Section 2.1, only the final pipe was monitored during the survey. This limitation prevents a comprehensive assessment of model structural uncertainty and restricts the evaluation to the outlet Out_1. However, this monitoring configuration is common in many urban catchments, where resource constraints often limit measurements to the final section of the drainage network.

3.1.2. Sensitivity analysis

Following the sensitivity analysis, the best parameter sets with a Maximum Error lower than 0.25 m were selected for the four models, to ensure consistency with Section 3.1.1. The Kolmogorov-Smirnov (KS) statistic was computed separately event by event to evaluate the differences between the cumulative distribution functions (CDFs). Pipe and sub-catchment roughness that revealed higher KS, manifested during the analysis of Event 6, were selected for further investigations.

As a result, Fig. 6 presents boxplots of pipe and sub-catchment roughness for Model A-D, based on their best-performing sets. This condition was satisfied by a limited number of parameter sets for Model B and Model D, whereas Model A and Model C exhibited huge parameter sets, also influenced by the fact that multiple events converged into these two classes.

The boxplots for Model A and C cover the same range of values, indicating limited sensitivity to the analysed parameters: the inclusion of boundary conditions when the system is stressed with normal events provides negligible improvements.

On the other hand, Model B and D, developed under extreme conditions, exhibit boxplots that reflect sensitivity to boundary conditions: our efforts to refine the model structure, reflected in Model D, have led to a clear shift in the distribution by preventing overfitting to data (Model B). The roughness parameter values in Model B appear to be biased in order to compensate for structural deficiencies in the model (Refsgaard et al., 2006). As noted by Tscheikner-Gratl et al. (2017), when a model structure lacks a certain component, calibration can often mask these errors at the cost of introducing compensatory "errors" in parameter estimates. A sensitivity check around the initial threshold was conducted, showing stable results regarding the distribution of behavioural parameters.

Our discussion so far has been driven by fortunate circumstances: the availability of hydraulic data collected during extreme conditions, that are however very rare and not always accessible (Kleidorfer et al., 2018) and the location of the measurement point, able to detect the dependencies between the two networks (Freni et al., 2009). This leads us to question whether similar results to those obtained with Model D could still be achieved by including the downstream boundary condition – even in absence of observed extreme events. To answer this question, we considered Model C, which was designed to reflect a scenario where data on extreme event are unavailable.

We generated Model Ensemble C using a Pareto front filtering approach, following the same method applied to generate Model Ensemble A. Six parameter sets were chosen from the Pareto front to represent the widest possible range of parameter variability along the front. Model Ensemble C was then subjected to Event 6: Fig. 7(c) shows the water level trends predicted by Model Ensemble C and Model Ensemble D. The results from Ensemble C exhibit a similar overall response pattern to those from Ensemble D, including the timing of the peak; however, greater variability is observed in the response. This is attributable to the rationale adopted for the models' selection (best parameter sets, wide range over the Pareto front). These results suggest that reliable model performance does not necessarily require the availability of observed rare extreme events.

Model Ensemble C can only exist when a downstream boundary condition is available. In the context of this work, the boundary condition is derived from the use of a full city-scale model; however, it could also be obtained by monitoring a strategically selected point in the network. This underlines the importance of strategic site selection when planning monitoring campaigns in urban drainage systems. Specifically, when the expected use of measurements is model calibration, maximizing the information content can substantially reduce model uncertainties (Huang et al., 2025; Kleidorfer et al., 2012).

Ultimately, we assessed the influence of assigning different boundary condition types at Out_1 – Normal versus Timeseries – on model results. For this purpose, we considered Model Ensemble D stressed by Event 1, 5, 6 and 9 using both boundary condition types. We found that for Events 1, 5, and 9, the maximum difference in water levels at the outlet remains below 0.75% of the maximum pipe height, indicating a negligible impact. In contrast, for Event 6, differences reached up to 15% of the maximum pipe height, highlighting a substantial influence of downstream water levels.

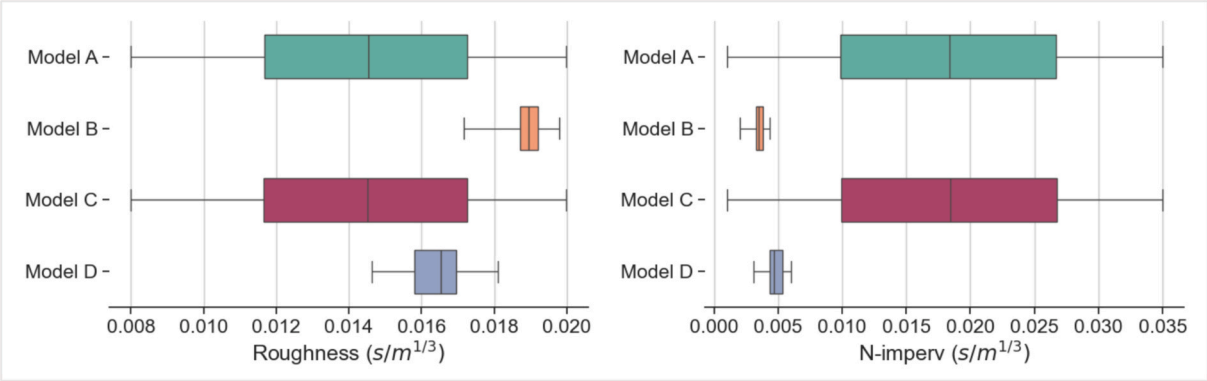

Fig. 6. Boxplots of behavioural parameters: pipe roughness and N-imperv parameters for Model A, B, C, and D based on their best parameters sets.

