https://doi.org/10.5194/egusphere-2022-615
Preprint. Discussion started: 25 July 2022

All models are wrong, but are they useful? Assessing reliability across multiple sites to build trust in urban drainage modelling

Agnethe Nedergaard Pedersen[1,2], Annette Brink-Kjær[1], and Peter Steen Mikkelsen[2]

[1] VCS Denmark, Vandværksvej 7, 5000 Odense C, Denmark
[2] DTU Sustain, Technical University of Denmark, Bygningstorvet, Bygning 115, 2800 Kgs. Lyngby, Denmark

Correspondence to: Agnethe Nedergaard Pedersen (anp@vandcenter.dk)

Abstract. Simulation models are widely used in urban drainage engineering and research, but they are known to include errors and uncertainties that are not yet fully realised. Within the herein developed framework, we investigate model adequacy across multiple sites by comparing model results with measurements for three model objectives: 'surcharges' (water level rises above defined critical levels related to basement flooding), 'overflows' (water levels rise above a crest level), and 'everyday events' (water levels stay below the top of pipes). We use multi-event hydrological signatures, i.e. metrics that extract specific characteristics of time series events in order to compare model results with the observations for the mentioned objectives through categorical and statistical data analyses. Furthermore, we assess the events with respect to sufficient or insufficient categorical performance, and good, acceptable, or poor statistical performance. We also develop a method to reduce the weighting of individual events in the analyses, in order to acknowledge uncertainty in model and/or measurements in cases where the model is not expected to fully replicate the measurements. A case study including several years of water level measurements from 23 sites in two different areas shows that only few sites score as "sufficient categorical performance" in relation to the objective 'overflow', and that sites do not necessarily obtain good performance scores for all the analysed objectives. The developed framework however highlights that it is possible to identify objectives and sites for which the model is reliable, and we also suggest methods for assessing where the model is less reliable and needs further improvement, which may be further refined in the future.

1 Introduction

Danish utility companies invest 800 million EUR annually in upgrading and rehabilitating urban drainage systems and 400 million EUR annually in operation of the existing systems (DANVA, 2021), which corresponds to 150 and 75 EUR per capita annually. The investments often rely on simulations with physics-based deterministic models, which are widely used in the urban drainage practice community for several decades. The model simulations are applied for many different purposes including e.g.: prioritising areas for redesign and optimization, making comparative assessments of optimal designs, comparing with measurements on a regular basis, and as important features in digital twins (Pedersen et al., 2021a). The models tends to gradually become more complex and include increasing levels of detail, and the model software to be equipped with professional output presentation interfaces, and thus expectations to the applicability of models from stakeholders such as municipal regulators and utilities are increasing (e.g. Fenicia and Kavetski, 2021). But still however, the uncertainty of the model results is (practically) not exhibited with the models. George Box once said: “All models are wrong, but some are useful” (Box, 1979). With increasing expectations to the models, the question to ask back is: “How useful are the models then?” The general aim of this paper is thus to explore systematic methods that can potentially be automated, for evaluating site-specific performance of large, detailed, distributed urban drainage models across a range of different model objectives.

In the future, digital twins of urban drainage systems are expected to be a part of the toolbox in many utility companies (SWAN, 2022). The model performance assessment tool developed in this paper applies to a living operational digital twin (Pedersen et al., 2021a), and is used to assess whether the model used within the digital twin is suitable for replicating certain events/situations. A living operational digital twin is a virtual copy of the current physical system, and can be held up and compared with reality, as measured through sensor observations from the system. Learnings from the evaluation of the operation model in relation to the real-world observations can be transferred to other model types, such as planning or design models, and thereby improve the basis for decision-making regarding future investments.

Several studies in the hydrology and environmental modelling community have highlighted and discussed model performance in relation to diagnosing of model fidelity (Gupta et al., 2014). This applies e.g. in relation to scientific research reporting (Fenicia and Kavetski, 2021) and in relation to choosing the best model parameters from several sets of parameters based on Principal Component Analysis (PCA) (e.g. Euser et al. (2013)). Determining a best model parameter set can be accommodated by e.g. post-processing of errors from historical data (Ehlers et al., 2019) or signature-based evaluation (Gupta et al., 2008). Research has focused on quantifying the uncertainty in model output, e.g. by using the Generalized Likelihood Uncertainty Estimator (GLUE) method (Beven and Binley, 1992). Developing error models that compensate for the lacking adequacy may provide better model results in the short term and for certain purposes, but the opportunity to detect the actual source of the underlying errors that dominate a model may be missed (Gupta et al., 2014, 2008; Pedersen et al., 2022).

The assessment of model adequacy relies on observations from the system. Monitoring is not always easy, as urban drainage systems are rough environments, e.g. with particles that can settle and clog equipment, and flow meters that have difficulties measuring flow correctly when the pipe is partly filled. Water level meters are acknowledged to be fairly accurate, but they can suffer from missing data values, which can be tackled using several methods, as described in Clemens-Meyer et al. (2021). In the present paper, model evaluation is conducted for sites with only level meters installed, because in the future low-cost level meters are expected to be installed at an increasing number of sites in urban drainage systems (Eggimann et al., 2017; Kerkez et al., 2016; Shi et al., 2021). This will provide an opportunity for using all of these observations in the model evaluation, provided that the model evaluation methodology is improved and more structured than today in hydrology (Gupta et al., 2012) and urban drainage.

In this paper, up to 11 years of level-measurements from 23 measurement sites located in two case areas operated by VCS Denmark (Danish utility company) are used to demonstrate several site-specific model evaluation methods. Every site can provide information about the model performance. However, manual inspection of the observations is practically impossible (too labour intensive), and the utility company therefore needs automatic calculation of the model performance (through comparison of model results with measurements) for specific conditions at each site. The utility company aims for such analyses to provide a geographical overview of the model performance across several model objectives, which can in the future be applied as an automatic and scalable tool across hundreds of measurement sites to help prioritise information and determine where and when further investigations are needed for error diagnosis of the models.

Determining whether a modelled time series is replicable, i.e. consistent with observations in real life, can be done by applying hydrologic signatures. Signatures are metrics that extract certain characteristics from time series of hydrological events, such as the peak level, or the duration of water level above a given level. Signatures have been applied in general hydrology for many years, and many different signatures (primarily based on flow as the measured variable) have been developed (McMillan, 2020a, b). Applying signatures from multiple events based on time series of measured water level have recently proven to be a promising tool for diagnosing errors in urban drainage models (Pedersen et al., 2022). By combining signatures with other variables characterising the direct (in-sewer) or indirect environment ("surrounding states"), tendencies can potentially be detected. With many sites and many signatures to analyse, it is easy to lose one's bearings, and an assessment of the model adequacy is therefore needed to (1) get an overview of model performance, and (2) prioritise where model diagnosing efforts should be placed – for different model objectives and for different measurement sites.

This paper goes in depth with how model performance can be assessed for large, detailed, distributed urban drainage models across a range of different model objectives. However, the recognition of 'acceptable' model and data uncertainties needs to be addressed. A model performance assessment should not be affected by events that lie way off target in relation to the known modelling capabilities and limitations. The urban drainage community has only recently started developing tools for anomaly detection in the urban drainage system caused by physical events, e.g. pipe blockage, pumps that did not start as planned, and other sudden activities in the system (Palmitessa et al., 2021; Clemens-Meyer et al., 2021). Uncertainty in input data, such as unrealistic representation of rain events due to spatial variability, has not to the authors awareness been investigated. This paper thus proposes a method for identifying events that models are not expected to replicate and that, therefore, should have less weight or be excluded from the model evaluation.

The investigated methods are explained in Section 2, including an overall framework for model adequacy assessment, three model objectives studied in detail ('surcharge', 'overflow' and 'everyday' events), the context definition (including signature definitions and method for weighting of individual events), the categorical and statistical methods used, and the overall criteria

https://doi.org/10.5194/egusphere-2022-615
Preprint. Discussion started: 25 July 2022

employed for assessing overall model performance. Section 3 describes the study area, model and data used, Section 4 presents and discusses our results, and Section 5 summarises our conclusions.

2 Methods

2.1 Framework for model adequacy assessment

To assess model performance, we suggest following five steps (Figure 1). A short introduction to each step will be provided here, and later subsections will go into details. The first step is to identify the overall model objective. We need to start by determining which objective it is that we wish to assess based on the model simulations, is it e.g. the model's ability to replicate an overflow or are we maybe more interested in everyday rain events? The next step is to establish the context, based on the model objective. Signatures related to overflows are not relevant when looking at everyday rain events, and vice versa. Categorical analysis is conducted as the next step, and here the events are categorised in accordance with the chosen objectives. For instance, if modelling of overflows is the objective, then we can identify if the modelled and observed water level of the event raises above a defined threshold, i.e. the crest level of an overflow structure. The true positives are the events where both model results and observations occur in the same category, and a statistical analysis of the true positive events can be carried out in order to assess whether they perform well. Finally, an assessment is made as to which 'traffic-light' categories the model belongs for each site. This procedure does not serve to fix anything, but solely to indicate how well the model is able to replicate the defined objectives.

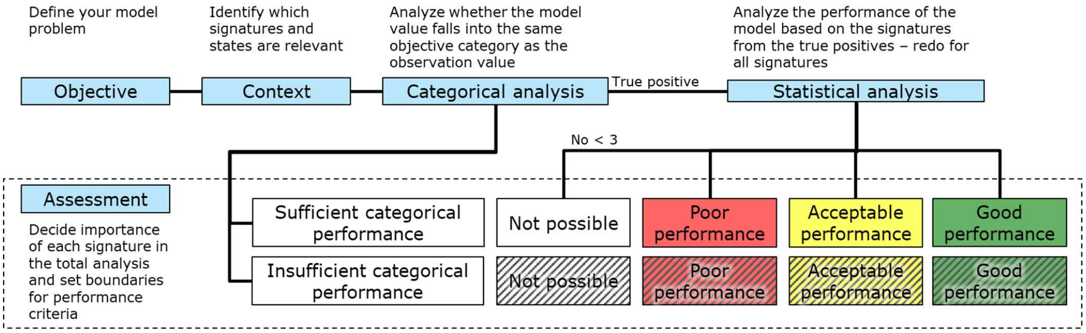

Define your model problem
Identify which signatures and states are relevant
Analyze whether the model value falls into the same objective category as the observation value
Analyze the performance of the model based on the signatures from the true positives – redo for all signatures
Objective
Context
Categorical analysis
True positive
Statistical analysis
No < 3
Assessment
Decide importance of each signature in the total analysis and set boundaries for performance criteria
Sufficient categorical performance
Insufficient categorical performance
Not possible
Poor performance
Acceptable performance
Good performance
Not possible
Poor performance
Acceptable performance
Good performance

Figure 1. Identified steps (blue boxes) for assessing model performance.

2.2 Model objectives

The utility company defined the objectives based on their interest in model output. The reliability of the operation model is analysed for three objectives in this paper: surcharges (water level rises above defined critical levels related to basement flooding), overflows (water levels rise above an overflow weir crest level), and everyday rain events (water levels stay below the top of pass-forward pipe). These objectives are especially important, as model results are used to support decision-making in relation to future investments. If the model does not mirror physical behaviour adequately, investments can made based on false premises - potentially leading to human injury or environmental damage that could have been prevented.

