Approaches for unsupervised identification of data-driven models for flow forecasting in urban drainage systems

Jóhannesson, Ari; Vezzaro, Luca; Mikkelsen, Peter Steen; Löwe, Roland

Published in:
Journal of Hydroinformatics

Link to article, DOI:
10.2166/hydro.2021.020

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

Journal of Hydroinformatics

Journal of Hydroinformatics Vol 23 No 6, 1368 doi: 10.2166/hydro.2021.020

Approaches for unsupervised identification of data-driven models for flow forecasting in urban drainage systems

Ari Jóhannesson [a,b,*], Luca Vezzaro [a,c], Peter Steen Mikkelsen [a] and Roland Löwe [a]

[a] Department of Environmental Engineering, Technical University of Denmark (DTU), Miljøvej B115, 2800 Kgs. Lyngby, Denmark
[b] Present address: Lyneborggade 29, 2300 København, Denmark
[c] Krüger A/S, Gladsaxevej 363, 2860 Søborg, Denmark
*Corresponding author. E-mail: arijoh94@gmail.com

AJ, 0000-0002-3041-6248; LV, 0000-0001-6344-7131; PSM, 0000-0003-3799-0493; RL, 0000-0002-5549-5456

ABSTRACT

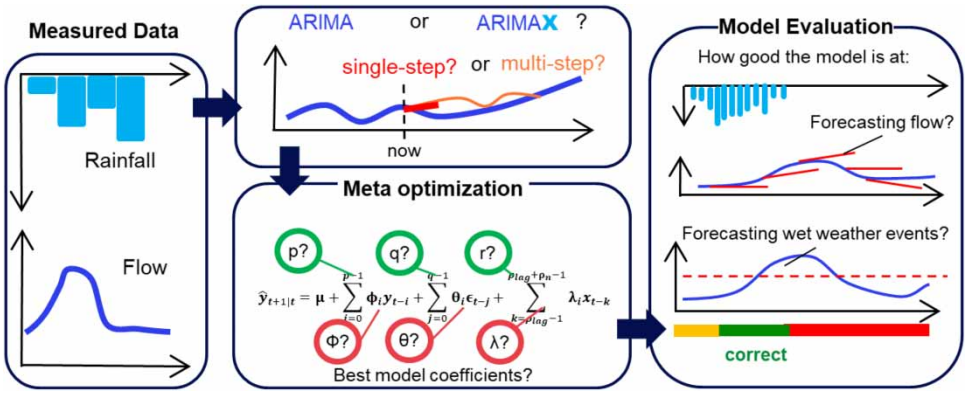
In this work, an unsupervised model selection procedure for identifying data-driven forecast models for urban drainage systems is proposed and evaluated. Specifically, we consider the case of predicting inflows to wastewater treatment plants for activating wet weather operation (aeration tank settling, ATS) using Box–Jenkins models. The model selection procedure considers different model structures and different objective functions. The hyperparameter search space is constrained based on the time of concentration in the catchment. Objective function criteria that minimize one-step-ahead as well as multi-step prediction errors are considered. Finally, we consider two criteria for unsupervised selection of the best-performing model. These measure the agreement of observed and predicted hydrographs (persistence index), as well as the binary exceedance of critical flow thresholds (critical success index (CSI)). Our work shows that forecast models can be developed in an unsupervised manner, and ATS activation is correctly forecasted in 60–90% of the events. The selected model structures reflect the physical behaviour of the catchment. Models should not be selected on operational criteria like the CSI due to a risk of overfitting. The degree to which rainfall input improves forecasts depends on the specific catchment, and the objective function criterion that should be used for coefficient estimation depends on the application context.

Key words: ARIMAX-type models, influent forecasting, time-series modelling, urban hydrology

HIGHLIGHTS

- Unsupervised model selection procedure for influent forecasting with ARIMA-type models.
- Development of hyperparameter selection criteria, i.e., physical/operational decision criterion.
- Comparisons of objective function criteria, i.e., single/multi-step forecasts.
- Evaluation of the impact of rainfall input on flow forecast quality.

GRAPHICAL ABSTRACT



INTRODUCTION

Urban drainage systems (UDS) are increasingly challenged as a result of on-going urbanization (European Commission 2016) and changing rainfall patterns (Arnbjerg-Nielsen 2012). Addressing these challenges with new infrastructural solutions (increased conveyance and storage) alone implies considerable costs and non-sustainable environmental impacts (Casal-Campos *et al.* 2015; Brudler *et al.* 2016). Green infrastructure, 'Water Sensitive Cities' and 'Sponge Cities' are, therefore, advocated, where water management is interlinked with many other services that a city provides to its citizens (Lund *et al.* 2019a; Wong *et al.* 2020). This development in urban water management occurs in parallel with the digitalization agenda, where the increasing availability of cheap sensor data enables new opportunities to exploit multifunctional aspects of urban water infrastructure through smart operation in real time (Kerkez *et al.* 2016; Sarni *et al.* 2019).

The potential of real-time operation of integrated urban wastewater infrastructures (Lund *et al.* 2019a; Wong *et al.* 2020) has been demonstrated in various contexts (García *et al.* 2015; Shishegar *et al.* 2018). These include using urban surfaces as a multifunction retention space for rainwater during heavy rainfall (Lund *et al.* 2019a), dynamically optimizing water distribution within the sewer system to minimize emissions to the environment (Fradet *et al.* 2011; Langeveld *et al.* 2013; Löwe *et al.* 2016), reducing the energy demand from sewer operations (Kroll *et al.* 2018) while maximizing the use of electricity from renewable sources (Stentoft *et al.* 2020), and increasing the treatment capacity of wastewater treatment plants (WWTPs) during and after wet weather events (e.g. aeration tank settling, ATS; Sharma *et al.* 2013). All of these applications either require or can be improved by the availability of a forecast of expected sewer flows, typically over forecast horizons ranging from 30 min up to 24 h.

Forecasting flows in UDS corresponds to translating observations and forecasts of rainfall into expected flows at specific locations, typically where the actuators used by real-time controls are placed. As described in García *et al.* (2015), UDS models can be subdivided into simulation- and control-oriented models. The first category includes detailed hydraulic models such as MIKE URBAN, SWMM, or Infoworks (so-called white-box models). Given their level of detail, these models are frequently considered too tedious and too slow to be applied for real-time operations, where decisions need to be taken within seconds to minutes. Control-oriented models include simplified conceptual hydrological models and data-driven models. Simplified conceptual models ensure important reductions in simulation times while still incorporating an understanding of the physical system into the model equations. These models have been applied both in a deterministic form (Vanrolleghem *et al.* 2005; Wolfs *et al.* 2013; Thrysøe *et al.* 2019) and in combination with data assimilation algorithms that enable the exploitation of flow sensor information to improve forecasts (Breinholt *et al.* 2011; Bach *et al.* 2016; Löwe *et al.* 2016; Lund *et al.* 2019b).

Data-driven models are a class of models without physical interpretation but much reduced computational requirements. These models are particularly useful to generate accurate forecasts in locations where sensor data are available. Their speed makes them an attractive choice in automated model selection procedures, where many models need to be tested. In urban hydrology, Box–Jenkins time-series models were used in a number of studies to forecast sewer flows and inflows to WWTPs (Tan *et al.* 1991; Pleau *et al.* 2005; Boyd *et al.* 2019). Zhang *et al.* (2018) used long short-term memory neural networks to predict inflow to WWTPs. Finally, Maleki *et al.* (2018) compared Box–Jenkins models and neural networks for predictions of WWTP inflow characteristics and found comparable performance between these model types. Data-driven models have also been used in catchment hydrology. For example, Phan & Nguyen (2020) used a combination of Box–Jenkins models and neural networks to forecast river water levels, and Valipour *et al.* (2013) applied Box–Jenkins models and neural networks to forecast reservoir inflows.