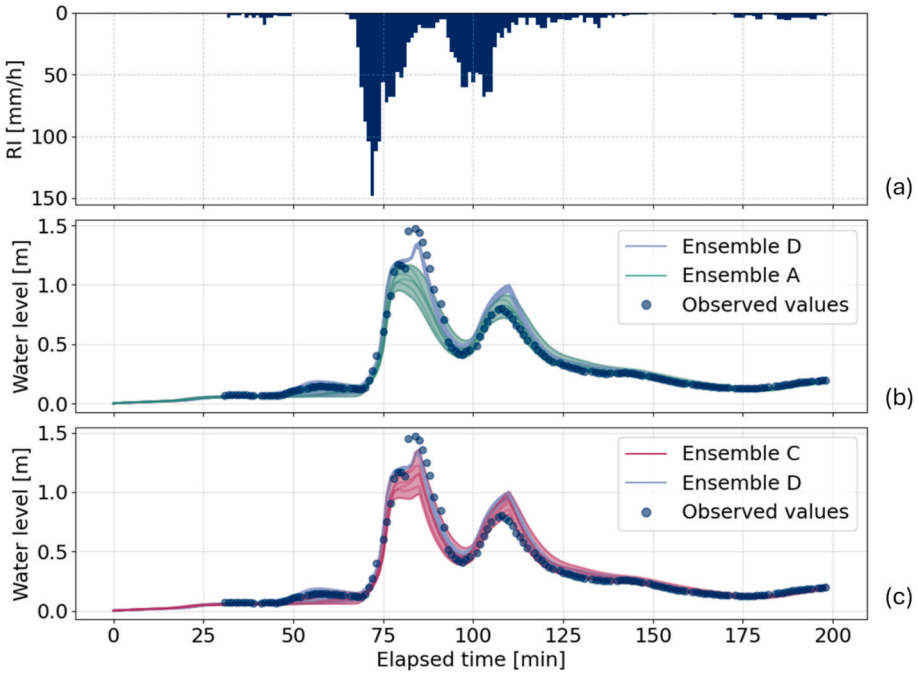

Fig. 7. Simulated and observed water level trends in the final section of Fossolo network for Event 6. (a) Rainfall intensity (mm/h) for Event 6. (b) Water level trend simulated by Ensemble A and D, with observed values. (c) Water level trend simulated by Ensemble C and D, with observed values.

3.1.3. Synthetic storm events based on Event 6

Model Ensemble A and Model Ensemble B were subjected to 1000 synthetic storms that have the same duration and total precipitation depth as Event 6. For each tested storm, a *Maximum Disagreement Matrix* was created (Fig. 8(a)): the matrix is symmetric and has zero values on the diagonal. In the matrix, three zones can be distinguished: Ensemble A vs Ensemble A, Ensemble A vs Ensemble B and Ensemble B vs Ensemble B.

Cumulative distribution functions (CDFs) of maximum disagreement have been calculated for all model combinations: Fig. 8(b) shows the model ensemble results for each group. The narrower green area indicates lower variability among the model output within that group (Ensemble B vs Ensemble B). This can be attributed to the fact that Ensemble B is composed of the best parameter sets of Event 6, whereas models within Ensemble A (blue zone) were selected to represent the wider range of parameter sets along the Pareto front. The separated analysis of intra-model-ensemble disagreement is included in Supplementary Information (S4).

We investigated the inter-model-ensemble groups to identify the storms responsible for the highest disagreements between Ensemble A and Ensemble B. At this scope, for each inter-ensemble group we selected the top 1 percentile (out of 1000 storms) with the highest model disagreement, focusing on storms causing large model ensemble divergences. This resulted in 10 storms per model combination. Since some storms appeared in multiple model combination, the final set included 32 unique storms responsible for the highest model disagreement.

Fig. 9(a) shows the maximum depth of precipitation observed over varying time windows during the storm events, plotted against the duration of the time window. A notable similarity emerges among the filtered storms that induce high model disagreement, highlighted in mustard. Fig. 9(b) shows, for the same filtered storms, the corresponding maximum rainfall intensity. This initial analysis allows to identify the threshold rainfall intensity for which Model Ensemble A and Model Ensemble B start to disagree.