A more in-depth description of the objectives that will be investigated in the paper is given below (cf. the illustration in Figure 2):

Surcharges: Situations where peak water levels in manholes rise above defined critical surcharge levels (CSL), In VCS the CSL is generally set at 1.5 m below ground level to mirror the typical level of basements, unless the crest level (CL) or top-of- pipe (TOP) is within the range of 1.5 m. The CSL must not be exceeded more often than every second year, to provide the optimal service level as required by the utility company (Odense Kommune, 2011).

Overflow: Situations where the water level rises above the crest level and overflows occur. The overflow criterion only applies to sites equipped with an overflow weir (either internal or external).

Everyday rain: Situations where water levels are below the top of the pass-forward pipe (TOP) or a crest level (CL) if this is lower. Such events occur for minor rain events that do not lead to exceedance of the pipe capacity, overflow or surcharge, but that occur quite often. In Denmark the design of a full-running pipe has to include a rain with a 2-year return period. The terminology relating to everyday rain is adopted from Sørup et al. (2016).

The above definition of model objectives means that the water levels in the range between the critical surcharge level (CSL)/overflow crest level (CL) and the top of the pipe (TOP) are not included in the analyses. This is intentional, because water levels in this range are often very dynamic due to the limited volume available in the manholes, and model-to-observation fits are thus expected to be poor in this range.

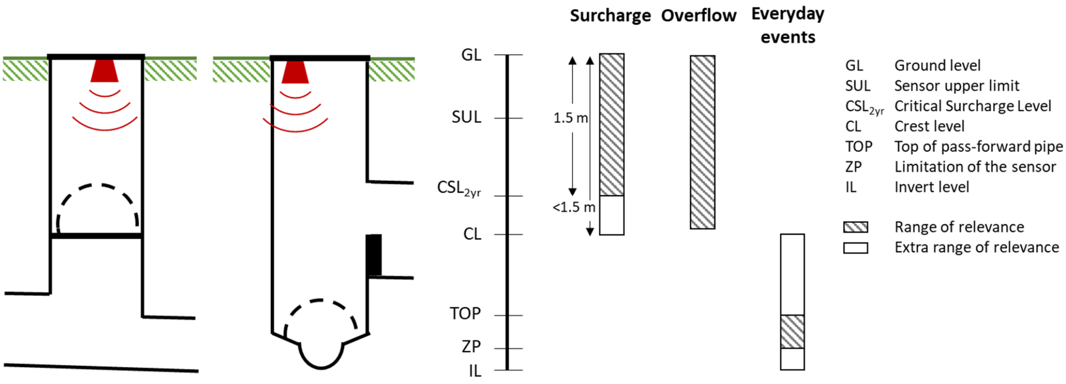


Figure 2. Illustration of the three model objectives in relation to water levels in the system. Overflow occurs for events where the water level rises the crest level (CL). Surcharge applied to events where the water level rises above the critical surcharge level (CSL), which is defined as 1.5 m below ground level (GL) - or, if crest level (CL) or top-of-pipe (TOP) is within the range of 1.5 m, then that level will be the CSL. Everyday events are when the water levels observed or modelled is below the top of the pipe (or CL if this level is lower) and above the invert level (IL) (or zero-point (ZP) if this level is higher).

2.3 Context definition

Methods for handling unrealistic anomalies can be found in literature (e.g. Clemens-Meyer et al., 2021). For this study the simple data cleaning approach by Pedersen et al. (2021b) was applied, using five techniques (low data quality determined by the SCADA system manufacturer, manually removed data, out of physical bounds considering the specific sensor, frozen sensor signal, and outlier data as assessed by an operator). Erroneous observation data were interpolated up to 5 min and otherwise replaced with NaN (not-a-number) values.

Since the three objectives are related to rain induced events and not dry weather conditions, a time-varying event definition was applied for the time series with a focus on water levels that are above the water level variations on a normal dry weather day influenced by infiltration-inflow (Pedersen et al., 2022). These events were found for the observed and modelled time series, separately and jointly; the joint events were applied in this study.

2.3.1 Signatures

Signatures are metrics that extract specific characteristics of a time series event, as illustrated for water levels in Figure 3. Peak level, duration and Area Under Curve (AUC) were previously described in Pedersen et al. (2022). Where the two first are standard parameters used in common practice, the third (AUC) calculates a ‘surrogate volume’ for an event from a reference level (similar to the area under a flow hydrograph, but with a different unit). Relevant signatures must be selected for each objective, in order to evaluate the model performance. Selection of signatures (Table 1) was here based on an assessment conducted by the author group, including discussion of this topic with a group of utility experts in hydraulic modelling. The relevant signatures for the objective “surcharge” are: peak level, duration above the CSL and AUC above CSL. Signatures relevant for the objective “overflow” are: duration above CL and AUC above CL. Peak level is not of interest for the overflow objective, and therefore it is not included. The relevant signatures for the objective “everyday events” are: peak level, number of peaks, the AUC between the zero-point (ZP)/invert level (IL) and the TOP/CL (the range of relevance, see Figure 2) and the maximum level rate of change (5 min smoothing window). These were chosen based on an assessment that they will provide valuable insights into the everyday event. Further analyses could have been conducted to support the relevance, but that lies outside the scope of this paper.

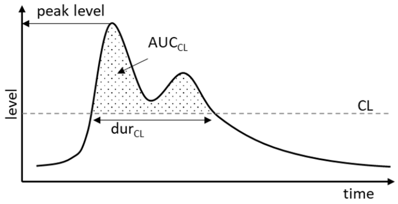


Figure 3. Simple representation of signatures.

Table 1. Relevant signatures included for the analysis of three defined model objectives.

Signature name	Description
Surcharge	
Peak level	The peak level of the event
Duration above CSL	Duration of time that the level is above the CSL
AUC above CSL	Area Under Curve, calculated with reference to the CSL
Overflow	
Duration above CL	Duration of time that the level is above the CL
AUC above CL	Area Under Curve, calculated with reference to the CL
Everyday event	
Peak level	The peak level of the event
No. of peaks	Number of local peaks identified. The time series is converted to a rolling median of 5 min, and peaks are identified by applying the scipy-code ‘find_peaks’ with a prominence of 3 cm and a width of 2 minutes (Virtanen et al., 2020).
AUC between ZP/IL – TOP/CL	Area Under Curve, calculated with lower threshold (maximum level of either ZP or IL) and upper threshold (minimum level of either TOP or CL)
Max level rate of change (5 min resolution)	The maximum level rate of change within a rolling window of 5 min duration.

2.3.2 Weighting of individual event in the categorical and statistical analyses

A model should, in theory, be able to replicate all measured events. However, in practice, this is often not the case, because uncertainties can potentially be identified in several model locations (including the model context, input, structure and parameters (Figure 4)), and because the system and the sensors can exhibit abnormal behaviour not included in the model setup. These events, where anomalies occur, constitute an outlier in relation to a model's performance, as the model is not made to handle these issues. Referring to Figure 4, this would indicate a location of uncertainty in the context area, as the model does not have the primary focus to handle all situations. It is also generally accepted that the model output is very sensitive to the input of rain, in terms of spatial coverage. When an intense storm event only affects a limited physical area, where only few rain gauges are monitoring the rainfall, the corresponding rainfall input may not be correct, generating high degrees of uncertainties in the model. This uncertainty may be reduced with rain radar input, but this possibility has not been investigated in this study. The same occurs for the sensors. If the sensor's range is shorter than the water level that it is measuring, this may affect the time series of the observations, but not necessarily all the signatures would be affected. An upper limit of the sensors may furthermore be exceeded by the water level, and the peak level will reach a limitation, and thereby be wrong. However, the signature duration above CL may not be wrong, if only the sensor is placed above the crest level.

In this study a method was developed to reduce the weighting of individual events in the categorical and statistical analyses, in order to acknowledge uncertainty in model and/or measurements in cases where the model is not expected to fully replicate the measurements. The weights for each event were, for this analysis, calculated based on the following rules.

- Events, where the peak level has reached the upper sensor limitation were given a weight of zero for the signatures: peak level, AUC above CL and AUC above CSL.
- Events, where there is a known system anomaly were given a weight of zero for all signatures. Known system anomalies were identified based on manual inspection of the outlier events.
- Events, where the rain input uncertainty is particularly high, quantified by the coefficient of variation (CV) of the rain depth of the rain gauges within a 5 km surrounding (Pedersen et al., 2022), were given a weight from zero to one. The weight w was calculated as $w = 1\text{-}CV$. For CV>0.5, the weight was calculated as $w = 0.5*(1\text{-}CV)$, and for CV>1, the weight was set to 0.

The weights were the same for the joint events, i.e. combining of both modelled and observed input.

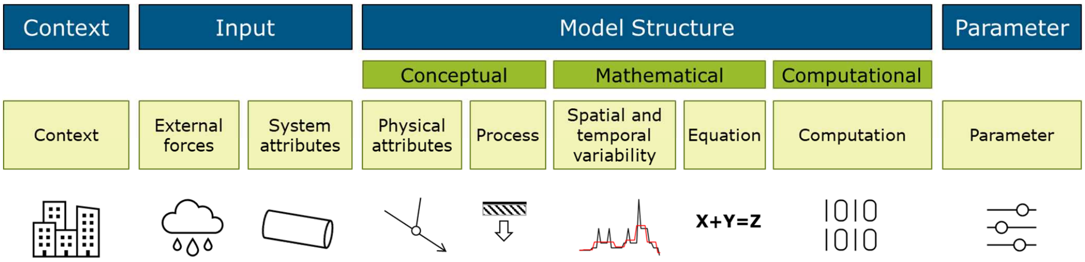


Figure 4. Illustration of the location of uncertainties in models. Adapted from a table in Pedersen et al. (2022)**.**

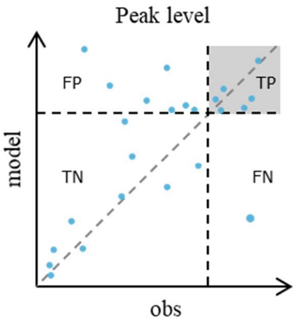
2.4 Categorical analysis

Analysing model performance for specific objectives suggests that events can fall into different categories (e.g. overflow or no overflow). For relevant signatures, the categorical analysis aims to identify for each event if observed and modelled results are above or below a given threshold, e.g. the crest level for the objective 'overflow' (Figure 5). If both a modelled and observed event is above the threshold, the event is categorised as a true positive (TP). If the number of TPs, relative to the false positives (FP - modelled, but not observed threshold exceedance) and false negatives (FN - not modelled, but observed threshold exceedance), is too low, then the model simulation of the events is not correct, categorically speaking, and the confidence, or trust in the model is low.

Several metrics can be applied to assess the categorical performance of the model, however for this analysis, where we are dealing with rare events, the metrics should not include the true negatives (TN). The metrics chosen is thus the Critical Success Index (CSI) (Bennett et al., 2013), which takes the true positives (TP) compared to all observed or modelled positives (TP, FP and FN).

$$CSI = \frac{TP}{TP+FP+FN} \tag{1}$$

For this categorical analysis, the weights introduced above were considered, so that events with *w*<0.5 were disregarded.