The aim of this work is to develop a procedure for the automated (or unsupervised) identification of data-driven models for forecasting flows in the UDS and specifically inflow to WWTPs. The purpose of the deveoped models is to control the activation of wet weather operation (ATS, see the section 'Data and catchments') at WWTPs. We choose to implement the model development procedure using Box–Jenkins models because they are widely used and readily available in the mathematical modelling software. The developed principles, however, can be transferable to other modelling approaches. Compared to existing procedures for the data-driven model development, we provide a number of major improvements:

1. A new approach to restrict the search space for rainfall input based on physical system understanding. Indeed, most of the existing studies in hydrology (Tan *et al.* 1991; Pleau *et al.* 2005; Boyd *et al.* 2019) select the inputs and the structure of the forecast model manually based on expert knowledge. For Box–Jenkins models, an inspection of autocorrelation functions would typically be performed (Box *et al.* 2008). In an unsupervised setting, this approach needs to be replaced by systematic

tests of different model structures. Box–Jenkins models traditionally offer three hyperparameters that can be used to systematically traverse the modelling space (Box *et al.* 2008; Hyndman & Khandakar 2008). Thus, when considering rainfall as an input to the model, the search space increases substantially and models may be difficult to identify. Therefore, we develop a new approach to restrict the search space for rainfall input based on physical system understanding.

2. A set of performance criteria to perfrom unsupervised model evaluation based on the intended model application. Criteria such as AIC that balance improved model fit on the training dataset against an increase in the number of model parameters are frequently used for Box–Jenkins models (Hyndman & Khandakar 2008) and have also been used in hydrology (Phan & Nguyen 2020). However, such criteria are not necessarily optimal for applications involving multi-step forecasting (Löwe *et al.* 2014). In general, different applications are dominated by different parts of flow hydrographs. For example, control schemes that minimize flooding require accurate forecasts of peak flows, while schemes for operating treatment plant inflow require accurate forecasts of water volume over some hours. Therefore, forecast models for different applications need to be selected based on different criteria. In this study, we focus on two criteria that, respectively, aim at reproducing the observed flow hydrograph and at maximizing the operational performance in the specific application, i.e., the frequency of how often ATS is activated correctly.
3. A comparison of different model error measures used to evaluate model performance. Given a model structure, the model coefficients (e.g., the coefficients in a time-series model, or the weights in a neural network) of the model need to be tuned such that the flow forecast is optimized for the specific case. Data-driven models come with standardized tuning criteria, such as maximum likelihood estimation, or the optimization of mean squared error or entropy (Tan *et al.* 1991; Boyd *et al.* 2019; Phan & Nguyen 2020). Similar to the previous point, the choice of tuning criterion, however, depends on the specific application. In particular, when developing a forecast model for flows in the UDS, the widely applied tuning procedures based on one-step-ahead predictions may not be robust, because sensor issues, pumping operation, and semi-random behaviour in water consumption cause noise with varying frequencies (Löwe *et al.* 2014). This work compares the effect on model calibration when using traditional error measures based on single-step forecasts or average forecast error over multiple lead times is minimized.
4. A model evaluation is based on performance indicators that are relevant for the final users of the model. Specifically, we evaluate the performance of our forecast models in terms of frequency of activation of wet weather operation (ATS) at a WWTP in Copenhagen, Denmark. The operation can, therefore, better relate to the forecast performance compared to when using traditional error measures based on residuals, such as SSE or RMSE, that were applied in similar studies (Boyd *et al.* 2019).

MATERIALS AND METHODS

In this section, we first provide an overview of the considered catchments, the decision-making problem, and the considered datasets. Subsequently, we provide details on the automated model selection procedure. In the end, a brief summary of the technical implementation is provided.

Data and catchments

Catchments and forecasting problems

In this work, models were fitted and evaluated on data from two locations in the Damhusåen catchment in Copenhagen, Denmark (Figure 1): at an upstream location (Dæmningen) and at a downstream location at the WWTP inlet (Damhusåen). Dæmningen is used as a wet weather control for Damhusåen WWTP, as it gives a warning of about 30 min to the plant operator (Breinholt & Sharma 2010). This transport time is confirmed by cross-correlation (see Figure S1 in Supplementary Material). The hydrographs at the two locations vary strongly. At the upstream location, the hydrographs are relatively smooth, driven mainly by gravity flow in pipe systems in suburban areas. At the downstream location, the hydrographs are more erratic due to inflows from dense city areas with fast response time and locally controlled 'on-off' pumps.

Historical rainfall observations were available in the form of C-band radar rainfall observations. The radar is located in Stevns, approximately 50 km south-west from Copenhagen. Data were gathered from the Danish Meteorological Institute (DMI). Historical forecast data were not available. Models were, therefore, fitted with 'perfect rainfall forecasts' (ex-post hindcasting) as well as entirely without rainfall input. A realistic forecasting performance can be expected to lie between these extremes.

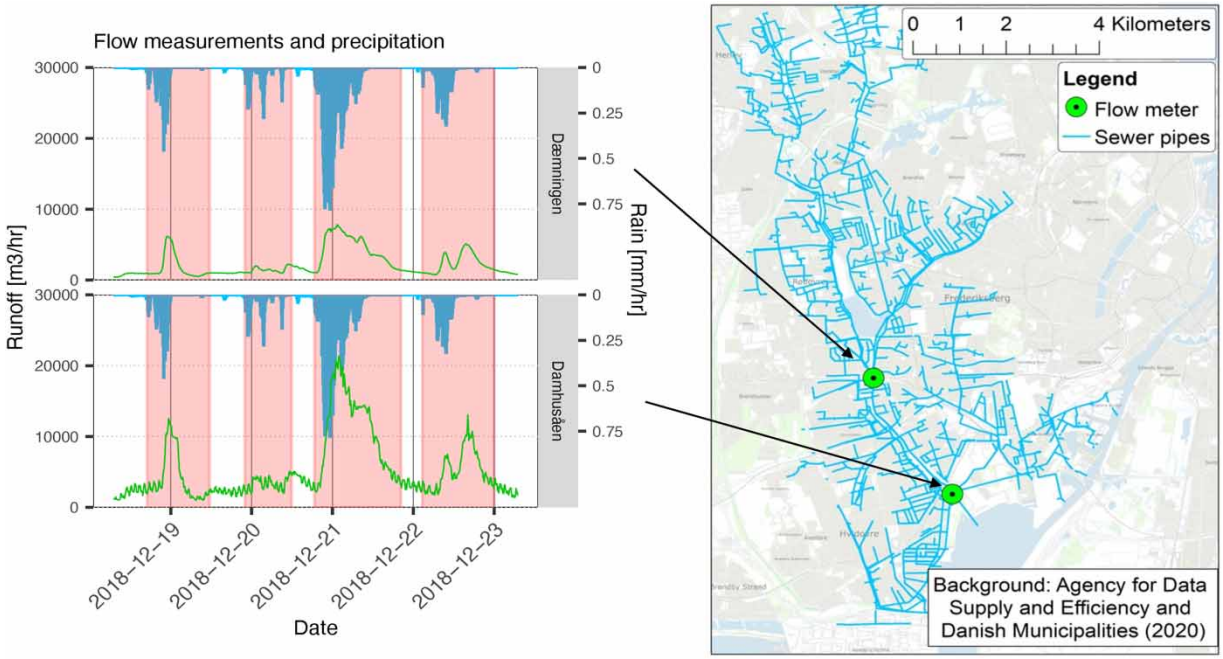


Figure 1 | The two studied flow meters are located in the Damhusåen catchment, Copenhagen. One flow meter is located upstream (Dæmningen) and one downstream at the WWTP inlet (Damhusåen). The left panel illustrates precipitation (blue) and runoff (green) in the two catchments. Valid rain events that are used for training and evaluating models are highlighted with red regions. The catchment area and the location of the flow meters are spatially visualized in the right panel. Please refer to the online version of this paper to see this figure in colour: http://dx.doi.10.2166/hydro.2021.020.