From the filtered events, we selected a subset to stress Model Ensemble D. Model Ensemble D was generated applying the same approach adopted for Model Ensemble B, by selecting the best parameter sets from Model D – maximum error lower than 0.25 m – resulting in an ensemble with a comparable number of parameter sets. The selection

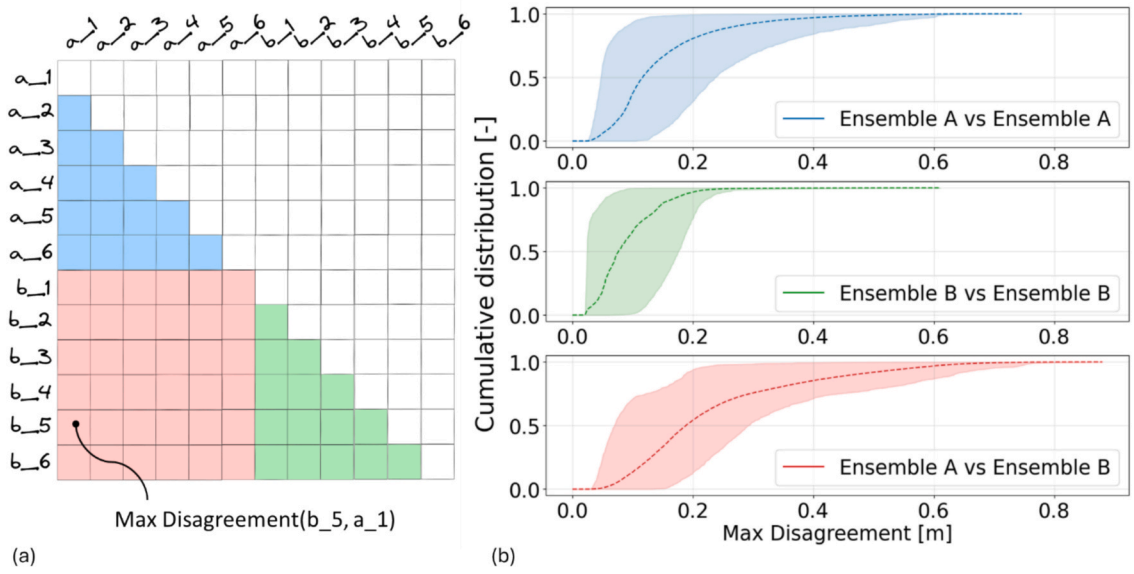

Fig. 8. (a) Maximum Disagreement Matrix. Three zones can be distinguished: Ensemble A vs Ensemble A (blue area), Ensemble A vs Ensemble B (salmon area), Ensemble B vs Ensemble B (green area). (b) CDFs of model disagreement for each group separately. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

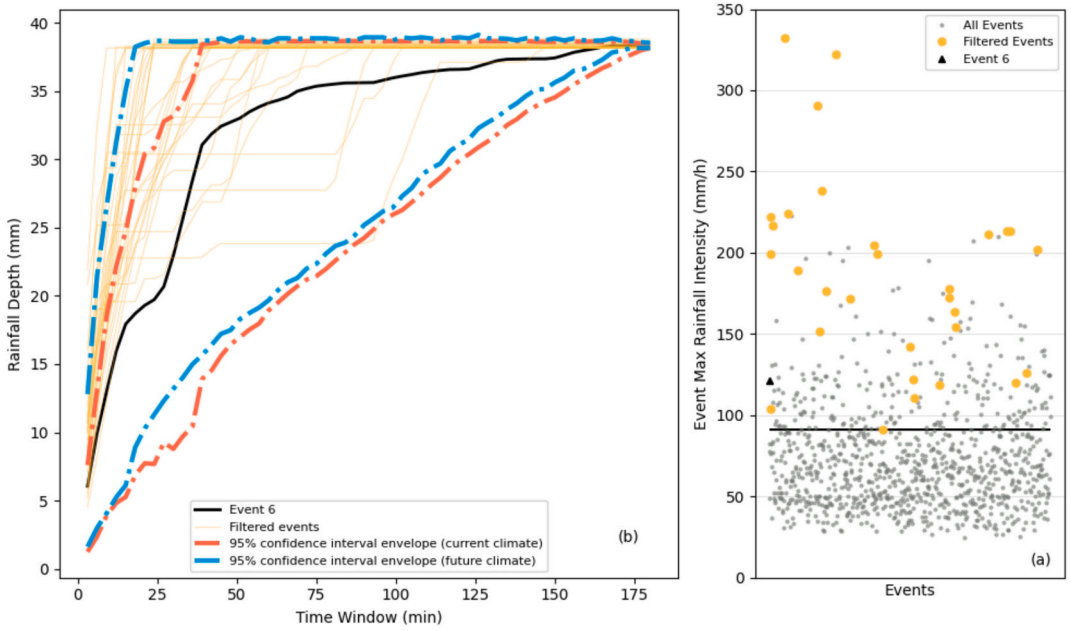

Fig. 9. Model Ensemble A and B stressed by a Storm Ensemble, built with same total precipitation depth and duration of Event 6. (a) Highest aggregate rain depth for different time windows. The values on the y axis show the total amount of rainfall observed within the corresponding time window shown in the x axis. (b) Maximum rainfall intensity observed for each event. Mustard circles show the filtered events. Horizontal line indicates the threshold value.

criterion was designed to identify storms that had an overall impact on the disagreement between ensembles, specifically, storms that triggered disagreement in at least 25% of the model combinations. This choice was also guided by the computational time required to run the full-city-scale model, which provides the downstream boundary condition for Model D. Details about the identified storms triggering the model ensemble disagreement and the number of times it occurred are included in the Supplementary Material (S5.1).