		Observation above threshold	
		No	Yes
Model above threshold	Yes	Σ False Positive (FP)	Σ True Positive (TP)
	No	Σ True Negative (TN)	Σ False Negative (FN)

Figure 5. Contingency tables (right) are used to categorise each event according to whether it is above or below a given threshold. The grey dashed line (left) is the 1:1 line, and the blue points are the signatures for each event (modelled vs. observed value).

2.5 Statistical analysis

The events categorised as true positives can be assessed statistically. Scatterplots of multi-event signature comparisons can be made with observation signature values for each event on the horizontal axis and modelled values on the vertical axis. An 1:1 line indicates the perfect model-to-observation fit (Figure 6), and three different ways of analysing the scatterplots were assessed in this paper: linear regression (Figure 6, left), an indicator function (Figure 6, middle) and the normalised root-mean-square-errors (RMSE) method (Figure 6, right).

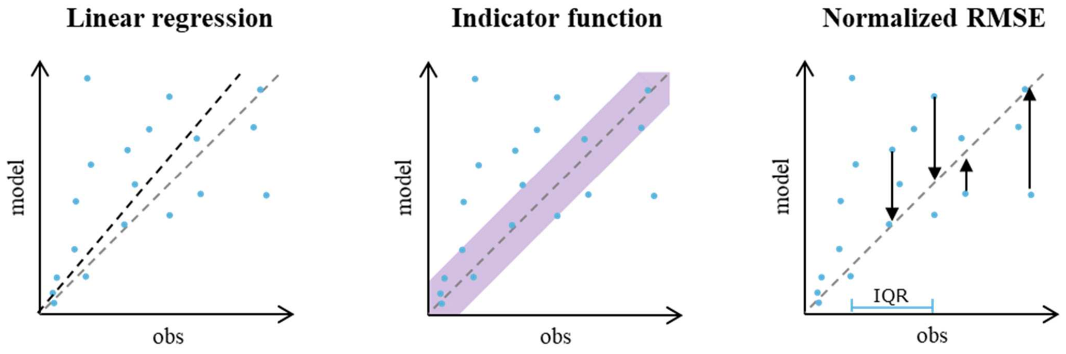


Figure 6. Three methods of statistical analyses of the true positive events, where both simulated and observed values are in the same category: linear regression (left), an indicator function (middle) and normalised root-mean-square-errors (RMSE) (right). The grey dashed lines are the 1:1 lines, and the blue points are the signatures for each event (modelled vs. observed values).

2.5.1 Linear regression

Linear regression is a simple statistical method to assess whether there is a correlation between two variables, in this case: observed and modelled signature values. If there is a cluster along a straight line, this will indicate a correlation of the two variables. A weighted least squares (WLS) regression model was used to include the individual event weights (cf. section 2.3.2), A fixed intercept of zero was used, as the theory indicates that this would be the optimal solution. The slope β was found by minimizing the weighted sum of squares (WSS) (eq. 2):

$$\beta \text{ from min } (WSS(\beta, w_e) = \sum_{e=1}^{n} w_e \left(y_{m,e} - y_{o,e}\beta\right)^2) \tag{2}$$

where y_o is the signature observation value, y_m is the signature modelled value, e is the event number and w_e is the weight of event e. The model is found adequate if the slope is 1. In practical terms, the WLS regression was conducted with the python package scikit-learn, Linear regression with sample weights (Pedregosa et al., 2011).

Four assumptions need to be valid to conduct weighted linear regression; linearity (linear relationship between y_o and y_m), homoscedasticity (the variance of the residuals is the same for all y_o), independence (the residuals are independent of each other), normality (the residuals are normally distributed) (Olive, 2017).

2.5.2 Indicator function

The score for the indicator function has a binary output: *I*. If the event is within the acceptance criteria (AC) (eq. 3) ("Indicator functions", Lectures on probability theory and mathematical statistics, 2022), indicated by the purple area in Figure 6, *I* will have a value of 1 (Eq. 3), and a total score across all events was calculated by considering the weights introduced in section 2.3.2 (Eq. 4):

$$I_{AC}(e) := \begin{cases} 1 & if\ e\ \in AC \\ 0 & if\ e\ \notin AC \end{cases} \tag{3}$$

$$score = \frac{\sum_{e=1}^{n} I_{AC} * w_e}{\sum_{e=1}^{n} w_e} \tag{4}$$

A good comparison gives a value of 1, as all events will be within the indicator function's acceptance criteria. The acceptance criteria were for this analysis made by a combination of a relative and an absolute criterion and were, for all sites, assessed to be the same as indicated in Table 2.

Table 2. Values representing the acceptance criteria for different signatures. The acceptance criteria are the combination of both the relative and absolute scale. The absolute values relate to the 1:1 line and the relative scale gives the slope-range of acceptance, *rv* is the relative value in this case 0.7. y_m is the modelled signature value and y_o is the observed signature value.

Relative scale	$rv > \frac{y_m}{y_o} < \frac{1}{rv}$
	$0.7 > \frac{y_m}{y_o} < 1.43$
Absolute scale	
Signature	
Peak level [m]	$y_m = y_o$ +/- 0.1
Duration above CSL [min]	$y_m = y_o$ +/- 20
Duration above CL [min]	$y_m = y_o$ +/- 20
AUC above CSL [m*min]	$y_m = y_o$ +/- 2
AUC above CL [m*min]	$y_m = y_o$ +/- 5
AUC between ZP/IL – TOP/CL [m*min]	$y_m = y_o$ +/- 10
No of peaks [-]	$y_m = y_o$ +/- 2
Max level rate of change (5 min resolution) [m/min]	$y_m = y_o$ +/- 0.01

2.5.3 Normalised RMSE

The RMSE function calculates the vertical distance (of modelled values) to the 1:1 line, to find the residuals in the model performance (Figure 6). RMSE is directly related to data and needs to be normalised to be (meaningfully) compared across sites and signatures. This can be achieved by dividing with e.g. the maximum value of the observations, or the interquartile range (IQR(y_0)) of the observations (the difference between the 75th and the 25th percentile). As the maximum value relies on extreme events, this is not considered as constituting a robust solution, and the IQR was thus chosen as the normalisation value (eq. 5):

$$RMSE(IQR) = \frac{\sqrt{\frac{1}{n}\sum_{e=1}^{n}(y_m - y_0)^2}}{IQR\ (y_0)} \tag{5}$$

The smaller the RMSE(IQR) value the better the model performance.

2.5.4 Score

An individual score (slope, score, or normalised RMSE, respectively) for each signature was calculated, which can be summarized to one score for the given objective.

$$Objective\ score_m = \frac{\sum_{s=1}^{z} score_{s,m}}{z} \tag{6}$$

score is the different score functions given by eq. 2, 4 and 5. *s* denotes signature, *m* is the method (either linear regression, indicator function or normalised RMSE), *z* is the total number of relevant signatures for the given objective. The optimal score is individual for the three methods, and direct comparison is therefore not possible.

2.6 Assessment criteria of the model performance score

As a last – and illustrative – step, the scoring of the model performance was categorised as indicated in Figure 1. The categorical analysis outcome was classified as either 'sufficient' or 'insufficient', and the statistical analysis outcome was classified by means of a 'traffic light' assessment (green=good performance, yellow= acceptable performance, red=poor performance). The criteria for which score falls into each assessment category were solely based on experience from the author group (Table 3). One consideration was not to be too 'hard' on the model, as it is expected that several other factors affecting the weights were not included in the calculations. There is limited prior experience dealing with criteria as such, as the method and the ambition to conduct multi-site analysis is new. Prospectively, future experience with the methods will strengthen the choice of criteria.

Table 3. The criteria for the categorisation of the model performance. Solely based on the utility company's preferences. x denotes the output from each method (slope, score, RMSE(IQR)).

	Linear regression	Indicator function	Normalised RMSE
Categorical			
Sufficient		CSI ≥ 0.6	
Insufficient		CSI < 0.6	
Statistical			
Ideal value	1	1	0
Green	0.70 > x < 1.43	x > 0.70	x < 0.60
Yellow	0.40 > x < 2.50	x > 0.40	x < 1.20
Red	0.40 ≤ x ≥ 2.50	x ≤ 0.40	x ≥ 1.20

3 Study area, model and data

The analysis covers two case areas: Bellinge and Dalum, in the utility company VCS Denmark's service area. The areas are characterised by being suburban areas with minor surface gradients. The urban drainage system analysed is a combined system and both areas are upstream from a main collecting pipe transporting the combined sewage to the treatment plant. The case areas for this study includes 23 sites with water level meters installed as highlighted in Figure 7. Some sites contain more than one level meter. The normal flow direction of the wastewater is illustrated, as are the combined sewer overflow locations (green dots). In Dalum, there are many 'ring-connections', where the combined sewage can be directed to several catchment areas in case of high water levels. This is however not indicated in the sketch.

The applied model is a semi-distributed 'integrated urban drainage model' (Bach et al., 2014). It includes a lumped-conceptual rainfall-runoff module that calculates runoff to a distributed, physics-based pipe flow module, computationally carried out in the software Mike Urban (DHI, 2020). The model setup is described in detail in Pedersen et al. (2021b) and is openly accessible. The rainfall-runoff is based on the time-area model (model A) – infiltration-inflow to pipes is not included. The model includes app. 3,500 nodes, and the imperviousness of sub-catchments was calculated based on a categorisation of the surface from satellite data using spectral analysis. Rain input was measured by two rain gauges in the proximity of the study area (Figure 1). The hydraulic reduction factor was set at 0.9 and the model was run continuously for app. 10 years (2010-2021) with a time step of 5 sec in the pipe-flow module. Water level time series from 23 sites in the study area are included in the analyses; these have durations between 2 and 11 years and include between 127 and 2,246 rain induced events with the event definition described in Pedersen et al. (2022). Observations and model output are in this paper presented in water level time series with a temporal resolution of 1 min.