The aim of the forecast models is to support the WWTP in initiating a process that can allow the WWTP to handle larger inflows than in normal operation, thus minimizing the risk of bypass and combined sewer overflow (CSO). This process is called ATS because the sludge is allowed to settle in the aeration tanks (Sharma *et al.* 2013). ATS increases the inflow capacity of WWTPs but decreases their treatment efficiency. This creates a decision-making problem where failure to activate ATS leads to unnecessary risk of CSO (because the inflow to the WWTP remains limited), while unnecessary activation of ATS leads to unwanted increased pollutant concentrations in the WWTP outflow (due to the reduced treatment efficiency). In practice, operators apply thresholds on the rate of inflow to the WWTP to decide whether ATS should be activated. The process requires between 30 and 45 min to activate and the threshold exceedance, therefore, needs to be forecasted with sufficient lead time.

Data

Data used in this study consist of both flow measurements and radar rainfall data that range from August 16th, 2017 to the end of December 31st, 2018. Flow measurements had a 2-min resolution, while rainfall data had a 10-min resolution. To synchronize these two temporal resolutions, the flow measurements were averaged to 10 min resolution.

Models were trained and evaluated only on rain events that are relevant for UDS operation. To identify these events, a rain intensity threshold of 0.01 mm/min was used to identify the start and stop times of specific rain events, i.e., when the intensity surpasses and falls below the threshold. This threshold was selected by assessing whether a reasonable selection of rain events was obtained in a visual inspection of the time series. Similar thresholds have been recommended in Löwe *et al.* (2016) and DWA (2006).

Rainfall events closer than 12 h to each other were merged to allow for the emptying of basins in the UDS and to return to dry-weather operation before the start of the next event. This event separation period was recommended by the operators of the system and validated by manually inspecting the flow time series. Figure 1 provides an example of a flow time series for a 5-day period.

Anomalies such as flatlines in inflow measurements and missing values were removed from the datasets. Around 3 and 9% of inflow and rainfall data, respectively, consisted of missing values. After pretreatment, all data were normalized by subtracting the lowest value from all data points in the time series and dividing them by the difference between the maximum and minimum values.

For an unbiased validation of the model fit, the datasets were split up into a training set and a validation set. The training set ranged from August 16th, 2017 to December 31st, 2017, while the validation set was the whole year of 2018. The training set contained 61 rain events covering both summer and winter periods. A training period of 4.5 months was considered a reasonable trade-off between the reducing computational expense and ensuring that the relevant hydrological variations are present in the training data. Subsequently, the model was validated on 105 events from an entire year to verify model performance across seasons. Training data of similar or shorter size has been used previously in previous work (Tan *et al.* 1991; Löwe *et al.* 2014).

Framework for unsupervised development of WWTP inflow forecast models

Hyperparameter search space in Box and Jenkins time-series models

We employed ARIMAX-type models for forecasting flow (Box *et al.* 2008; Madsen 2008). In their general form, these are illustrated in Equation (1), where $\hat{y}_{t+1|t}$ is the flow value to be forecasted, assuming that observations are available until time step t. y_t is the flow observation for time step t, $\epsilon_t = y_t - \hat{y}_{t|t-1}$ is the one-step-ahead forecast error of the model observed at time step t, and x_t is the rainfall observation at time step t. The equation has four terms, the mean μ, an autoregressive (AR) term with parameter ϕ, a moving average term (MA) with parameter θ, and an exogenous input term (X) with parameter λ. The parameters (sometimes referred to as coefficients) μ, ϕ, θ, and λ are dimensionless and need to be estimated. The model orders p, q, and r can be seen as hyperparameters that define the length of the polynomials. An additional model parameter d describes the order of differencing (Madsen 2008) of the flow time series, which is a process frequently employed to remove non-stationarity from the series.

$$\hat{y}_{t+1|t} = \mu + \sum_{i=0}^{p-1} \phi_i y_{t-i} + \sum_{j=0}^{q-1} \theta_j \epsilon_{t-j} + \sum_{k=0}^{r-1} \lambda_k x_{t-k} \tag{1}$$

There is no theoretical limitation on the values of p, q, and r, and the models can, therefore, in principle contain an infinite number of parameters. In practice, most processes can, however, be described with p and q below 10. The rainfall inputs x_{t-k} that should be included in the model will depend on the time delay between rainfall and runoff in the specific catchment, as well as on the shape of the typical rainfall–runoff hydrograph. Larger time delays imply that x_{t-k} for small time lags k are not relevant to include in the model. In addition, flow from catchments that cover a larger area will be sensitive to rainfall from both the more distant and the near past, while the reaction from smaller catchments can be characterized by a narrow rainfall interval. We exploited these effects by introducing hyperparameters that, depending on the characteristics of a catchment, define the minimal value ρ_{lag} that should be considered for the time lag k, and the number ρ_n of rainfall input terms x_{t-k} that should be included in the model. This approach limits the number of coefficients that need to be estimated for the rainfall inputs to a window of size ρ_n that covers the time period that is relevant for the catchment and enables the automated search for optimal model structures based on these hyperparameters with a limited range of values. Equation (2) summarizes the structure of the resulting model.

$$\hat{y}_{t+1|t} = \mu + \sum_{i=0}^{p-1} \phi_i y_{t-i} + \sum_{j=0}^{q-1} \theta_i \epsilon_{t-j} + \sum_{k=\rho_{\text{lag}}-1}^{\rho_{\text{lag}}+\rho_n-1} \lambda_i x_{t-k} \tag{2}$$

Criteria for estimating model parameters (coefficients)

Traditionally, the coefficients of ARIMA-type models are calibrated by minimizing the single-step residuals, which can be done by using the sum-of-squares objective function (Equation (3)).

$$SSE(y, \hat{y}) = \sum_{t=1}^{n} (y_{t+1} - \hat{y}_{t+1|t})^2 \tag{3}$$

where y_t and $\hat{y}_t$ are the observations and predictions at the time point t in a time series of length n.

For estimating the model coefficients of a multi-step forecast, a scoring criterion SC_t presented in Löwe *et al.* (2014) was used, which takes a weighted average over several forecast horizons ($h = 1, 2, \ldots, k$) (Equation (4)). The maximum forecast

horizon was set to $k = 10$. This criterion encourages forecast models that reflect the physical system behaviour by penalizing deviations from the hydrograph on longer forecast horizons.

$$SC = \sum_{t=1}^{n} \frac{1}{\sum_{h=1}^{k} (k-h+1)} \left(\sum_{h=1}^{k} (k-h+1)(y_{t+h} - \hat{y}_{t+h|t})^2 \right) \tag{4}$$

The implementation of the criterion in Equation (4) requires that forecasts covering multiple forecast horizons are generated for each time step during model training. This approach cannot be implemented efficiently using standard time-series modelling packages in R. We, therefore, implemented an approach based on matrix multiplication that exploits that ARIMA models in multi-step predictions re-use information from previous forecast horizons in an iterative manner. This approach is documented in the Supplementary Material (S3, Figure S3).