The comparison among Model Ensemble A, B, and D stressed for the same subset of storms (15 storms) is presented in Fig. 10. The figure shows the respective survival distribution of the ratio between water level and maximum pipe depth at the outlet section, with a detailed focus (see inset) on the tail region, which corresponds to higher water level values.

From Fig. 10 the following considerations can be introduced:

- Ensemble A exhibits greater intra-model disagreement, as evidenced in tail region, with an average Kolmogorov-Smirnov (KS) distance of 0.073 between their CDFs (Supplementary Material S5.1). This can be explained by the procedure for selecting parameters set: the parameters set were chosen from different locations on the Pareto front, to represent a large range of model behaviors. The creation of that ensemble aims at covering the range of observed behavior in the different calibration events.

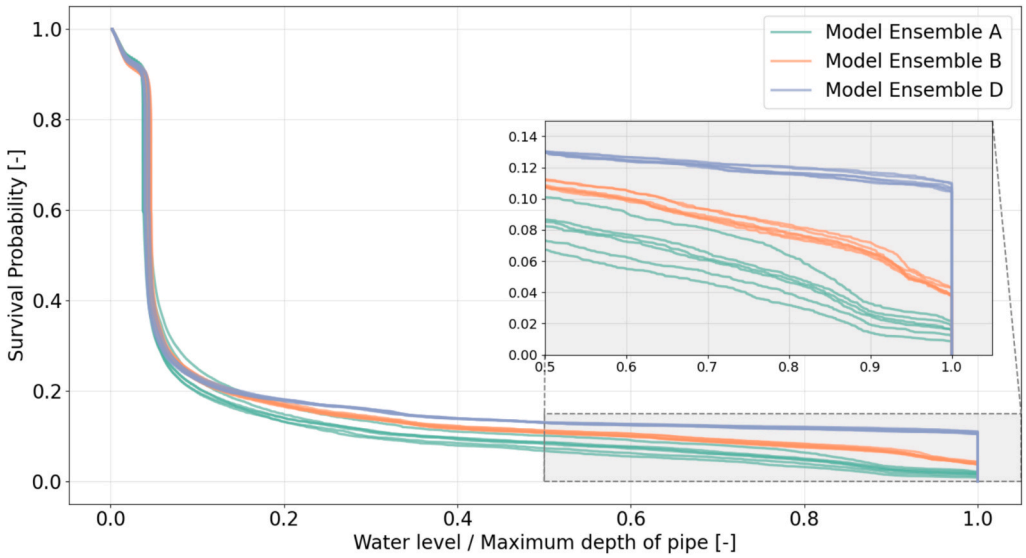

Fig. 10. Survival distribution of dimensionless water level in the final pipe for Model Ensemble A, B and D, stressed by the same subset of ensemble storms, with a focus on the tails (x-value > 0.5) in the figure inset.

- Ensemble D demonstrates strong intra-model agreement (average KS distance equal to 0.024). It also predicts a higher occurrence of full pipe conditions, over 10% of cases, compared with Model Ensemble A and Model Ensemble B, which estimate probabilities of approximately 2% and 4%, respectively.
- Ensemble B also shows good intra-model agreement: in the tail region, the average value of KS distance is equal to 0.033 (Supplementary Material S5.1). While it fails at accurately predicting full pipe conditions, its distributions are closer to those of ensemble D.

When it comes to a modelling choice, modelers may question the relevance or added value of Model Ensemble D, which requires simulations with a full city-scale model, compared with Model Ensemble B, which is faster in terms of computation and may be considered as "close enough" if not examined in detail.

The choice lies in model reliability. Indeed, Ensemble B reaches the full pipe condition in some cases, but possibly for wrong reasons (Kirchner, 2006), meaning that Ensemble B behaves more like a black box model: effective for prediction but limited in explaining what's happening in terms of physical processes. Ensemble D, on the other hand, can be considered an explanatory model: a model aiming at representing the physical processes and suitable for scenario implementation (Knoben and Spieler, 2022). Ultimately, the choice to use a model such as Ensemble B should be left to the modeller. The ensemble may be considered as suitable to predict high flows in the sewer. On the other hand, it may not be suitable to predict full pipe conditions, and full pipe conditions have a different implication than high water level (Bennett et al., 2013).

In the context of our analysis, we explored some of the risks associated with using a predictive model rather than an explanatory one. It appears clear that Model Ensemble D predicted a full pipe condition that can result in high probability of flooding, an outcome that Model Ensemble B fails to reliably reproduce. At the same time, other issues such as inaccurate runoff estimation in subcatchments due to incorrect parameterization, which are not explored in this study, may also occur.

3.2. Impact of the structural source of uncertainty under changing conditions

To estimate the variability in the occurrence of structural uncertainty under future condition, we evaluated the proportion of extreme events triggering the boundary conditions. Model disagreement was assessed by subjecting Model Ensemble A and B to a future ensemble of storm events. Similarly to the methodology implemented in current condition, the Maximum Disagreement Matrix has been built by computing the maximum absolute disagreement for simulated water levels between each combination of model ensembles.