Further description of sites and models implemented in Mike Urban and SWMM are available in Pedersen et al. (2021b) for Bellinge, and Pedersen et al. (2022) provides a description of three sites in these areas: F67F47Y, G73F010 and G71F05R_LevelBasin, including information about the horizontal area of the structure, the levels, crest widths and imperviousness and total areas.

https://doi.org/10.5194/egusphere-2022-615
Preprint. Discussion started: 25 July 2022

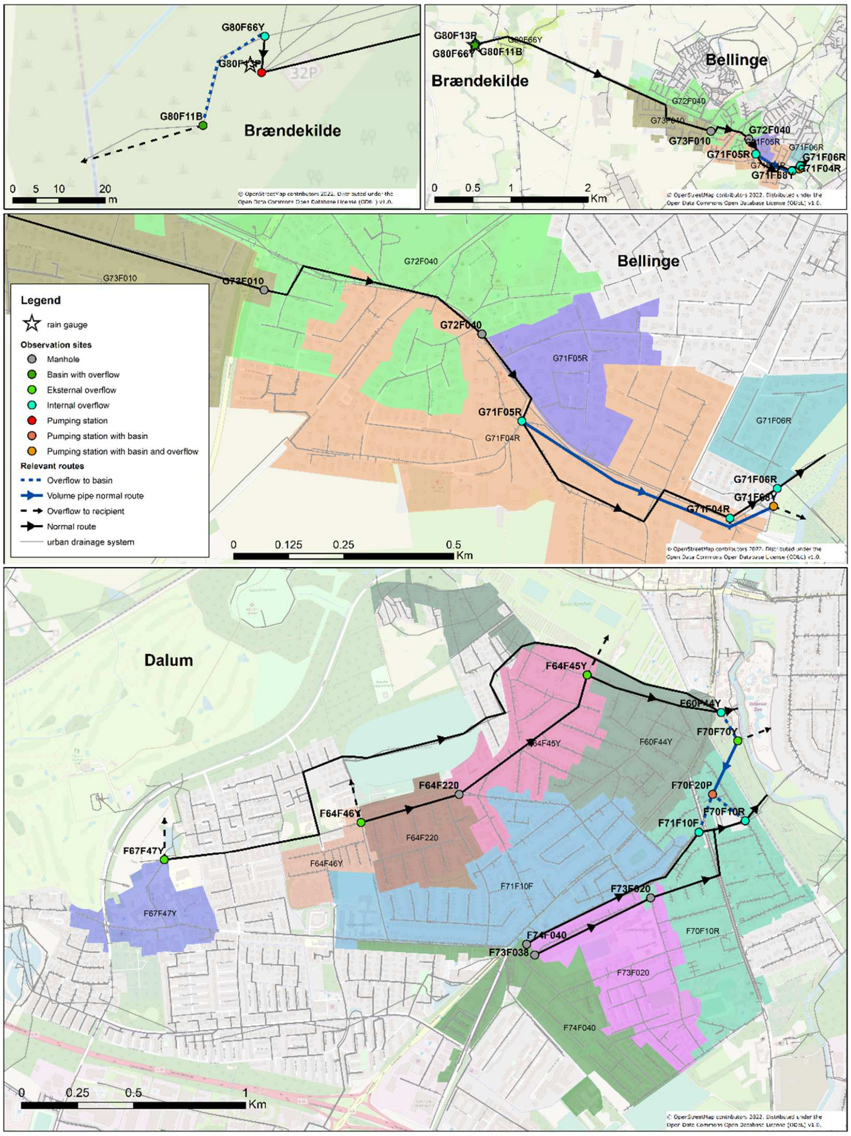


Figure 7. Observation sites in the case areas. Many internal connections between the different areas are present in Dalum, but these
330 **are not illustrated in the figure. The node names in the centre of the sub-catchments refer to the connected manholes downstream. Background map is from** OpenStreetMap (2022) (© OpenStreetMap contributors 2022. Distributed under the Open Data Commons Open Database License (ODbL) v1.0.)**.**

4 Results and discussion

4.1 Direct time series comparisons for different objectives

The time series analysed were split into events, as illustrated for six examples of events in Figure 8. The different events are examples of surcharge, overflow and everyday events in the different columns, and the rows indicate two different sites (F64F46Y (Figure 8a-c) and F70F70Y_LevelSump (Figure 8d-f)). Observed water levels (red dots and line) and modelled water levels (blue lines) are plotted, and weights (cf. section 2.3.2) and signatures related to the specific objectives are shown in the top right-hand corner of each panel. The grey areas illustrate the range in which the peak of the observed and modelled water levels should be, to be considered true positives and be included in the statistical analysis. Each event can be visually interpreted, and different model replications can be seen for the events. Figure 8e illustrates how the model replicates the objective overflow even though the model does not replicate the rest of the event. In this specific case, the event is not replicated well below crest level due to a missing global control setting for the pump emptying this site. The opposite is seen for Figure 8a, where the objective surcharge is not replicated very well (a heavy overestimation of the peak level by the model), but the rest of the event shows better performance. However, this is of no interest if we are aiming to figure out the performance for critical surcharge levels. Figure 8d-f clearly illustrates a difference in the lower sensor limitation, ZP and IL. As illustrated in Figure 8f, the area of interest is limited to the level between ZP and TOP. This also illustrates that with the given objectives in this paper, the water level range between TOP to CL for site F70F70Y_LevelSump will not be included in the analyses.

In the top right-hand corner of all panels in Figure 8, the weights of the events are indicated, cf. section 2.3.2. “Weight rain” refers to weights calculated from the spatial rainfall information, and “weight obs.” refers to weights determined from information about known system anomalies and the sensor upper limitation. Figure 8b and Figure 8d show a very low weight on the rain, indicating that these events will not be valued highly in the further analysis. Furthermore, it is seen that Figure 8d has reached the upper sensor limitation and the entire event will therefore get a weight equal to 0 for peak level and AUC above CSL. The duration is not affected by the sensor upper limitation and therefore this signature will still count in the further analysis of this signature, however still with a weight of 0.23 from the rain gauge uncertainty.

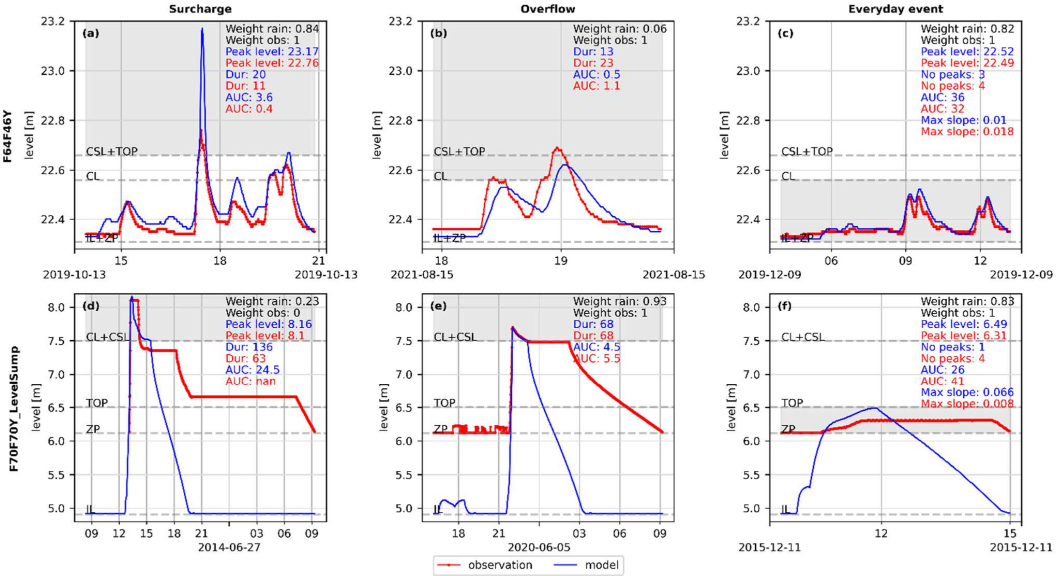


Figure 8. Examples of timeseries plotted for two sites (rows): F<64F46Y (a-c) and F70F70Y_LevelSump (d-f). The objectives are illustrated in the columns as surcharge, overflow and everyday events respectively. The values indicated in the top right-hand corner of each subplot indicate the values of the signatures for the three objectives for the modelled event (blue) and the observed event (red). The weights of the rain gauge uncertainty are indicated as well as the 'weight obs' which is a combination of the weights from known system anomalies and indications of whether the sensor reached an upper limitation.

4.2 Illustration of different categorical and statistical scores

To illustrate the procedure of categorical and statistical analysis, multi-event signature comparison plots are shown for three sites for the objective "overflow" in Figure 9. Multi-event signature comparison plots for all signatures and objectives across all 23 sites can be found in the supplementary material. Each column in Figure 9 illustrates a signature, peak level, duration above CL and AUC above CL, respectively. The grey areas on the peak level comparison plots illustrate the range of true positives for the objective "overflow". The weight of the events is indicated by the coloured scale, where events assigned a weight of 0 (red colour) do not have any influence on the output of the methods. The true positives for the objective "overflow" are plotted in the last two columns. Important elements are illustrated for the three investigated statistical analyses: linear regression (with slope as a black dashed line), indicator function (with the purple area as acceptance criteria) and the normalised RMSE (with the IQR indicated by vertical blue lines (Q25 and Q75)). The score values from the three methods are indicated below each subplot.

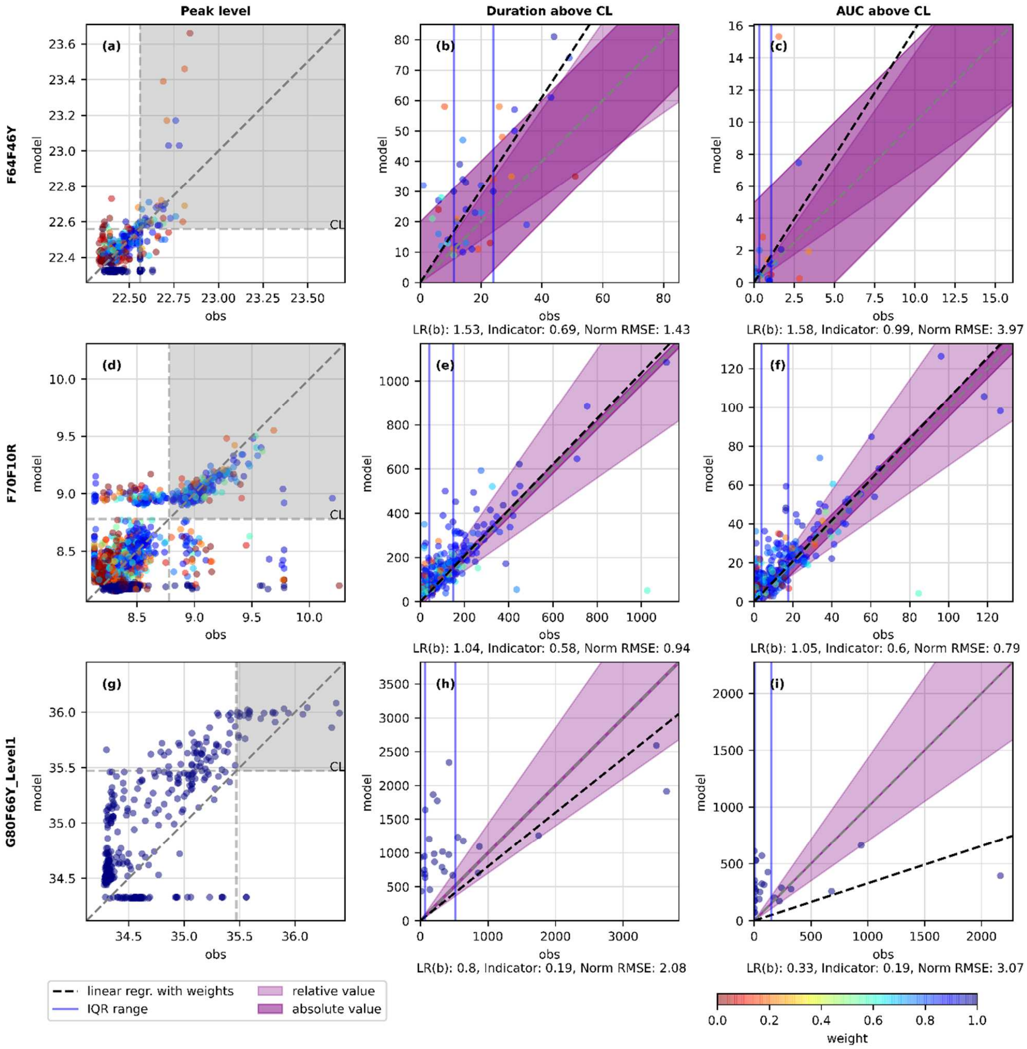


Figure 9. Multi-event signature comparison plots for different sites (rows) and signatures (columns). Left column: the true positive ranges are indicated (grey area), where peak level is above CL for both observed and modelled events. Middle and right columns: the true positive events are the only events plotted for the two signatures: duration above CL and AUC above CL. Illustration of important elements of for the three methods: linear regression (dashed black line), indicator function (acceptance criteria with purple) and normalised RMSE (blue lines indicating IQR). The weight of each event is illustrated with the colour bar. The 1:1 line is plotted with a grey dashed line. The score of the statistical methods is indicated below the multi-event signature comparison plot.