The optimization method used to minimize the objective functions was the dynamically dimensioned search algorithm (DDS) presented in Tolson & Shoemaker (2017). The DDS search is a heuristic algorithm, specifically designed to solve computationally expensive optimization problems with many parameters. The search starts as a global search where all model parameters are randomly varied but then narrows down to a local search, eventually varying only one parameter. All parameters were constrained to the interval −5 and 5, and starting parameters were selected randomly from a Gaussian distribution with a mean of 0 and a variance of 1. DDS requires the user to define the number of objective function evaluations, which was set to 2,500 for all models.

Criteria for selecting the best-performing model

We considered two criteria for evaluating which combinations of hyperparameters of the model should be selected and for evaluating model performance in general. The first criterion is the persistence index (PI) (Bennett *et al.* 2013), which measures whether the model can reproduce the observed flow hydrograph by comparing the forecast error against a benchmark forecast defined as the last available flow observation:

$$PI_h = 1 - \frac{\frac{1}{n}\sum_{t=1}^{n} (y_{t+h} - \hat{y}_{t+h|t})^2}{\frac{1}{n}\sum_{t=1}^{n} (y_{t+h} - y_t)^2} \tag{5}$$

where $\hat{y}$ represents a prediction of inflow y at time step t, in a time series of length n, and h represents the forecast horizon in time steps. A perfect model yields $PI = 1$, while $PI < 0$ suggests that the forecast model cannot outperform the constant value benchmark.

The second criterion focuses on the application of the forecast model to the decision problem of activating wet weather operation of the WWTP (ATS). The threshold for ATS activation was defined together with the operators of the WWTP and set to $y_t = 5{,}000\,\mathrm{m}^3/\mathrm{h}$ at both locations. As unnecessary activation and failure to activate ATS in due time leads to unwanted pollution of the environment, evaluating the frequency of these undesirable situations enables the assessment of the model performance in an operational setting. This assessment was implemented as a contingency table with three states that were defined together with the operators (Courdent *et al.* 2017, 2018):

1. *Correct forecast (TP, true positive)*: If ATS activation was predicted within a 75 min interval (from 60 min before measured flow exceeded the threshold to 15 min after).
2. *False alarm (FP, false positive)*: ATS event forecasted, but the measured flow did not exceed the threshold. This case leads to suboptimal WWTP performance with a worsening of its pollutant removal efficiency.
3. *Missed (FN, false negative)*: if ATS activation was predicted more than 15 min after the measured flow exceeded the threshold or if ATS activation was not predicted at all. This case can negatively affect the WWTP performance by causing discharge of partially untreated wastewater (bypass), reduction of the efficiency of biological removal processes, loss of active biomass, and consequent long-term worsening of the WWTP operational capability.

The time windows used in the classification of the ATS events are schematized in Figure 2 and were defined in collaboration with the plant operator. Forecasting an ATS too early results in FP.

The critical success index (CSI) (Bennett *et al.* 2013) was used to measure how often ATS events were forecasted correctly and would ideally take a value of 1:

$$CSI = \frac{TP}{TP + FP + FN} \tag{6}$$

Both model selection criteria (Equations (5) and (6)) can be evaluated on different forecast horizons. Selecting models on longer forecast horizons is likely to encourage more physically realistic behaviour of the forecasts because high forecast skill cannot be obtained without generating multi-step flow predictions that follow the general shape of the observed hydrograph (Löwe *et al.* 2014). On the other hand, this approach might compromise forecast performance on shorter horizons. To identify which trade-off is reasonable for our problem, we computed PI and CSI for lead times of 30, 60, and 90 min as well as the average over these horizons (AVG-PI). The PI criterion generally increases with a larger forecasting horizon, and selecting a model based on the average PI across different forecasting horizons is done to help lower the risk of non-identifiability. Subsequently, we analysed the performance of models selected based on different lead times.

Technical implementation of model selection

ARIMA and ARIMAX-type models were identified by performing a meta-model search (meta-optimization) on a pre-defined hyperparameter search space. The search was composed of three main steps:

1. Selecting various sets of hyperparameters (p, d, q, ρ_n, ρ_{lag}) from a pre-defined hyperparameter search space, generating various 'hyper-models'.
2. Estimating coefficients of the hypermodels by minimizing an objective function (SSE or SC).
3. Evaluating the calibrated models on a validation dataset based on two evaluation criteria (PI and CSI).

The hyperparameter search space was defined as $p = \{0, 1, 2, 3, 4, 5, 6, 7, 8\}$, $d = \{0, 1\}$, $q = \{0, 1, 2, 3, 4, 5, 6, 7, 8\}$, $\rho_n = \{2, 4, 6, 8, 10\}$, and $\rho_{\text{lag}} = \{5, 10, 15\}$. The process was carried out for models with (ARIMAX) and without (ARIMA) rainfall input, as well as for different objective function criteria (single/multi-step forecasts) to allow for a systematic comparison of model types. The models were implemented in R and trained in a high-performance cluster environment where up to 120 hyperparameter combinations could be evaluated in parallel. Calibration for each model generally took under 10 min. In total, 2,592 models were calibrated for each catchment, resulting in a total computation time of approximately 100 CPU h.

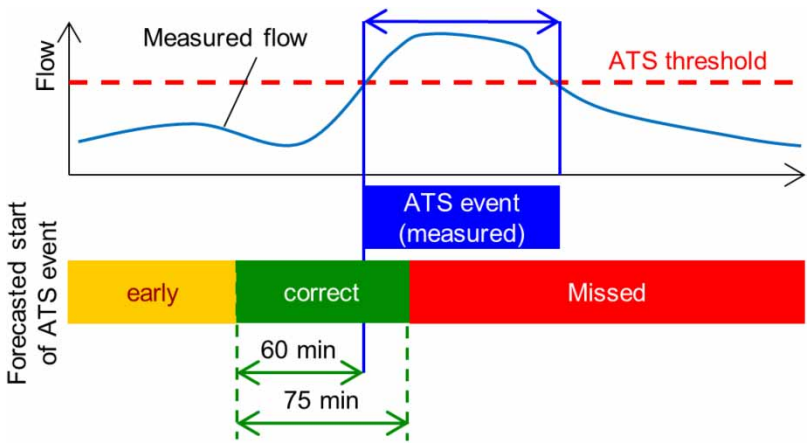


Figure 2 | Schematization of the time-thresholds used to classify ATS events, starting from the moment when the measured flow exceeds the ATS threshold (5,000 m^3/h).

RESULTS AND DISCUSSION

In this section, the effect of different model configurations on forecast performance is evaluated. The first subsection provides a visual overview of the effect of different procedures for selecting and training forecast models. The subsequent sections then go in-depth with each of the analysed configurations. Sections S4–S6 in the Supplementary Material provide tables with a detailed overview of model structures and performance when applying different criteria for the estimation and selection of forecast models.

Physical forecast behaviour for different models

Figure 3 illustrates multi-step forecasts for a single rain event in the two considered catchments. Panel (a) compares the best-performing models with and without external rainfall input. Panel (b) compares the best-performing models that were trained using single and multi-step criteria (SSE and SC), and panel (c) compares models that were selected considering the persistence index and critical success index (PI and CSI), respectively (PI was applied as a selection criterion in panels (a) and (b)). All models shown in Figure 3 were selected on a 90-min forecast horizon.

Figure 3(a) shows that models without rainfall input can behave similar to models with rainfall input, but cannot very well forecast the increasing limb of the hydrograph. This can be clearly seen from the hydrographs from Damhusåen (Figure 3(a)). Where precipitation is increasing (for example, around 00:00–06:00), the ARIMAX models can quite successfully forecast the increasing runoff, while the ARIMA models remain more stable. From Figure 3(b), it can be seen that models trained using a multi-step criterion (Equation (4)) seem to better reproduce the physical behaviour of the system on longer forecast horizons. As before, this is more obvious for the catchment of Damhusåen. From Figure 3(c), it is clear that if models are selected based on the CSI score, there is a high risk of overfitting and forecasts not reflecting the physical behaviour of the system. This is clearly illustrated as models selected on the CSI are just sharp peaks while selecting on PI results in a model behaviour that much more closely resembles the physical behaviour and shape of the hydrograph.