For each inter-ensemble group, we selected the top 1st percentile (out of 1000 storms) with the highest model disagreement, resulting in 10 storms per model combination. Since some storms appeared in multiple model combination, the final set included 49 unique storms responsible for the highest model disagreement. As expected, under the assumption that the network structure remains unchanged in future scenarios, the tipping/threshold rainfall intensity of the filtered storms remains consistent with that observed under current conditions (Supplementary Material, S7). Fig. 11 compares the number of storms that trigger the disagreement between model ensembles for current and future climate conditions and quantifies differences in storm frequency. The analysis is divided into two subsections: the first considers Model Ensemble B and Model Ensemble A excluding a_1 model, while the second focuses on Model Ensemble B and a_1 model.

The positive difference in the number of storms that trigger the disagreement from future to current condition suggests that the risk associated with the use of overfitting models may increase, highlighting the importance of advancing knowledge to ensure that model-based predictions are reliable for the correct reasons (Gupta et al., 2012). Highest disagreement is observed when comparing Model Ensemble B and Model Ensemble A excluding the a_1 model: notably, a_1 shows greater agreement with Ensemble B respect to Ensemble A. This pattern was depicted in the analysis of intra-model-ensemble disagreement within Ensemble A (Supplementary Material, S5).

Focusing on the highest model ensemble disagreement window, a Maximum Disagreement threshold equal to 0.5 m is selected to present the results. The comparison between Model Ensemble B and Model Ensemble A, excluding a_1 model, shows an average increase of 175 storms from current to future conditions, nearly tripling the number of disagreement-triggering events. This pattern is consistent across thresholds close to 0.5 m: within the 0.40–0.60 m range, the increase remains between 2.3 and 3 times the values in current climate. This increase is related to the type of event generated under future conditions. Indeed, through the use of state-of-the-art bias corrected, convection permitting model outputs for Bologna, combined with a statistical downscaling model, results in a higher proportion of events

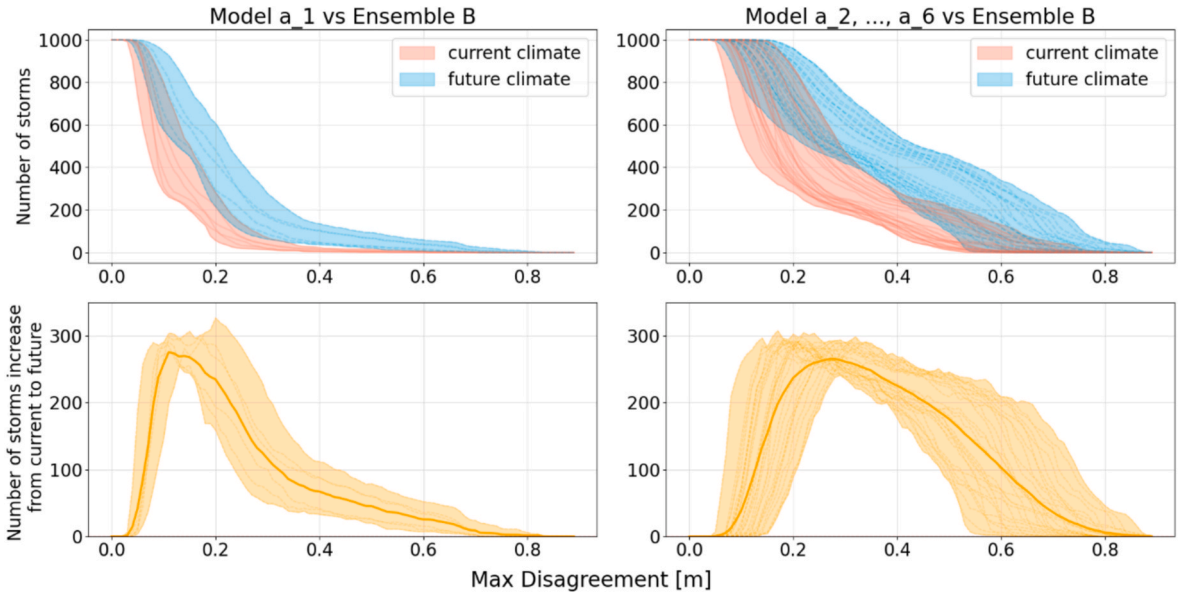

Fig. 11. Separated analysis of Model Ensemble B vs model a_1 (left column) and Model Ensemble B vs Model Ensemble A excluding model a_1 (right column). Upper row: Number of storms exceeding a given threshold of model ensembles disagreement, quantified as the maximum absolute difference in water level in the final section, for future (blue) and current (tomato) climate conditions. Lower row: increase in the number of storms from current to future condition for different threshold of disagreement. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

similar to convective storms under future conditions. These events further trigger the boundary condition of the system. These findings question the validity of urban drainage models under changing conditions. While in the current conditions the proportion of events triggering boundary conditions may be considered acceptable, the substantial increase projected under future conditions suggests a reassessment.

3.3. The use of model disagreement as a measure of structural uncertainty: limitation and potential

The methodology developed to investigate structural uncertainty on urban drainage models, and more specifically to evaluate the adequacy of downstream boundary conditions under extreme events and changing climate conditions, is based on the use of multi-model ensembles combined with a stressor to investigate model disagreement.

