Reducing uncertainties in urban drainage models by explicitly accounting for timing errors in objective functions

Ico Broekhuizen, Günther Leonhardt and Maria Viklander

Department of Civil, Environmental and Natural Resources Engineering, Luleå University of Technology, Luleå, Sweden

ABSTRACT

Traditional hydrological objective functions may penalize models that reproduce hydrograph shapes well, but with some shift in time; especially for urban catchments with a fast hydrological response. Hydrograph timing is not always critical, so this paper investigates alternative objective functions (based on the Hydrograph Matching Algorithm) that try to mimic visual hydrograph comparison. A modified version of the Generalized Likelihood Uncertainty Estimation is proposed to compare regular objective functions with those that account for timing errors. This is applied to 2-year calibration and validation data sets from an urban catchment. Results show that such objective functions provide equally reliable model predictions (they envelop the same fraction of observations), but with more precision, i.e. smaller estimated uncertainty of model predictions. Additionally, identifiability of some model parameters improved. Therefore objective functions based on the Hydrograph Matching Algorithm can be useful to reduce uncertainties in urban drainage modelling.

1. Introduction

Calibration of urban hydrodynamic models requires an objective function, i.e. a mathematical way of quantifying the performance of a proposed set of parameters values by comparing the simulated values it produces to corresponding observed values (e.g. Bennett et al. 2013). The choice of objective function determines how well a certain model fits with the observations, and as such it will influence the outcomes of urban drainage modelling studies (Deletic et al. 2012). In principle, any output or state variable of the model for which corresponding measurements are available can be included in the objective function. Types of data that have been used include sewer flow rates, sewer water depths (Tscheikner-Gratl et al. 2016), duration and volume of combined sewer overflow (e.g. Kleidorfer et al. 2009; Thordahl et al. 2008) or soil water content (Easton et al. 2007). Sewer flow rates (i.e. hydrographs) are the most commonly used and will therefore be the focus of this paper.

Different objective functions emphasize different parts or properties of the hydrograph (Bennett et al. 2013) and can therefore lead to different results in calibration. The implications of this general observation have not been studied extensively within the urban drainage context. Barco, Wong, and Stenstrom (2008) demonstrated trade-offs between peak flow and total volume performance when combining these in a weighted objective function. They found that adding instantaneous flow rates to this objective function was not useful for improving the model performance in terms of peak flow error and flow volume error, because the errors in instantaneous flow rates were mainly sensitive to timing which was not affected by the calibration parameters. In the same vein, Zhang and Li (2015) demonstrated how using different objective functions affected which model parameters could be identified.

As touched upon by Barco, Wong, and Stenstrom (2008), one important aspect in which objective functions differ is how sensitive they are to errors in timing of the simulated hydrograph compared to the observed one. Timing differences may be caused by drift between the clocks of different data loggers, moving storms reaching the rain gauge and (different parts of) the catchment at different times, or limitations of the model structure and setup. The human eye has a good ability to recognize similar patterns in different hydrographs that traditional objective functions may not be able to. Measures such as the total runoff volume error or the error in peak flow are completely insensitive to when the runoff occurs, with the result that a simulated hydrograph may have a perfect score despite not resembling the observed hydrograph. Other measures (e.g. Nash-Sutcliffe model efficiency or mean absolute error) compare observed and simulated values for the same time points, and therefore they will penalize a simulated hydrograph for which the shape exactly matches the observed hydrograph, but with some shift in time. This effect is especially pronounced if the falling and rising limbs of the hydrograph are steep or if the time scale of the rainfall-runoff process is short, both of which are often the case in urban catchments with many impervious surfaces. Objective functions that implicitly penalize timing differences may therefore not provide the most suitable measure to compare the performance of different candidate sets of model parameters.
parameters, and considering the timing error in the objective function may be beneficial. The relevance of errors along the time-axis depends on the purpose of the model, e.g., it can be more relevant to accurately estimate how high flow rates and runoff volumes will be than knowing the exact time that these occur. Hydrograph-specific approaches to including timing errors in objective functions are e.g., the Hydrograph Matching algorithm (HMA; Ewen 2011) and the Series Distance concept (Ehret and Zehe 2011; Seibert, Ehret, and Zehe 2016). Both draw connecting lines (‘rays’) between comparable points on the simulated and observed hydrograph, similar to the way a hydrologist might visually compare two hydrographs. Longer rays indicate a larger mismatch between the two hydrographs.