Structural differences between models selected on different criteria were also identified. In Sections S5 and S6 in the Supplementary Material, this is shown by selecting the 10 best-performing models based on both error metrics and various forecast horizons. When selected on PI, the best-performing models have a higher number of MA parameters (q) compared to the number of AR parameters (p) which is usually quite low. Additionally, most of the models selected on PI are differentiated, although some 60-min forecast models for Damhusåen are undifferentiated. When selected on the CSI, the only clear structure that can be seen is that most of the models are not differentiated, resulting in non-stationary behaviour.

Value of rainfall input for forecasting

Figure 4 shows the average PI and CSI skill scores of the top-5 performing models without (ARIMA) and with (ARIMAX) rainfall input in the Dæmningen and Damhusåen catchments. The figure illustrates that in all cases, models using rainfall input are superior when compared to the ones that use no rainfall input. The difference is more significant in Damhusåen compared to Dæmningen, as the Damhusåen catchment exhibits more pronounced peaks and a generally less smooth behaviour of the flow series, caused partly by fast rainfall–runoff from dense city areas (cf. Figures 1 and 3).

Considering the model structures with the highest forecast performance (Supplementary Material, Sections S5 and S6), we see that the selected hyperparameters (ρ_{lag} and ρ_n) reflect the behaviour of the catchment. The pronounced peaks in Damhusåen are mostly described by short lag times (ρ_{lag} in the order of 5) and narrow window sizes (ρ_n in the order of 2–4), while slightly longer lag times (ρ_{lag} in the order of 5–15) and wider window sizes for the rainfall input (ρ_n in the order of 2–10) are favoured in the Dæmningen catchment.

Effects of single- vs. multi-step parameter calibration

Figure 5 compares models fitted on the single-step objective function criterion (SSE) to models fitted on the multi-step objective function criterion (SC). Considering selecting on PI skill scores in panel (a), the traditional single-step objective function criterion generally results in higher score values for shorter forecasting horizons (30 min) while for longer forecasting horizons, a multi-step objective function criterion is preferred. In addition, PI values in the Damhusåen catchment are generally very low for short forecast horizons, while they are more balanced in the Dæmningen catchment. This implies that the model selection procedure favours multi-step models in the Damhusåen catchment where 7 of the 10 best-performing models were trained using the SC_t criterion, while the opposite result is obtained in the Dæmningen catchment (Supplementary Material, Section S4 and Figure S4).

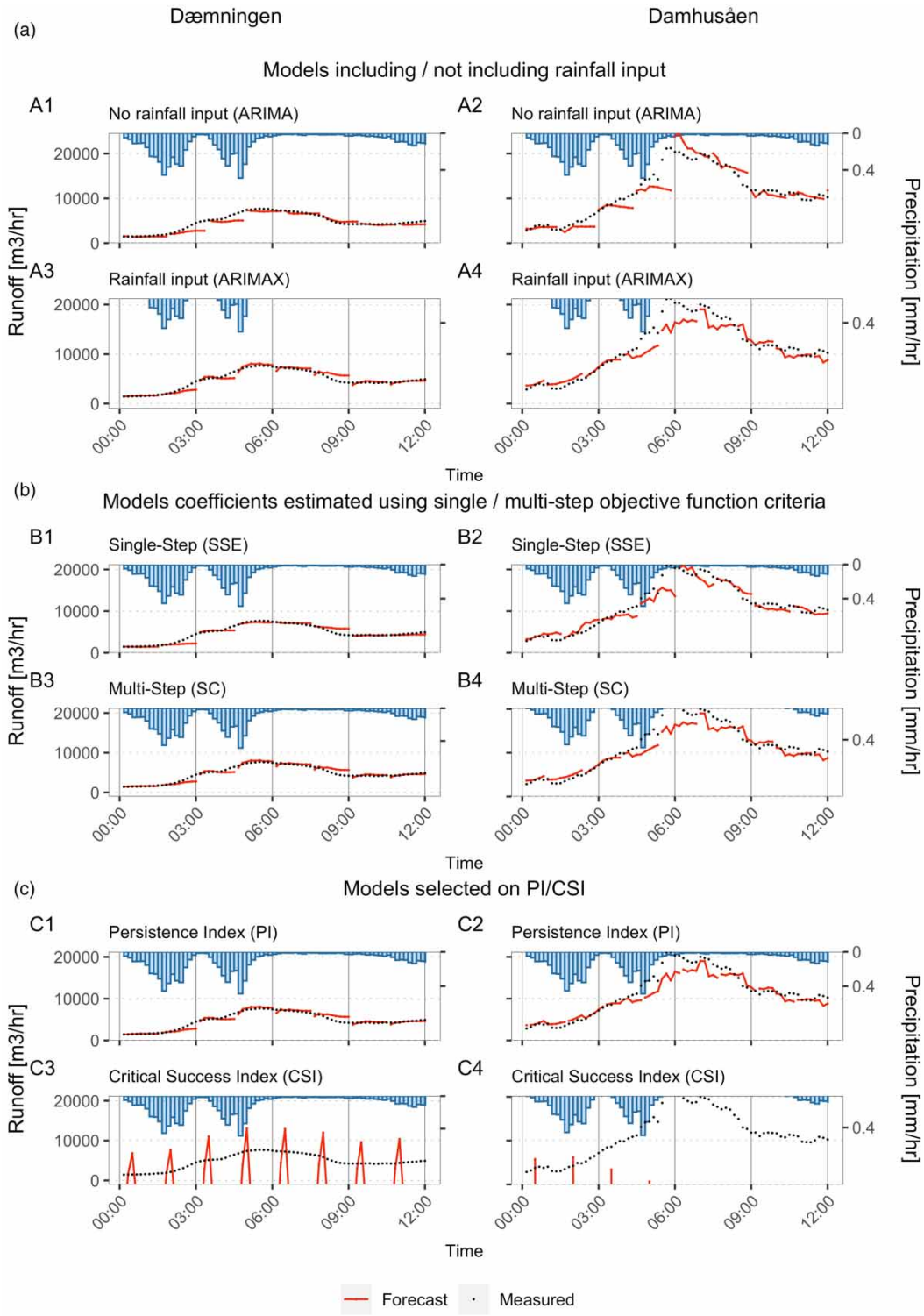


Figure 3 | Forecasts for both locations (Dæmningen; left, Damhusåen; right) visualized when selected on different criteria. Panel (a) compares models selected with and without rainfall as an external regressor (ARIMA/ARIMAX) and panel (b) compares models estimated using single- and multi-step criteria (SSE/SC); both panels' models were selected based on the PI. Panel (c) compares models selected on the PI and the CSI. All models were selected on a 90-min forecasting horizon. Axes are re-used for each panel. From the figure, it can be seen that ARIMAX models are capable of forecasting the increasing limb of the hydrograph and that multi-step models (SC) can better capture the physical behaviour of the system for longer horizons. Finally, it can be seen that selecting on the CSI risks overfitting where the model does not reflect the physical behaviour of the system.