https://doi.org/10.5194/egusphere-2022-615
Preprint. Discussion started: 25 July 2022

For linear regression, Figure 9c illustrates an event at modelled value 15, which is identified with an uncertain rain gauge input, given the weight of zero and therefore does not affect the slope. The slope from the linear regression is, on the other hand, affected in Figure 9i, where a few very large observation values force a low slope gradient even though modelled values seem to be higher at low observation values. For Figure 9e and f the slope is very close to the 1:1 line, indicating a close to perfect fit with the model. For the indicator function, the purple area illustrates the area where the acceptance criteria are met. Events that are within this area have an indictor value of 1, and their weights are counted in the numerator of Eq. (1), whereas the sum of all the weights are counted in the denominator. The area of acceptance is of great importance as can be seen for F64F46Y (Figure 9c), where many events are inside the acceptance criteria, which is also indicated by the score of 0.99. The absolute acceptance criteria are the same for all sites for each signature. It can be discussed if the acceptance criteria should be the same for all sites, or if there is site-specific interest that should be taken into account, especially when the utility company sets this evaluation into operation. The IQR are illustrated with the vertical blue lines and looking at Figure 9f, a range of app. 15 m*min is seen, which is applied to normalise the RMSE. If the error is larger than the IQR, the normalised RMSE will not be within zero to one, and it is therefore difficult to compare values between sites and signatures.

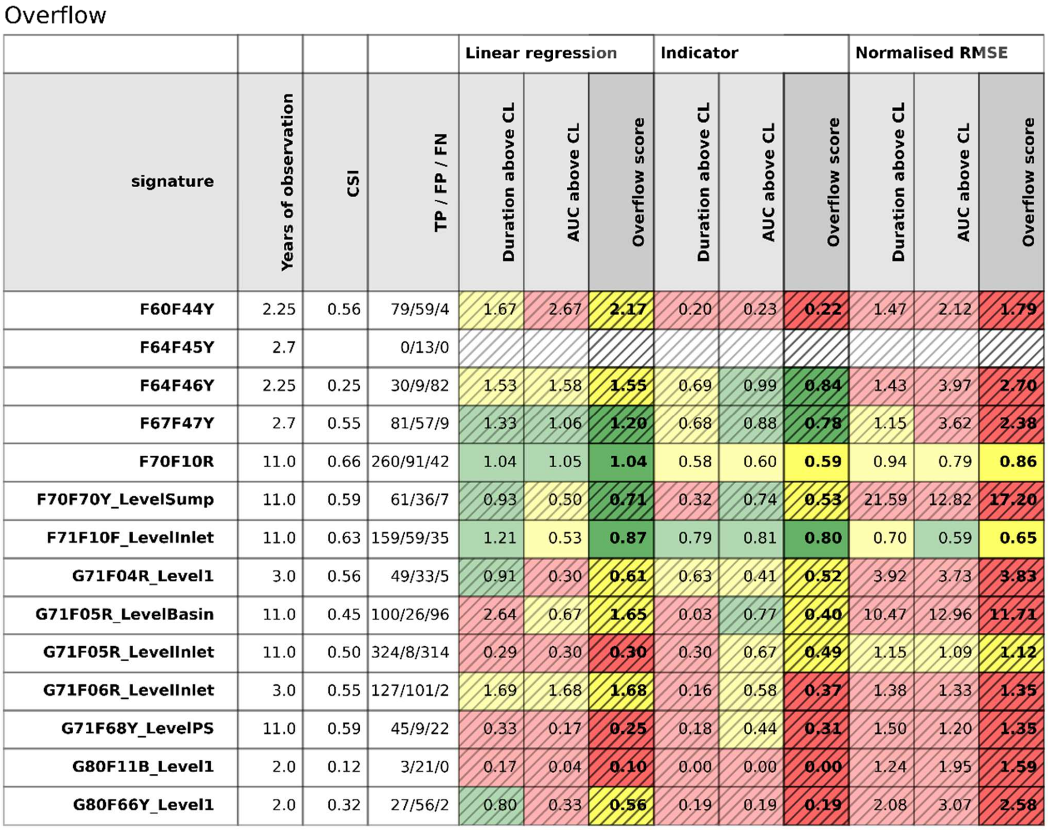
The last site G80F66Y_Level1 in Figure 9g-h-i, does not show any weights. For this site, there is only one rain gauge within a 5 km distance of the upstream catchment, and the coefficient of variation cannot be calculated for a single rain gauge. All events thus have the same weight of 1.

4.3. Comparison of methods for statistical analysis

Each categorical and statistical method relies on different metrics as shown in Figure 9. In Figure 10 the results from for the objective ‘overflow’ based on the three statistical methods are shown, with hatched and color-coded scorings (Figure 1, Table 3). Here results are shown only for the 14 sites where overflow is either modelled or measured. The ‘overflow scores’ are highlighted and were calculated as the average of the two signatures indicated with more transparent colours (Eq. 6), The differences between the methods are large, e.g. where the normalised RMSE method does not generate a very low and optimal score. The first thing to notice is that we have three sites, F70F10R, F70F70Y_LevelSump and F71F10F_LevelInlet where the categorical analysis shows sufficient performance (cells with no hatch, CSI > 0.6). From the CSI values provided in Figure 10, many values appear above 0.5, indicating that at least half of both modelled and observed positive events were simulated in the same category. However, the threshold for CSI was set to be above 0.6 to have sufficient categorical performance (Table 3), and therefore it is not enough. Two of the sites, F70F10R and F71F10F_LevelInlet perform categorically well (no hatched cells) and as acceptable or good for all three statistical methods (yellow and green colours only) when focusing on the objective ‘overflow’ (Figure 10). Looking at G80F66Y_Level1 in Figure 10 illustrates that the duration is assessed to be good for the linear regression method. However, when looking at Figure 9h a different reality emerges. The variance of data is large and because of three large observation values below the 1:1 line, the slope is, coincidentally, within the ideal range. The linear regression method could therefore potentially be further improved by including the variance in data in the assessment.

https://doi.org/10.5194/egusphere-2022-615
Preprint. Discussion started: 25 July 2022

415 Each method has advantages and disadvantages as they are based on selected statistical metrics favouring specified features in the modelled and observed signatures. The identified pros and cons for each method are indicated in Table 4, as well as suggestions for improvements of the methods.



Overflow

				Linear regression			Indicator			Normalised RMSE		
signature	Years of observation	CSI	TP / FP / FN	Duration above CL	AUC above CL	Overflow score	Duration above CL	AUC above CL	Overflow score	Duration above CL	AUC above CL	Overflow score
F60F44Y	2.25	0.56	79/59/4	1.67	2.67	**2.17**	0.20	0.23	**0.22**	1.47	2.12	**1.79**
F64F45Y	2.7		0/13/0									
F64F46Y	2.25	0.25	30/9/82	1.53	1.58	**1.55**	0.69	0.99	**0.84**	1.43	3.97	**2.70**
F67F47Y	2.7	0.55	81/57/9	1.33	1.06	**1.20**	0.68	0.88	**0.78**	1.15	3.62	**2.38**
F70F10R	11.0	0.66	260/91/42	1.04	1.05	**1.04**	0.58	0.60	**0.59**	0.94	0.79	**0.86**
F70F70Y_LevelSump	11.0	0.59	61/36/7	0.93	0.50	**0.71**	0.32	0.74	**0.53**	21.59	12.82	**17.20**
F71F10F_LevelInlet	11.0	0.63	159/59/35	1.21	0.53	**0.87**	0.79	0.81	**0.80**	0.70	0.59	**0.65**
G71F04R_Level1	3.0	0.56	49/33/5	0.91	0.30	**0.61**	0.63	0.41	**0.52**	3.92	3.73	**3.83**
G71F05R_LevelBasin	11.0	0.45	100/26/96	2.64	0.67	**1.65**	0.03	0.77	**0.40**	10.47	12.96	**11.71**
G71F05R_LevelInlet	11.0	0.50	324/8/314	0.29	0.30	**0.30**	0.30	0.67	**0.49**	1.15	1.09	**1.12**
G71F06R_LevelInlet	3.0	0.55	127/101/2	1.69	1.68	**1.68**	0.16	0.58	**0.37**	1.38	1.33	**1.35**
G71F68Y_LevelPS	11.0	0.59	45/9/22	0.33	0.17	**0.25**	0.18	0.44	**0.31**	1.50	1.20	**1.35**
G80F11B_Level1	2.0	0.12	3/21/0	0.17	0.04	**0.10**	0.00	0.00	**0.00**	1.24	1.95	**1.59**
G80F66Y_Level1	2.0	0.32	27/56/2	0.80	0.33	**0.56**	0.19	0.19	**0.19**	2.08	3.07	**2.58**

Figure 10. Results from the categorical and the three statistical analysis methods for the objective overflow. Color-coding as described in Figure 1 and Table 3. 420

Table 4. The advantages, disadvantages and improvements are outlined for the three methods.

Linear regression		
	Advantages	· Illustrates nicely the "direction" of the error (above or below slope 1) · Few parameters, easy to understand · Weights are easy to include
	Disadvantages	· The large values count too much · Heteroscedasticity occurs, and confidence intervals cannot be applied
	Improvements	· Include the variance of the slope · The weight function could include the signature value to downgrade the importance of the large ones
Indicator function		
	Advantages	· Weights are easy to include · Do include an acceptable residual
	Disadvantages	· General absolute values in acceptance criteria are hard to set · Do not tell anything about the size of the residual
	Improvements	· Make site-specific acceptance criterion instead of the general acceptance criteria
Normalised RMSE		
	Advantages	· Include the size of residual
	Disadvantages	· Hard to make comparable if the residual is larger than the variance in data · If normalised by max value, dependent on very extreme value · Do not apply weights · Truncated data which are not distributed normally · Hard to find an optimal normalisation range. If we rely on maximum values, they do not occur that often, and would make a skewness towards long-monitored sites. The residuals are larger than the IQR. This makes it hard to compare across sites and signatures, as the score will be above 1.
	Improvements	· Finding a suited normalisation value would improve the method