After a qualitative analysis on the urban catchment considered and the identification of a key structural hypothesis susceptible to change, we formulated different models with different hypothesis and structures (Model A & C and Model B and D). Such a strategy belongs to model ensembles, and in particular multi-model ensembles. Multi-model ensembles have been used in hydrology and for flood modelling (Cloke and Pappenberger, 2009; Quintana-Romero and Leandro, 2022) but are more commonly used in climate science (Abramowitz et al., 2019). The key idea is that the use of different independent models allows for a better representation of uncertainty, and for instance epistemic uncertainty. The independence of models is often debated since models have similar parametrizations and conceptualizations. In this study, the model structures are purposely similar as we are investigating the impact of a difference in structure. Accordingly, we are not interpreting the multi-model ensemble prediction as a range of uncertainty to account for, but rather as a disagreement between the model's hypotheses. It means that investigation is needed to reduce the range of uncertainty.

The second aspect of our methodology is the use of a stressor to investigate model disagreement. In this study, we consider climate change as a stressor. In particular we investigated the increased frequency of model disagreement resulting from an increase in temperature. This approach does not look at the entire range of impact of climate change, but rather at the shape of hyetograph produced by a temperature driven Multiplicative Random Cascade (MRC) model. As the MRC is a statistical downscaling method, we consider an ensemble of storms to account for the variability of outputs. The MRC model for Bologna was calibrated with observed data and further evaluated against outputs from convection permitting climate models. It showed good agreement between convection permitting and statistical downscaling for precipitation produced under summer temperature (Supplementary Material – S6). Beyond the present case study, we suggest investigating the use of weather generators to stress multi-structure ensemble of urban drainage models. Indeed, climate inputs are one of the main inputs for urban drainage models, consequently the change of climate conditions could trigger model disagreement. For instance, in climates subject occasional drought, soil freezing, or flooding, it may be necessary to investigate how those event occurrences affect modelling results and how these effects may evolve under climate change.

We suggest that the use of such a framework – combining an ensemble of models with different structures to a stressor to trigger tipping points toward model disagreement – could help with accounting for structural uncertainty when developing urban drainage models in other case studies. While structural uncertainty has been discussed in the literature, it may not be yet integrated into practice within the urban drainage community. Our case study shows an example of overlooked structural uncertainty in a model reused multiple times, where the impact of that uncertainty may increase under climate change. We argue that reuse of models i) for purposes that differ from their original developments, and ii) in conditions that differ from those originally considered, should be supported by explicit information on the limits of the model's underlying hypothesis.

4. Conclusions

In this work, we explored the concept of model structure uncertainty through the development of four models with different configurations, each calibrated either by including or excluding extreme events. From these models, we generated model ensembles by adopting two criteria: the best-performing parameter set and parameter sets selected along the Pareto front.

The use of model ensembles proved to be a useful approach for exploring structural sources of uncertainty due to inadequate boundary conditions. We analyzed the responses of Model Ensembles A and B

when subjected to both current and future storm ensembles. Their responses start to disagree beyond a specific model tipping point: below this threshold, associated with the characteristics of rainfall events stressing the models, no tangible disagreement was observed. This divergence served as an alarm for deeper system investigation, which was subsequently conducted using the full city-scale model (Ensemble D).

In this sense, the approach presented in this work challenges the general modelling practice of propagating uncertainty solely through a large number of parameter sets and interpreting uncertainty only as the spread of results they produce. Here, although the ensembles tested are composed of a relatively small number of models, each model is well-understood and selected based on its structural relevance. Their divergence therefore provides targeted insights that can inform specific actions or modelling decisions.

We demonstrated that knowing the model structure prior to parameter optimization is essential to avoid overfitting to data. In our work, "knowing the model structure" is reflected in the role of the downstream boundary condition which, despite appearing as a minor modelling choice, significantly influences the modelling results by altering the parameter sets. Without this awareness, there is a significant risk of developing black-box models. While Model Ensemble B, good for prediction, may not be reliable when subjected to a different context (both current and future storm ensembles), the explanatory model (Model Ensemble D) not only provides accurate predictions but also maintains physical consistency, even when subjected to a different scenario.

As a final remark, the study emphasizes the importance of designing monitoring surveys that enable conscious model understanding and effective calibration. Aligning planning and execution of urban drainage monitoring surveys – including site selection, measurement devices, timeframe of the survey – with the needs of urban drainage modelling allows for the development of reliable models.

CRediT authorship contribution statement

Margherita Evangelisti: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Vincent Pons:** Writing – review & editing, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Spyros Pritsis:** Writing – review & editing, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Vittorio Di Federico:** Writing – review & editing, Funding acquisition. **Franz Tscheikner-Gratl:** Writing – review & editing, Conceptualization. **Marco Maglionico:** Writing – review & editing, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This project has received funding from the European Union's Horizon Europe research and innovation program StopUP (grant 101060428).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2026.135490.

Data availability

Data will be made available on request.

References

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R., Schmidt, G.A., 2019. ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing. Earth Syst. Dyn. 10, 91–105. https://doi.org/10.5194/esd-10-91-2019.