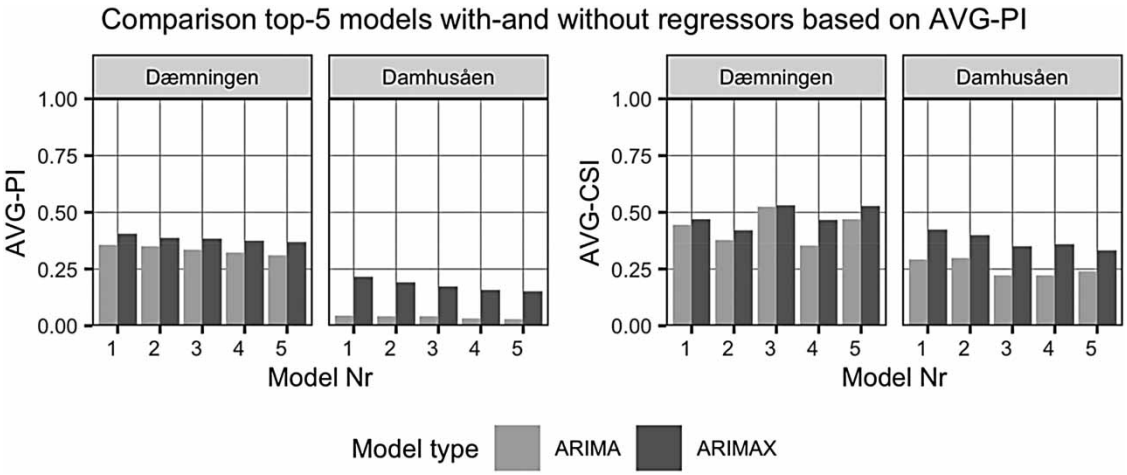


Figure 4 | The top-5 models based on the average persistence index over various horizons (AVG-PI) and average critical success index (AVG-CSI) over various horizons with and without rainfall input. Models with rainfall input (ARIMAX) achieve better scores than models without rainfall (ARIMA).

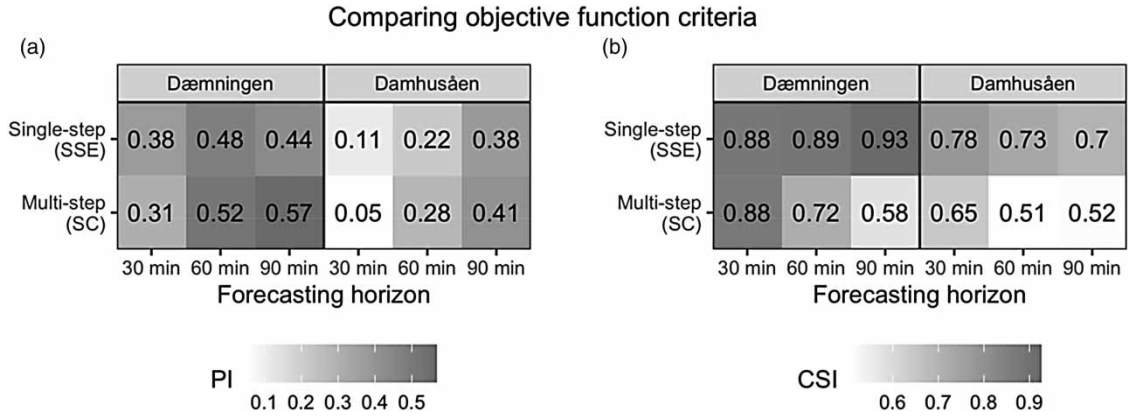


Figure 5 | Model performances of single/multi-step (SSE/SC) estimated models compared for different forecast horizons. Panel (a) shows the PI skill obtained for different forecast horizons when training forecast models using the single-step and multi-step objective function criteria, respectively. Panel (b) shows the corresponding CSI values. Darker colours correspond to higher score values. Please refer to the online version of this paper to see this figure in colour: http://dx.doi.10.2166/hydro.2021.020.

From Figure 5(b), where models are selected on CSI (in contrast to AVG-PI in Figure 4), it can be observed that the CSI is lower for models fitted with the multi-step objective function criterion (SC) compared to models fitted with the single-step objective function criterion (SSE). This is explained by the fact that the multi-step objective function criterion prevents overfitting and punishes models that do not follow the shape of the hydrograph (Figure 3(b)). The single-step models are more prone to overfit, which is clear from the reduced PI values for longer lead times.

Efficient model selection criteria

Figure 6 shows how models selected on PI estimated for different forecast horizons (30, 60, and 90 min, AVG-PI) perform on other forecast horizons. Each model is selected on a specific forecasting horizon and shown as a certain point shape, and a line connects its performance for different forecasting horizons.

We have not considered models selected based on the CSI criterion, as the previous sections clearly illustrated that this criterion favours overfitted models that perform strongly for this specific criterion and a specific forecast horizon. Performance strongly degrades when considering other forecast horizons or evaluation criteria. In addition, no clear tendencies in the selected model hyperparameters can be identified if the model selection is performed based on CSI (Supplementary Material, Section S4). This underlines that the criterion leads to models that do not reflect the physical system behaviour.

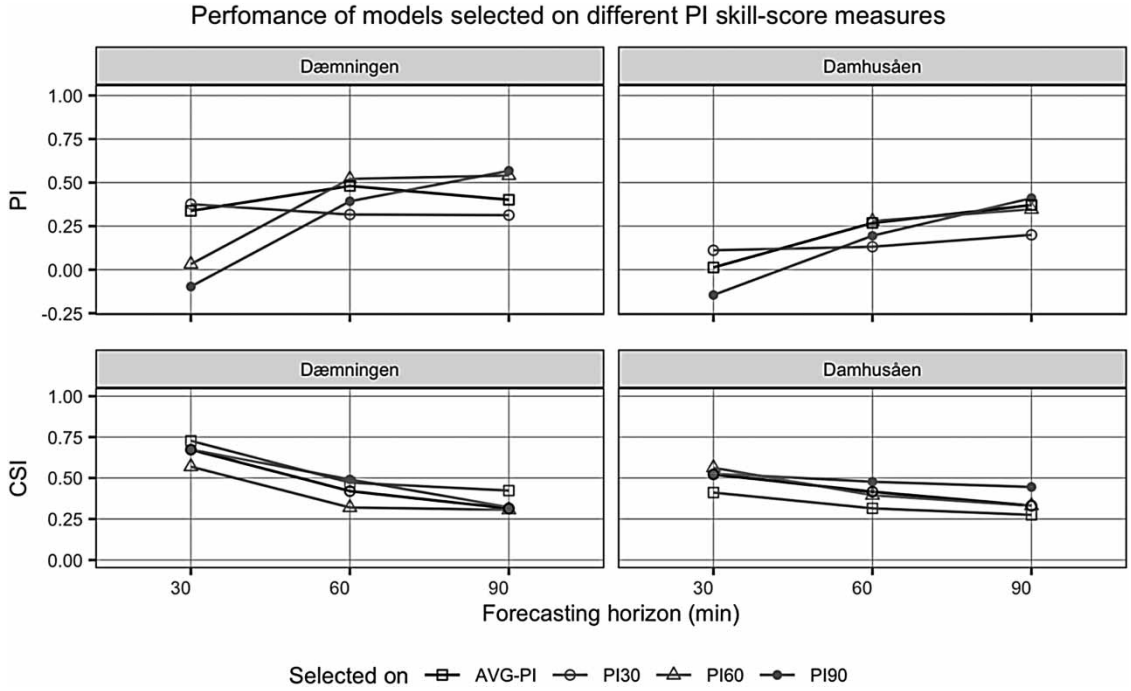


Figure 6 | PI and SCI scores for different forecast horizons of models that were selected using the PI for different forecast horizons (an average over the forecast horizons). The PI skill score increases with longer horizons due to the error caused by the benchmark being greater. To prevent this overfitting, an average of all PI scores should be considered. Concerning the CSI, selecting on the PI for shorter horizon results in a better CSI score.