4.4 Model performance for all sites and for all objectives

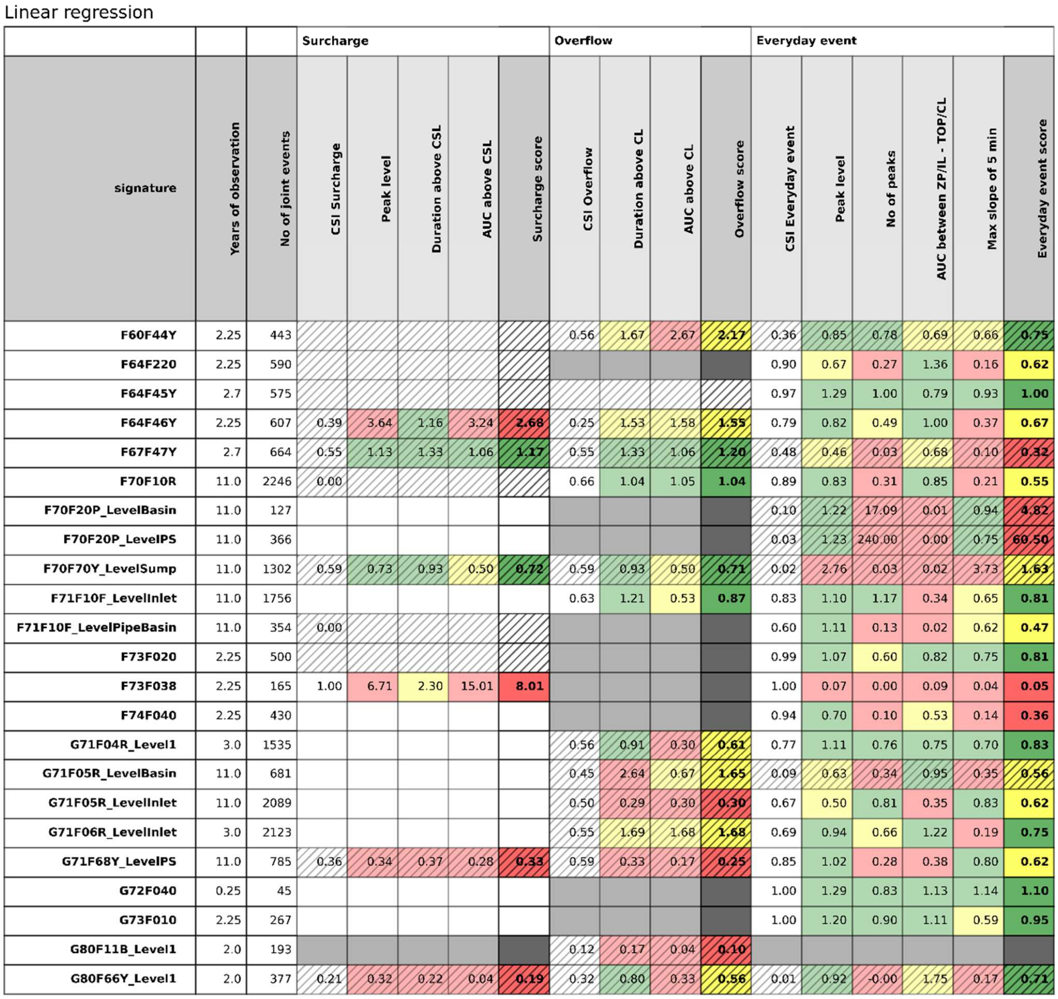
In Figure 11, the results for all three objectives using the 'linear regression' statistical analysis method are summarised across all 23 sites. Looking at the performance score for all objectives, it can be easily highlighted where the model performs well for different objectives. What is first seen is that for many sites a better performance score is obtained for the objective "everyday events" than for the objectives "surcharge" and "overflow". This is not surprising results for VCS Denmark, as until now the utility company compares model results with observations manually. Events falling within the range of "everyday event" were much more often applied in comparisons, as they, by nature, occur more often. And when they fit - or detective work showed no more misunderstandings in the model - it was simply assumed that this also applies to the surcharge and overflow events. This was not a correct assumption, as is seen in Figure 11. When the model is primarily applied for planning and design purposes, i.e. for the overflow and surcharge, this is a wakeup call for the utility company to change its practices. The sites furthest upstream in the catchment area (F67F47Y, F64F45Y, F73F038, F74F040, G80F66Y_Level1, see Figure 7) generally perform more poorly than the rest of the sites. This can be due to the fact that the outliers observed cannot be "averaged-out" downstream, but also that the upstream structures (pipes, manholes) most often are much smaller in diameter than structures at downstream sites, and that water level variations are thus most probably more dynamic at upstream sites. In theory, the model should be able to simulate the system just as well in the upstream sites, and there should be an increasing awareness of this issue.

For "surcharge", many blank fields are seen in Figure 11, as not much data is available for these more extreme events. The categorical analysis shows an insufficient performance for many sites, indicated by a hatched white cell - often meaning that the model simulated a surcharge event that was not observed in reality.

Regarding the different signatures, the number of peaks and the different AUCs generally perform more poorly than the other signatures. The AUC is a combination of both a level and a time unit and can therefore be, due to complexity, harder to simulate, but it could also be that diagnostic tools (Pedersen et al., 2022) can identify where the errors occur so that the model can be improved.

The overall score for each objective was calculated as the average of the relevant signatures (as described in eq. 6). However, when regression slopes are extreme, as e.g. for the number of peaks for everyday events at the site F70F20P_LevelPS (regression slope=240), the overall score will naturally be affected. A very low score of e.g. 0.02 will not affect the objective score as much but is naturally an extreme as well. It can be discussed if the objective score should be an average of the relevant signatures, the maximum or the minimum, or more advanced calculations should be implemented, but no matter what is agreed upon, one must also have an eye on the regression slopes for different signatures, as well as on the uncertainty of these.

The assessment of the 'traffic-light' could have been analytically conducted by interviewing several experts or making further analysis. Because the methodology is rather new, as is the signature method that the work is based on, the hypothesis is that experts would not yet be familiar with this analysis, and efforts with obtaining input from more experts would thus not be fruitful. For now, this final step was limited to be an assessment based on the authors' experience.



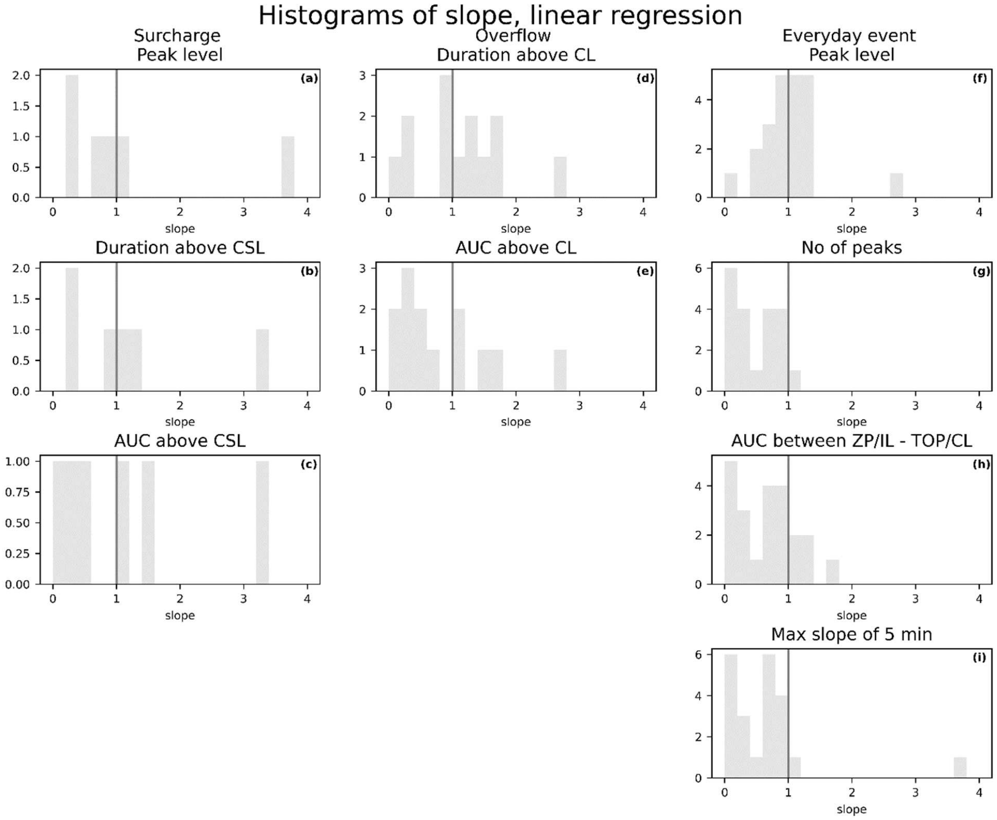
Linear regression

			Surcharge					Overflow				Everyday event					
signature	Years of observation	No of joint events	CSI Surcharge	Peak level	Duration above CSL	AUC above CSL	Surcharge score	CSI Overflow	Duration above CL	AUC above CL	Overflow score	CSI Everyday event	Peak level	No of peaks	AUC between ZP/IL - TOP/CL	Max slope of 5 min	Everyday event score
F60F44Y	2.25	443						0.56	1.67	2.67	**2.17**	0.36	0.85	0.78	0.69	0.66	**0.75**
F64F220	2.25	590										0.90	0.67	0.27	1.36	0.16	**0.62**
F64F45Y	2.7	575										0.97	1.29	1.00	0.79	0.93	**1.00**
F64F46Y	2.25	607	0.39	3.64	1.16	3.24	**2.68**	0.25	1.53	1.58	**1.55**	0.79	0.82	0.49	1.00	0.37	**0.67**
F67F47Y	2.7	664	0.55	1.13	1.33	1.06	**1.17**	0.55	1.33	1.06	**1.20**	0.48	0.46	0.03	0.68	0.10	**0.32**
F70F10R	11.0	2246	0.00					0.66	1.04	1.05	**1.04**	0.89	0.83	0.31	0.85	0.21	**0.55**
F70F20P_LevelBasin	11.0	127										0.10	1.22	17.09	0.01	0.94	**4.82**
F70F20P_LevelPS	11.0	366										0.03	1.23	240.00	0.00	0.75	**60.50**
F70F70Y_LevelSump	11.0	1302	0.59	0.73	0.93	0.50	**0.72**	0.59	0.93	0.50	**0.71**	0.02	2.76	0.03	0.02	3.73	**1.63**
F71F10F_LevelInlet	11.0	1756						0.63	1.21	0.53	**0.87**	0.83	1.10	1.17	0.34	0.65	**0.81**
F71F10F_LevelPipeBasin	11.0	354	0.00									0.60	1.11	0.13	0.02	0.62	**0.47**
F73F020	2.25	500										0.99	1.07	0.60	0.82	0.75	**0.81**
F73F038	2.25	165	1.00	6.71	2.30	15.01	**8.01**					1.00	0.07	0.00	0.09	0.04	**0.05**
F74F040	2.25	430										0.94	0.70	0.10	0.53	0.14	**0.36**
G71F04R_Level1	3.0	1535						0.56	0.91	0.30	**0.61**	0.77	1.11	0.76	0.75	0.70	**0.83**
G71F05R_LevelBasin	11.0	681						0.45	2.64	0.67	**1.65**	0.09	0.63	0.34	0.95	0.35	**0.56**
G71F05R_LevelInlet	11.0	2089						0.50	0.29	0.30	**0.30**	0.67	0.50	0.81	0.35	0.83	**0.62**
G71F06R_LevelInlet	3.0	2123						0.55	1.69	1.68	**1.68**	0.69	0.94	0.66	1.22	0.19	**0.75**
G71F68Y_LevelPS	11.0	785	0.36	0.34	0.37	0.28	**0.33**	0.59	0.33	0.17	**0.25**	0.85	1.02	0.28	0.38	0.80	**0.62**
G72F040	0.25	45										1.00	1.29	0.83	1.13	1.14	**1.10**
G73F010	2.25	267										1.00	1.20	0.90	1.11	0.59	**0.95**
G80F11B_Level1	2.0	193						0.12	0.17	0.04	**0.10**						
G80F66Y_Level1	2.0	377	0.21	0.32	0.22	0.04	**0.19**	0.32	0.80	0.33	**0.56**	0.01	0.92	-0.00	1.75	0.17	**0.71**

Figure 11. Table of scores for linear regression with weighted events. The colours refer to the overall performance score; good (green), acceptable (yellow) and poor (red). The white area is where there are not enough 'true positives' to evaluate a score (no<3, cf. Figure 2). The hatched areas refer to the categorical analysis, where too many events are not true positive, meaning that they are not modelled or observed. The grey/black area indicate where analysis is not possible due to physical constraints at the site, e.g. that not all sites have a crest level and evaluation of overflow is thus not possible

.