Artina, S., Maglionico, M., Marinelli, A., Raffaelli, G., Anzalone, C., Lanzarini, S., Guzzinati, E., 1997. Le misure di qualità nel bacino urbano Fossolo. L'acqua.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. Environ. Model. Software 40, 1–20. https://doi.org/10.1016/j.envsoft.2012.09.011.

Broekhuizen, I., Muthanna, T.M., Leonhardt, G., Viklander, M., 2019. Urban drainage models for green areas: Structural differences and their effects on simulated runoff. J. Hydrol. X 5, 100044. https://doi.org/10.1016/j.hydroa.2019.100044.

Caillaud, C., Somot, S., Alias, A., Bernard-Bouissières, I., Fumière, Q., Laurantin, O., Seity, Y., Ducrocq, V., 2021. Modelling Mediterranean heavy precipitation events at climate scale: an object-oriented evaluation of the CNRM-AROME convection-permitting regional climate model. Clim. Dyn. 56, 1717–1752. https://doi.org/10.1007/s00382-020-05558-y.

Chrysochoidis, V., Gruber, G., Hofer, T., Mikkelsen, P.S., Vezzaro, L., 2025. Rethinking modelling of particulate pollutants in combined sewer overflows (CSOs): a focus on model structure. J. Hydrol. 659, 133239. https://doi.org/10.1016/j.jhydrol.2025.133239.

Cloke, H.L., Pappenberger, F., 2009. Ensemble flood forecasting: a review. J. Hydrol. 375, 613–626. https://doi.org/10.1016/j.jhydrol.2009.06.005.

De Paola, F., Giugni, M., Pugliese, F., 2018. A harmony-based calibration tool for urban drainage systems. Proceed. Instit. Civil Eng. - Water Manage. 171, 30–41. https://doi.org/10.1680/jwama.16.00057.

Del Giudice, D., Reichert, P., Bareš, V., Albert, C., Rieckermann, J., 2015. Model bias and complexity – Understanding the effects of structural deficits and input errors on runoff predictions. Environ. Model. Software 64, 205–214. https://doi.org/10.1016/j.envsoft.2014.11.006.

Deletic, A., Dotto, C.B.S., McCarthy, D.T., Kleidorfer, M., Freni, G., Mannina, G., Uhl, M., Henrichs, M., Fletcher, T.D., Rauch, W., Bertrand-Krajewski, J.L., Tait, S., 2012. Assessing uncertainties in urban drainage models. Phys. Chem. Earth, Parts a/b/c 42–44, 3–10. https://doi.org/10.1016/j.pce.2011.04.007.

Freni, G., Mannina, G., 2010. Bayesian approach for uncertainty quantification in water quality modelling: the influence of prior distribution. J. Hydrol. 392, 31–39. https://doi.org/10.1016/j.jhydrol.2010.07.043.

Freni, G., Mannina, G., Viviani, G., 2009. Assessment of data availability influence on integrated urban drainage modelling uncertainty. Environ. Model. Software 24, 1171–1181. https://doi.org/10.1016/j.envsoft.2009.03.007.

Freni, G., Sambito, M., Piazza, S., 2024. Bayesian Model Averaging Approach for Urban Drainage Water Quality Modelling. In: Gourbesville, P., Caignaert, G. (Eds.), Advances in Hydroinformatics—SimHydro 2023, Volume 2. Springer Nature Singapore, Singapore, pp. 217–228.

Gupta, H.V., Clark, M.P., Vrugt, J.A., Abramowitz, G., Ye, M., 2012. Towards a comprehensive assessment of model structural adequacy. Water Resour. Res. 48, 2011WR011044. https://doi.org/10.1029/2011WR011044.

Höge, M., Guthke, A., Nowak, W., 2019. The hydrologist's guide to Bayesian model selection, averaging and combination. J. Hydrol. 572, 96–107. https://doi.org/10.1016/j.jhydrol.2019.01.072.

Huang, Y., Xiong, J., Zhang, J., Zheng, F., Ji, Y., Gupta, H., 2025. A robust method to simultaneously place sensors and calibrate parameters for urban drainage pipe system models using Bayesian decision theory. Eng. Appl. Comput. Fluid Mech. 19, 2473992. https://doi.org/10.1080/19942060.2025.2473992.

Jakeman, A.J., Letcher, R.A., Norton, J.P., 2006. Ten iterative steps in development and evaluation of environmental models. Environ. Model. Software 21, 602–614. https://doi.org/10.1016/j.envsoft.2006.01.004.

Kirchner, J.W., 2006. Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology. Water Resour. Res. 42. https://doi.org/10.1029/2005wr004362.

Kleidorfer, M., Leonhardt, G., Rauch, W., 2012. Identifiability analysis in conceptual sewer modelling. Water Sci. Technol. 66, 1467–1474. https://doi.org/10.2166/wst.2012.330.

Kleidorfer, M., Tscheikner-Gratl, F., Vonach, T., Rauch, W., 2018. What can we learn from a 500-year event? experiences from urban drainage in Austria. Water Sci. Technol. 77, 2146–2154. https://doi.org/10.2166/wst.2018.138.