We observe a general trend of the PI skill score to increase with the forecast horizon. The reason for this behaviour is that the error of the benchmark forecast (last known observation) increases for increasing forecast horizons. As a result, models selected on PI60 or PI90 may be overfitted and perform poorly on shorter forecast horizons (Supplementary Material, Sections S2 and S3). Selecting models based on their average PI for different forecast horizons (AVG-PI) reduces the overfit and leads to more robust models with robust CSI values.

DISCUSSION

Recommendations for the automated identification of flow forecast models

Our results suggest that it is possible to identify flow forecast models in an automated manner. Although models were trained on rain events that differed in size, no significant deviation in forecasting performance was observed (Supplementary Material, Section 1 and Figure S1). In terms of the criterion for parameter estimation, Figure 5 suggests that the most suitable approach may depend on the specific application. Generally, the multi-step estimation criterion (SC) favours models that better reflect the physical behaviour of the system, but the CSI of the forecast models will, in some cases, be inferior to those estimated using the traditional single-step procedure (SSE). The SSE criterion allows for greater flexibility of the predicted hydrographs. This may or may not be desirable, as the forecasts achieve higher skill in an operational sense. However, the forecasted hydrograph over multiple time steps will not necessarily reflect the shape of the observed hydrograph.

The selection of the best-performing model should, in an unsupervised procedure, use a criterion that is based on measuring the deviation from the hydrograph (such as PI) rather than an operational decision criterion such as CSI, as the latter carries a high risk of overfitting. The PI criterion generally increases as a function of the forecast horizon, as the performance of the constant value benchmark decreases. This implies a risk of non-identifiability if the forecast model is selected based on PI scores obtained for large forecast horizons. We were able to mitigate this problem by selecting models based on their average performance for different forecast horizons.

The hyperparameters of the selected models exhibited clear trends. For instance, selected models were mostly differentiated ($d = 1$) and had low AR term (p) and high MA term (q). Additionally, rainfall input improved forecasts and was mostly centred around specific lag values, that, however, differed between the catchments. Hence, recalibration of the models can likely be done more efficiently by considering a smaller search space.

Limitations

This work focused on investigating what factors need to be considered when selecting Box–Jenkins-type models in an unsupervised manner. The work focused on flow measurements of two locations in Copenhagen, but the presented approach will likely be applicable to other catchments with similar physical characteristics.

Due to various changing factors such as changing temperature and groundwater levels, hydrological models occasionally need to be recalibrated (Troutman *et al.* 2017). Future work could investigate how often a recalibration should be performed and how much data is needed.

There are known data uncertainties, for example in the rainfall data, that will always be present in a real-world context. Radar rainfall input used in this work does not necessarily correspond to the true rainfall that is observed at the surface and information on the estimate of the uncertainty is not provided to the model. Such uncertainties imply that models that react less strongly to rainfall input will be selected and increase the uncertainty of flow forecasts. If desired, uncertainty bands for the forecasts can be generated based on the variance of model residuals (Madsen 2008). Previous work demonstrated that the uncertainty of rainfall forecasts can reduce the efficiency of control schemes for the sewer system by 20–30% (Löwe *et al.* 2016).

Outlook

We applied a simple grid search to systematically evaluate the hyperparameter search space at the cost of computational expense. Hyperparameter optimization routines would enable a reduction of the computation time required for training the models. Heuristic search algorithms such as DDS (Tolson & Shoemaker 2017) or DREAM (Vrugt 2016) could be employed. Additionally, other criteria for model selection such as hydrological signatures (Shafii & Tolson 2015) could be promising. We would, however, expect that the consideration of such criteria is subject to similar challenges as identified in our work, i.e., selecting models based on very specific hydrograph features leads to a high risk of overfitting. Finally, as rain input does not need to be limited to a temporal resolution of the flow, rainfall-based daily intensities could be investigated. This might help the model capturing seasonal variations in flows.

We employed Box–Jenkins models for forecasting, but our procedures for defining rainfall input variables, tuning the model, and evaluating model performance are generic and can be transferred directly to other data-driven forecasting approaches such as artificial neural networks (Zhang *et al.* 2018; Crotti *et al.* 2020).

CONCLUSIONS

We developed a method for unsupervised (automated) identification of data-driven models that forecast inflow to wastewater treatment facilities. Based on the results, we conclude that:

1. We can obtain reliable forecasts of WWTP inflow with models that were identified in an unsupervised manner.
2. To facilitate unsupervised model identification, the search space for precipitation variables that should be included in the forecast model can be constrained using meta-variables for the number of input variables and the time lag that should be considered.
3. There is a high risk of overfitting if best-performing models are selected based on operational criteria (e.g. the frequency of flow threshold exceedance). Model selection criteria should instead ensure that the model generates physically meaningful hydrographs.
4. Different approaches for tuning model coefficients need to be considered in initial stages of an unsupervised model selection procedure. A multi-step objective function leads to a more physically realistic behaviour of the forecasts but can lead to reduced forecast performance in terms of operational criteria. Thus, different objective functions may be optimal in different applications.
5. Precipitation as an external regressor improves the performance of the models, but the improvement can be very small if the time of concentration of a catchment is greater than the considered forecast horizon. Precipitation input may, thus, not be necessary to generate short-term forecasts (<2 h) for larger urban catchments.

6. Unsupervised model selection involves substantial computational efforts. However, once a suitable model was identified, recalibration to new observations can be performed quickly, because the hyperparameters selected in the unsupervised procedure exhibit clear trends that are linked to catchment characteristics.

ACKNOWLEDGEMENTS

We thank Carsten Thirsing and BIOFOS for the provision of flow data and the DMI for the provision of radar rainfall data. Both datasets were made available as part of the Water Smart Cities project, which was supported by Innovation Fund Denmark (grant no. 5157-00009B).

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

REFERENCES

Arnbjerg-Nielsen, K. 2012 Quantification of climate change effects on extreme precipitation used for high resolution hydrologic design. *Urban Water Journal* **9** (2), 57–65.

Bach, H., Baattrup-Pedersen, A., Holm, E. P., Jensen, P. N., Larsen, T., Ovesen, N. B., Petersen, M. L., Sand-Jensen, K. & Styczen, M. 2016 *Faglig udredning om grødeskæring i vandløb*. Available from https://mst.dk/media/114793/faglig-udredning-om-groedeskaering-i-vandloeb-d-14062016.pdf (accessed 19 October 2021).

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. *Environmental Modelling and Software* **40**, 1–20.

Box, G. E. P., Jenkins, G. M. & Reinsel, G. C. 2008 *Time Series Analysis: Forecasting and Control*. Wiley, Hoboken, NJ, USA.

Boyd, G., Na, D., Li, Z., Snowling, S., Zhang, Q. & Zhou, P. 2019 Influent forecasting for wastewater treatment plants in North America. *Sustainability (Switzerland)* **11** (9), 1764.

Breinholt, A. & Sharma, A. K. 2010 *Case Area Baseline Report: Copenhagen and Lynette Fællesskabet*.

Breinholt, A., Thordarson, F. O., Møller, J. K., Grum, M., Mikkelsen, P. S. & Madsen, H. 2011 Grey-box modelling of flow in sewer systems with state-dependent diffusion. *Environmetrics* **22** (8), 946–961.

Brudler, S., Arnbjerg-Nielsen, K., Hauschild, M. Z. & Rygaard, M. 2016 Life cycle assessment of stormwater management in the context of climate change adaptation. *Water Research* **106**, 394–404.

Casal-Campos, A., Fu, G., Butler, D. & Moore, A. 2015 An integrated environmental assessment of green and gray infrastructure strategies for robust decision making. *Environmental Science and Technology* **49** (14), 8307–8314.