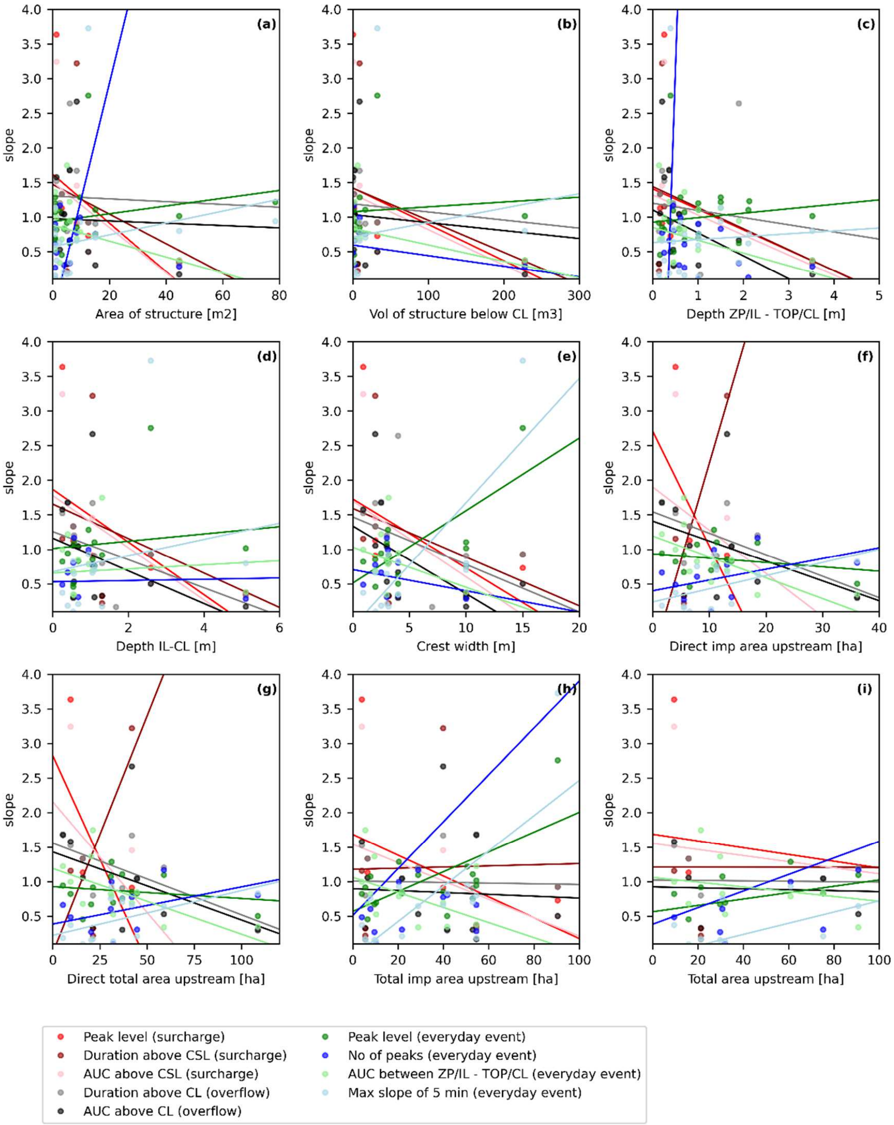
Figure 12 shows histograms of regression slopes across sites for all objectives and signatures, illustrating the consistency of the resulting slope from the linear regression throughout the sites. The ideal would, of course have been values around slope 1, but this is not the case. The peak level for "everyday events" (Figure 12f) is nicely represented as a normal distribution, but 470 others such as AUC above CL (Figure 12e) are distributed more densely towards low slopes. The histogram does not show the result of the categorical analysis, but only the statistical analysis of the true positives (there are 6, 13 and 22 true positives for the three objectives, which also appears from Figure 12).



475 **Figure 12. Histograms of the regression slope across sites from the statistical analysis (linear regression). Each histogram applies to one individual objective and signature. The slope is only considered if the number of true positives is above 1, as in Figure 11. Slopes higher than 4 are not seen in the histograms. The ideal slope is 1, and a normal distribution would appear with a mean of 1, and the values can go from 0-1 if model values are underestimated, and from 1-infinity if model values are overestimated.**

4.4 Multi-site correlations

Figure 13 shows the correlation between different site-specific variables and the multi-event signature comparison of regression slopes resulting from the linear regression. Each site is plotted with a dot, and linear regression lines are plotted to highlight any tendencies. Notice that this analysis does not take the categorical analysis into consideration. Generally, the picture is not consistent and clear. However, as illustrated in Table 4, the slope calculated from the linear regression may include a high uncertainty itself. The correlation analysis can therefore support the diagnostics of slopes. The slope of the regression line is therefore interesting. For the signatures in the objective everyday event (blue and green colours), the signatures: 'number of peaks' and 'AUC' show a relation toward the variables concerning the connected catchment (Figure 13f-i). With increasing area (either impervious or total, direct or total upstream connected) the model tends towards overestimation (positive signature comparison slope). The number of peaks for everyday events (dark blue colours), are generally highly affected by variables, e.g. Figure 13a, c and h. It is, however, necessary to give awareness to the calculation of this signature. Sensitivity is very high and could probably be improved. The signatures related to overflow, duration and AUC above CL (black and grey colours), generally follow the same trend. Interestingly, the increasing crest width (Figure 13e) seems to result in underestimated model results, meaning that the model underestimates the durations of the overflow event, if the crest level becomes too high. The depth range between the ZP/IL to TOP/CL (Figure 13c) shows the same tendency. The larger the range, the more the model tends to underestimate what is observed. For the signatures related to surcharge (red colours), the tendency lines are very inconsistent. Only six sites are included in the analysis of surcharge events, as seen in Figure 11, because not all sites have a value for the variable. It is therefore assessed that the number of values is too low to extract knowledge from these signatures. Generally, this correlation analysis would be strengthened by including more sites than the 23 provided in this paper, and by developing analytical methods that address uncertainty better than in this work.



500 **Figure 13. Correlation between a variable on the horizontal axis and linear regression slope from the multi-event signature comparison on the vertical axis (given in Figure 11). A dot represents a site. Linear regression lines are fitted to the dots to spot any tendencies. If the slope on the vertical axis is above 1, the signature at the site is overestimated in the model, whereas slopes below 1 indicate signatures that are underestimated.**

4.5 Communication to the utilities

505 This analysis is intended to be applied in the service area of the utility company for at least 165 sites and an easy overview is therefore needed. To communicate the performance score, maps will be generated because these provide a good overview of performance in relation to where the sites are located (see example in Figure 14). Together with score tables (Figure 11) and multi-event signature comparison plots (Figure 9), these will provide a strong basis for communication concerning the reliability of models. Performance scores for the three statistical methods can be found in the supplementary material, as well 510 as performance maps for all objectives using the linear regression method.

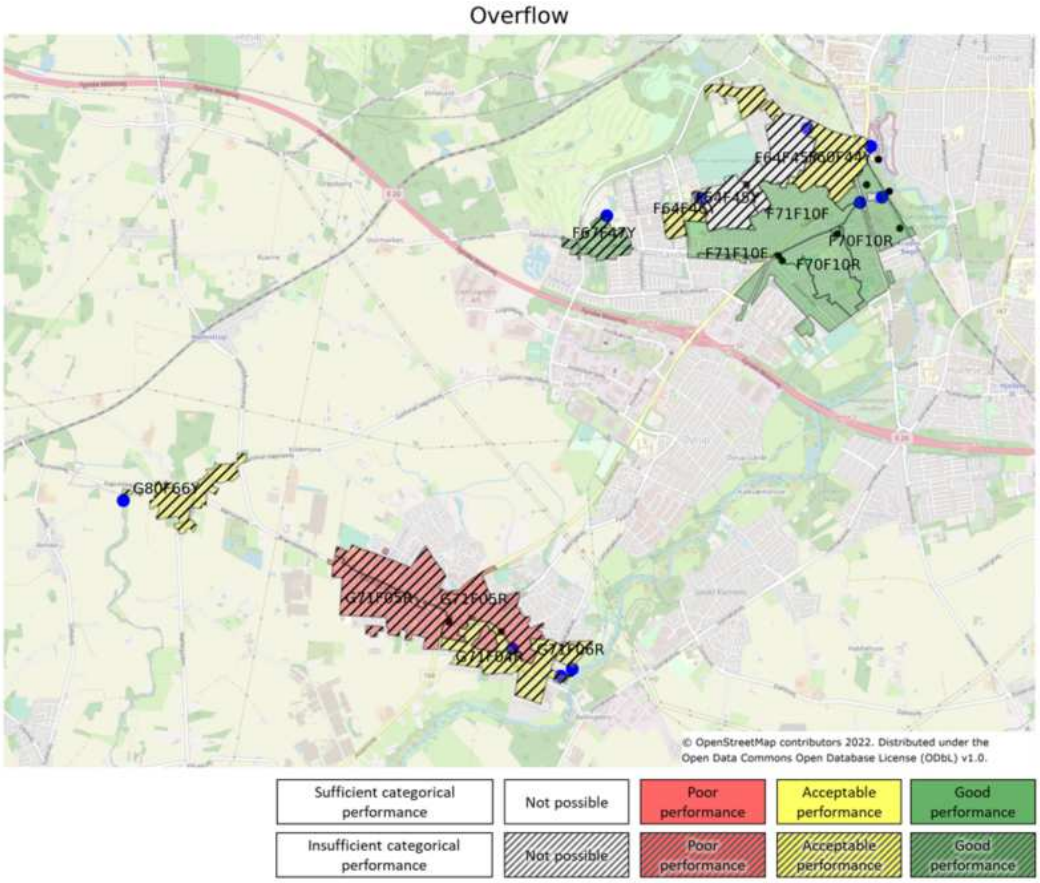


Figure 14. Map of the performance for overflow using the method of linear regression. The upstream catchment area of the site is mapped, and the naming in the catchment refers to the overflow structure that is mapped. The catchment area represents the case areas. The urban areas in between the catchment areas are not connected to the case areas, as they have a separate stormwater 515 **system. Background map is from** OpenStreetMap (2022)**.**

5 Conclusion

Large, detailed, distributed simulation models are widely applied in utility companies. They support decision-making relating to future investments and are applied to support the daily operation of the current urban drainage systems. When routinely comparing model results with in-sewer level measurements from an increasing number of sites, uncertainties previously not realised become visible. The slightly provocative question "how useful are the models then?" highlights the need for methods to systematically assess and investigate model performance. Can we always rely on the model results in urban drainage modelling? The herein developed 5-step framework suggests a method for answering this question. The steps are as follows: objective identification, context identification, categorical analysis, statistical analysis, and assessment. The method is based on the methodology of signatures, which are metrics extracting certain characteristics of a time series event. Observed and modelled signatures can be compared across many events. Three objectives were identified for this study: 'surcharge', 'overflow' and 'everyday events', and relevant signatures for each of these objectives were determined, including the 'Area Under Curve' (AUC) which is a 'surrogate volume' for an event determined from a reference level (similar to the area under a flow hydrograph, but with a different unit). The employed signatures were for surcharge: peak level, duration above CSL (critical surcharge level), AUC above CSL; for overflow: duration above CL (crest level), AUC above CL; and for everyday events: peak level, no. of peaks, AUC between TOP/CL (top of pipe, or crest level) and ZP/IL (zero-point of sensor, or invert level). Each event was analysed to determine whether observed and modelled values occur in the same category, e.g. if both observed and modelled results show an overflow event. Only the categorical true positive events were further subject to statistical analysis, to assess how 'correct' the model is. Three methods of multi-event statistical analysis of signatures were proposed and investigated in this study: linear regression, an indicator function, and a normalised RMSE method. The final step assessed the values obtained in the categorical and statistical analysis and placed the signatures for each site into distinct categories: for all events sufficient or insufficient categorical performance, and for true positive events also good, acceptable or poor statistical performance.