Knoben, W.J.M., Spieler, D., 2022. Teaching hydrological modelling: illustrating model structure uncertainty with a ready-to-use computational exercise. Hydrol. Earth Syst. Sci. 26, 3299–3314. https://doi.org/10.5194/hess-26-3299-2022.

Kreikenbaum, S., Krejci, V., Fankhauser, R., Rauch, W., 2004. Berücksichtigung von Unsicherheiten in der Planung. GWA Gas, Wasser, Abwasser 84, 587–594.

Leimgruber, J., Steffelbauer, D.B., Krebs, G., Tscheikner-Gratl, F., Muschalla, D., 2018. Selecting a series of storm events for a model-based assessment of combined sewer overflows. Urban Water J. 15, 453–460. https://doi.org/10.1080/1573062X.2018.1508601.

Mannina, G., Viviani, G., 2010. An urban drainage stormwater quality model: model development and uncertainty quantification. J. Hydrol. 381, 248–265. https://doi.org/10.1016/j.jhydrol.2009.11.047.

Marinelli, A., Maglionico, M., Artina, S., 1997. Water quality simulation in an urban drainage catchment. In: Presented at the Proceedings of the European Water Resources Association Conference, pp. 383–390.

Pedersen, A.N., Brink-Kjær, A., Mikkelsen, P.S., 2022a. All models are wrong, but are they useful? Assessing reliability across multiple sites to build trust in urban drainage modelling. Hydrol. Earth Syst. Sci. 26, 5879–5898. https://doi.org/10.5194/hess-26-5879-2022.

Pedersen, A.N., Pedersen, J.W., Borup, M., Brink-Kjær, A., Christiansen, L.E., Mikkelsen, P.S., 2022b. Using multi-event hydrologic and hydraulic signatures from water level sensors to diagnose locations of uncertainty in integrated urban drainage models used in living digital twins. Water Sci. Technol. 85, 1981–1998. https://doi.org/10.2166/wst.2022.059.

Pichler, M., 2022. swmm-api: API for reading, manipulating and running SWMM-Projects with python. https://doi.org/10.5281/zenodo.7054804.

Pons, V., Benestad, R., Sivertsen, E., Muthanna, T.M., Bertrand-Krajewski, J.-L., 2022. Forecasting green roof detention performance by temporal downscaling of precipitation time-series projections. Hydrol. Earth Syst. Sci. 26, 2855–2874. https://doi.org/10.5194/hess-26-2855-2022.

Pritsis, S., Pons, V., Rokstad, M.M., Clemens-Meyer, F.H.L.R., Kleidorfer, M., Tscheikner-Gratl, F., 2024. The role of hyetograph shape and designer subjectivity in the design of an urban drainage system. Water Sci. Technol. 90, 920–934. https://doi.org/10.2166/wst.2024.261.

Quintana-Romero, T., Leandro, J., 2022. A method to devise multiple model structures for urban flood inundation uncertainty. J. Hydrol. 604, 127246. https://doi.org/10.1016/j.jhydrol.2021.127246.

Raffa, M., Adinolfi, M., Reder, A., Marras, G.F., Mancini, M., Scipione, G., Santini, M., Mercogliano, P., 2023. Very High Resolution Projections over Italy under different CMIP5 IPCC scenarios. Sci. Data 10, 238. https://doi.org/10.1038/s41597-023-02144-9.

Refsgaard, J.C., van der Sluijs, J.P., Brown, J., van der Keur, P., 2006. A framework for dealing with uncertainty due to model structure error. Adv. Water Resour. 29, 1586–1597. https://doi.org/10.1016/j.advwatres.2005.11.013.

Reichert, P., 2012. Conceptual and Practical Aspects of Quantifying Uncertainty in Environmental Modelling and Decision Support. International Congress on Environmental Modelling and Software. 276.

Rieckermann, J., 2016. There is nothing as practical as a good assessment of uncertainty. QUICS Blog. URL https://quicsblog.wordpress.com/2016/12/22/there-is-nothing-as-practical-as-a-good-assessment-of-uncertainty/ (accessed 7.3.25).

Rossman, L., 2010. Storm Water Management Model User's Manual Version 5.0. US Environmental Protection Agency.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., Tarantola, S., 2010. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. Comput. Phys. Commun. 181, 259–270. https://doi.org/10.1016/j.cpc.2009.09.018.

Tscheikner-Gratl, F., Lepot, M., Moreno-Rodenas, A., Schellart, A., 2017. QUICS D.6.7 - A Framework for the application of uncertainty analysis. https://doi.org/10.5281/zenodo.1240926.

Van Der Keur, P., Henriksen, H.J., Refsgaard, J.C., Brugnach, M., Pahl-Wostl, C., Dewulf, A., Buiteveld, H., 2008. Identification of Major sources of uncertainty in Current IWRM Practice. Illustrated for the Rhine Basin. Water Resour. Manag. 22, 1677–1708. https://doi.org/10.1007/s11269-008-9248-6.

Walker, W.E., Harremoës, P., Rotmans, J., van der Sluijs, J.P., van Asselt, M.B.A., Janssen, P., Krayer von Krauss, M.P., 2003. Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. Integr. Assess. 4, 5–17. https://doi.org/10.1076/iaij.4.1.5.16466.