Courdent, V., Grum, M. & Mikkelsen, P. 2017 A gain–loss framework based on ensemble flow forecasts to switch the urban drainage–wastewater system management towards energy optimization during dry periods. *Hydrology and Earth System Sciences* **21** (5), 2531–2544.

Courdent, V., Grum, M. & Mikkelsen, P. S. 2018 Distinguishing high and low flow domains in urban drainage systems 2 days ahead using numerical weather prediction ensembles. *Journal of Hydrology* **556**, 1013–1025.

Crotti, G., Leandro, J. & Bhola, P. K. 2020 A 2D real-time flood forecast framework based on a hybrid historical and synthetic runoff database. *Water (Switzerland)* **12** (1), 114.

DWA 2006 Available from: https://webshop.dwa.de/de/dwa-a-118-hydraulische-bemessung-3-2011.html (accessed 19 June 2021).

European Commission 2016 *Statistics on cities, towns and suburbs*. Available from https://op.europa.eu/en/publication-detail/-/publication/da0b33d3-764f-11e6-b076-01aa75ed71a1 (accessed 5 March 2021).

Fradet, O., Pleau, M. & Marcoux, C. 2011 Reducing CSOs and giving the river back to the public: innovative combined sewer overflow control and riverbanks restoration of the St. Charles River in Quebec City. *Water Science and Technology* **63** (2), 331–338.

García, L., Barreiro-Gomez, J., Escobar, E., Téllez, D., Quijano, N. & Ocampo-Martinez, C. 2015 Modeling and real-time control of urban drainage systems: a review. *Advances in Water Resources* **85**, 120–132.

Hyndman, R. J. & Khandakar, Y. 2008 Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software* **27** (3), 1–22.

Kerkez, B., Gruden, C., Lewis, M., Montestruque, L., Quigley, M., Wong, B., Bedig, A., Kertesz, R., Braun, T., Cadwalader, O., Poresky, A. & Pak, C. 2016 Smarter stormwater systems. *Environmental Science and Technology* **50** (14), 7267–7273.

Kroll, S., Fenu, A., Wambecq, T., Weemaes, M., van Impe, J. & Willems, P. 2018 Energy optimization of the urban drainage system by integrated real-time control during wet and dry weather conditions. *Urban Water Journal* **15** (4), 362–370.

Langeveld, J., Benedetti, L., de Klein, J. J. M., Nopens, I., Amerlinck, Y., van Nieuwenhuijzen, A., Flameling, T., van Zanten, O. & Weijers, S. 2013 Impact-based integrated real-time control for improvement of the Dommel River water quality. *Urban Water Journal* **10** (5), 312–329.

Löwe, R., Mikkelsen, P. S. & Madsen, H. 2014 Stochastic rainfall-runoff forecasting: parameter estimation, multi-step prediction, and evaluation of overflow risk. *Stochastic Environmental Research and Risk Assessment* **28** (3), 505–5016.

Löwe, R., Vezzaro, L., Mikkelsen, P. S., Grum, M. & Madsen, H. 2016 Probabilistic runoff volume forecasting in risk-based optimization for RTC of urban drainage systems. *Environmental Modelling & Software* **80**, 143–158.

Lund, N. S. V., Borup, M., Madsen, H., Mark, O., Arnbjerg-Nielsen, K. & Mikkelsen, P. S. 2019a Integrated stormwater inflow control for sewers and green structures in urban landscapes. *Nature Sustainability* **2** (11), 1003–1010.

Lund, N. S. V., Madsen, H., Mazzoleni, M., Solomatine, D. & Borup, M. 2019b Assimilating flow and level data into an urban drainage surrogate model for forecasting flows and overflows. *Journal of Environmental Management* **248**, 109052.

Madsen, H. 2008 *Time Series Analysis*. Chapman and Hall/CRC, Boca Raton, FL, USA.

Maleki, A., Nasseri, S., Aminabad, M. S. & Hadi, M. 2018 Comparison of ARIMA and NNAR models for forecasting water treatment plant's influent characteristics. *KSCE Journal of Civil Engineering* **22** (9), 3233–3245.

Phan, T.-T.-H. & Nguyen, X. H. 2020 Combining statistical machine learning models with ARIMA for water level forecasting: the case of the Red river. *Advances in Water Resources* **142**, 103656.

Pleau, M., Colas, H., Lavallee, P., Pelletier, G. & Bonin, R. 2005 Global optimal real-time control of the Quebec urban drainage system. *Environmental Modelling & Software* **20** (4), 401–413.

Sarni, W., White, C., Webb, R., Cross, K. & Glotzbach, R. 2019 *Digital Water – Industry Leaders Chart the Transformation Journey*.

Shafii, M. & Tolson, B. A. 2015 Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives. *Water Resources Research* **51** (5), 3796–3814.

Sharma, A. K., Guildal, T., Thomsen, H. A. R., Mikkelsen, P. S. & Jacobsen, B. N. 2013 Aeration tank settling and real time control as a tool to improve the hydraulic capacity and treatment efficiency during wet weather: results from 7 years' full-scale operational data. *Water Science and Technology* **67** (10), 2169.

Shishegar, S., Duchesne, S. & Pelletier, G. 2018 Optimization methods applied to stormwater management problems: a review. *Urban Water Journal* **15** (3), 276–286.

Stentoft, P. A., Vezzaro, L., Mikkelsen, P. S., Grum, M., Munk-Nielsen, T., Tychsen, P., Madsen, H. & Halvgaard, R. 2020 Integrated model predictive control of water resource recovery facilities and sewer systems in a smart grid: example of full-scale implementation in Kolding. *Water Science and Technology* **81** (8), 1766–1777.

Tan, P. C., Berger, C. S., Dabke, K. P. & Mein, R. G. 1991 Recursive identification and adaptive prediction of wastewater flows. *Automatica* **27** (5), 761–768.

Thrysøe, C., Arnbjerg-Nielsen, K. & Borup, M. 2019 Identifying fit-for-purpose lumped surrogate models for large urban drainage systems using GLUE. *Journal of Hydrology* **568**, 517–533.

Tolson, B. A. & Shoemaker, C. A. 2017 Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research* **43**, W01413.

Troutman, S. C., Schambach, N., Love, N. G. & Kerkez, B. 2017 An automated toolchain for the data-driven and dynamical modeling of combined sewer systems. *Water Research* **126**, 88–100. http://dx.doi.org/10.1016/j.watres.2017.08.065.

Valipour, M., Banihabib, M. E. & Behbahani, S. M. R. 2013 Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. *Journal of Hydrology* **476**, 433–441.

Vanrolleghem, P. A., Benedetti, L. & Meirlaen, J. 2005 Modelling and real-time control of the integrated urban wastewater system. *Environmental Modelling & Software* **20** (4), 427–442.

Vrugt, J. A. 2016 Markov chain Monte Carlo simulation using the DREAM software package: theory, concepts, and MATLAB implementation. *Environmental Modelling and Software* **75**, 237–316.

Wolfs, V., Villazon, M. F. & Willems, P. 2013 Development of a semi-automated model identification and calibration tool for conceptual modelling of sewer systems. *Water Science and Technology* **68** (1), 167–175.

Wong, T. H. F., Rogers, B. C. & Brown, R. R. 2020 Transforming cities through water-sensitive principles and practices. *One Earth* **3**, 436–337.

Zhang, D., Hølland, E. S., Lindholm, G. & Ratnaweera, H. 2018 Hydraulic modeling and deep learning based flow forecasting for optimizing inter catchment wastewater transfer. *Journal of Hydrology* **567**, 792–802.

First received 10 February 2021; accepted in revised form 6 October 2021. Available online 19 October 2021