A method was furthermore developed to reduce the weighting of individual events in the statistical analyses, in order to acknowledge uncertainty in model and/or measurements in cases where the model is not expected to fully replicate the measurements. The weight calculation includes: rain gauge uncertainty due to spatial variability in the proximity of the site, known system anomalies, and sensor limitations. The three methods in the statistical analysis were investigated, which highlighted the differences between the methods but also areas in which further improvements can be made; these improvements notably include tests for statistical significance, inclusion site-specific assessment criteria, and normalising performance scores across methods, objectives and sites.

A case study covering 23 sites in two areas was conducted using the developed framework, which highlighted a number of model uncertainties for certain sites. For the objective "overflow", only three among 14 investigated sites were categorized as exhibiting sufficient categorical performance, whereas the remaining 11 sites had too many events where the observed and modelled signatures fell into different categories. Generally, the model performed better for everyday events, compared to surcharge and overflow events, which is not surprising due to the previous tradition of model validation in the local utility company (VCS Denmark). With the developed method, the models are useful for some signatures, but clearly not useful for others, especially for some sites. Further improvements may include a general assessment of the performance criteria, as well as more elaborate statistical analysis, as suggested in the paper. Our results point to a general need for more research on model performance and error detection methods that can be applied when comparing simulation results from large, detailed, distributed urban drainage models with observations from tens, hundreds and perhaps thousands of sensor locations.

Code availability. Code is not made publicly available; readers should contact the corresponding author for details.

Data availability. Data is available for some areas as described in Pedersen et al. (2021b).

Author contributions. All authors jointly contributed to conceptualising and designing the study, discussing results, and drafting or revising the manuscript. ABK and ANP designed the observation programme. ANP, ABK and PSM discussed the framework together. ANP simulated model results, collected and curated the observation data, and prepared the software tools for model performance evaluation. ANP prepared the initial draft manuscript and the visualisations. ANP, ABK and PSM revised the manuscript and prepared the final submitted paper. ABK and PSM supervised the project.

Competing interests. The authors declare that they have no conflict of interests.

Acknowledgements. The authors would like to thank Morten Borup (Krüger A/S), Lasse Engbo Christiansen (Statens Serum Instititut) and Jörg Rickermann (EAWAG) for valuable discussions prior to and during this study. The authors would also like to thank colleagues in VCS Denmark for contributing with valuable information about sensors and field measurements, when necessary. The work was carried out in collaboration between the Technical University of Denmark and VCS Denmark through an Industrial Ph.D. project partly funded by VCS Denmark and Innovation Fund Denmark (grant no. 8118-00018B).

References

Bach, P. M., Rauch, W., Mikkelsen, P. S., McCarthy, D. T., and Deletic, A.: A critical review of integrated urban water modelling - Urban drainage and beyond, Environmental Modelling & Software, 54, 88–107, https://doi.org/10.1016/j.envsoft.2013.12.018, 2014.

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., and Andreassian, V.: Characterising performance of environmental models, Environmental Modelling and Software, 40, 1–20, https://doi.org/10.1016/j.envsoft.2012.09.011, 2013.

Beven, K. and Binley, A.: The future of distributed models: Model calibration and uncertainty prediction, Hydrological Processes, 6, 279–298, https://doi.org/10.1002/hyp.3360060305, 1992.

Box, G. E. P.: Robustness in the Strategy of Scientific Model Building, in: Robustness in Statistics, edited by: Launer, R. L. and Wilkinson, G. N., ACADEMIC PRESS, INC., 201–236, https://doi.org/10.1016/B978-0-12-438150-6.50018-2, 1979.

Clemens-Meyer, F. H. L. R., Lepot, M., Blumensaat, F., Leutnant, D., and Gruber, G.: Data validation and data quality assessment, in: Metrology in Urban Drainage and Stormwater Management: Plug and Pray, edited by: Bertrand-Krajewski, J.-L., Clemens-Meyer, F., and Lepot, M., IWA Publishing, 327–390, https://doi.org/10.2166/9781789060119_0327, 2021.

DANVA: Vand i tal 2021, 2021.

Mike Urban: https://www.mikepoweredbydhi.com, last access: 17 August 2020.

Eggimann, S., Mutzner, L., Wani, O., Schneider, M. Y., Spuhler, D., Moy De Vitry, M., Beutler, P., and Maurer, M.: The Potential of Knowing More: A Review of Data-Driven Urban Water Management, Environmental Science and Technology, 51, 2538–2553, https://doi.org/10.1021/acs.est.6b04267, 2017.

Ehlers, L. B., Wani, O., Koch, J., Sonnenborg, T. O., and Refsgaard, J. C.: Using a simple post-processor to predict residual uncertainty for multiple hydrological model outputs, Advances in Water Resources, 129, 16–30, https://doi.org/10.1016/j.advwatres.2019.05.003, 2019.

Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., and Savenije, H. H. G.: A framework to assess the realism of model structures using hydrological signatures, Hydrol. Earth Syst. Sci, 17, 1893–1912, https://doi.org/10.5194/hess-17-1893-2013, 2013.

Fenicia, F. and Kavetski, D.: Behind every robust result is a robust method: Perspectives from a case study and publication process in hydrological modelling, Hydrological Processes, 35, 1–9, https://doi.org/10.1002/hyp.14266, 2021.

Gupta, H. V., Wagener, T., and Liu, Y.: Reconciling theory with observations: elements of a diagnostic approach to model evaluation, Hydrological Processes, 22, 3802–3813, https://doi.org/10.1002/hyp.6989, 2008.

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., and Ye, M.: Towards a comprehensive assessment of model structural adequacy, Water Resources Research, 48, 1–16, https://doi.org/10.1029/2011WR011044, 2012.

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., and Andréassian, V.: Large-sample hydrology: A need to balance depth with breadth, Hydrology and Earth System Sciences, 18, 463–477, https://doi.org/10.5194/hess-18-463-2014, 2014.

Kerkez, B., Gruden, C., Lewis, M., Montestruque, L., Quigley, M., Wong, B., Bedig, A., Kertesz, R., Braun, T., Cadwalader, O., Poresky, A., and Pak, C.: Smarter stormwater systems, Environmental Science and Technology, 50, 7267–7273, https://doi.org/10.1021/acs.est.5b05870, 2016.

McMillan, H. K.: A review of hydrologic signatures and their applications, Wiley Interdisciplinary Reviews: Water, https://doi.org/10.1002/wat2.1499, 2020a.

McMillan, H. K.: Linking hydrologic signatures to hydrologic processes: A review, Hydrological Processes, 34, 1393–1409, https://doi.org/10.1002/hyp.13632, 2020b.

Odense Kommune: Spildevandsplan 2011-2022, 2011.

Olive, D. J.: Linear Regression, Springer International Publishing, 1–494 pp., https://doi.org/10.1007/978-3-319-55252-1, 2017.

OpenStreetMap: www.openstreetmap.org, last access: 3 February 2022.

Palmitessa, R., Mikkelsen, P. S., Borup, M., and Law, A. W. K.: Soft sensing of water depth in combined sewers using LSTM neural networks with missing observations, Journal of Hydro-Environment Research, 38, 106–116, https://doi.org/10.1016/j.jher.2021.01.006, 2021.

Pedersen, A. N., Borup, M., Brink-Kjær, A., Christiansen, L. E., and Mikkelsen, P. S.: Living and Prototyping Digital Twins for Urban Water Systems: Towards Multi-Purpose Value Creation Using Models and Sensors, Water (Basel), 13, 592, https://doi.org/10.3390/w13050592, 2021a.

Pedersen, A. N., Pedersen, J. W., Vigueras-Rodriguez, A., Brink-Kjær, A., Borup, M., and Mikkelsen, P. S.: The Bellinge data set: open data and models for community-wide urban drainage systems research, Earth System Science Data, 13, 4779–4798, https://doi.org/10.5194/essd-13-4779-2021, 2021b.

Pedersen, A. N., Pedersen, J. W., Borup, M., Brink-Kjær, A., Christiansen, L. E., and Mikkelsen, P. S.: 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 Science and Technology, 85, 1981–1998, https://doi.org/10.2166/wst.2022.059, 2022.

Pedregosa, F., Grisel, O., Weiss, R., Passos, A., Brucher, M., Varoquax, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., and Brucher, M.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, https://doi.org/10.48550/arXiv.1201.0490, 2011.

Shi, B., Catsamas, S., Kolotelo, P., Wang, M., Lintern, A., Jovanovic, D., Bach, P. M., Deletic, A., and McCarthy, D. T.: A low-cost water depth and electrical conductivity sensor for detecting inputs into urban stormwater networks, Sensors, 21, https://doi.org/10.3390/s21093056, 2021.

Sørup, H. J. D., Lerer, S. M., Arnbjerg-Nielsen, K., Mikkelsen, P. S., and Rygaard, M.: Efficiency of stormwater control measures for combined sewer retrofitting under varying rain conditions: Quantifying the Three Points Approach (3PA), Environmental Science and Policy, 63, 19–26, https://doi.org/10.1016/j.envsci.2016.05.010, 2016.

SWAN: Digital Twin Readiness Guide, UK, 2022.

“Indicator functions”, Lectures on probability theory and mathematical statistics: https://www.statlect.com/fundamentals-of-probability/indicator-functions, last access: 2 March 2022.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A. pietro, Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C. N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D. A., Hagen, D. R., Pasechnik, D. v., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G. A., Ingold, G. L., Allen, G. E., Lee, G. R., Audren, H., Probst, I., Dietrich, J. P., Silterra, J., Webber, J. T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J. L., de Miranda Cardoso, J. V., Reimer, J., Harrington, J., Rodríguez, J. L. C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N. J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P. A., Lee, P., McGibbon, R. T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., et al.: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nature Methods, 17, 261–272, https://doi.org/10.1038/s41592-019-0686-2, 2020.