The use of such objective functions that account for timing errors in model calibration has only received limited attention. Liu et al. (2011) used a wavelet-based method to evaluate the magnitude of timing errors in hydrographs and were able to use this information to apply corrections which improved the model predictions. Similarly, Towler and McCreight (2020) showed how wavelets can be used to quantify timing errors and to aid in model diagnostics. However, these two studies only considered evaluating pre-existing simulated hydrographs and did not apply the wavelet techniques during model calibration. Hemri and Klein (2017) found that hydrograph-specific approaches such as HMA and Series Distance as well as the more general-purpose Dynamic Time Warping were useful for identifying training periods when forecasting water level for river-navigation purposes. Lerat and Anderssen (2015) demonstrated that their method for allowing for timing errors in calibration was able to reconstruct (in calibration) synthetic outflow data using a simple model when the input rainfall was shifted in time. Neither study considered aspects such as the uncertainty in the model predictions or parameter estimates, nor did they use urban catchments where timing errors may be particularly important as discussed in the previous paragraph. Therefore the goal of this paper is to investigate if objective functions that account for timing errors can improve model calibration performance in terms of how well the model matches the observations (predictive uncertainty) and how well the different model parameters can be identified (parametric uncertainty). This is investigated using actual observed data, so that the effect of e.g., model structure deficiencies is implicitly included.

2. Materials and methods

2.1. Study catchment, data, and model

This study used a 10.2 ha urban catchment in Luleå, Sweden with an imperviousness of approx. 37%. 1-minute observations of rainfall and catchment outflow were available (with only limited gaps) for the snow-free periods from June 2016 to November 2019; details on the measurement campaign are available in earlier works (Broekhuizen 2019; Broekhuizen et al. 2020). Uncertainties in the flow measurements were estimated by applying the results of laboratory benchmarks for the level and velocity (Aguilar, McDonald, and Dymond 2016) to the circular pipe using the law of propagation of uncertainties (e.g., Bertrand-Krajewski and Bardin 2002). The periods 2016–2017 and 2018–2019 were used as calibration and validation period respectively. To avoid large influence of dry periods on the results, only observations ≥ 1 l/s were used. In this way the start and end of runoff for each event are included in the evaluation, but not dry periods between events. The reduction of the number of observation points (4.3% of the original number) also improves the performance of the HMA (see §2.2) considerably.

A previously developed finely discretized SWMM model (Broekhuizen et al. 2020) was used, where each of the 146 subcatchments was completely impervious or completely pervious and physical subcatchment properties such as width and slope could be estimated directly (see e.g., Krebs et al. 2014; Petrucci and Bonhomme 2014). The calibration parameters are listed in Table 1. The number of calibration parameters was lowered compared to Broekhuizen et al. (2020), since the focus of the current study was on the comparison of different approaches to model evaluation, rather than detailed fine-tuning to maximize model performance. A multiplier for the rainfall input is applied to allow adjusting runoff volume. This is more typically done by calibrating e.g., the (percentage) impervious area, but since the model was setup using high-resolution GIS data it was preferred to estimate imperviousness and catchment area directly.

Since this study was based on actual field data, some timing errors (resulting from e.g., rainfall events reaching the gauge and catchment at different times or drift between the clocks of the data loggers used) are expected to be present in the data, but the precise extent is unknown. Given the goals of this study, two scenarios with additional time drift (to simulate data loggers having slow or fast clocks) were also included. Following the scenarios used by Dotto et al. (2014), the two time drift scenarios assume that the logger for the flow data drifts linearly at a rate of ±0.14 min/day and is resynchronized every four weeks.

2.2. Objective functions

Three different types of objective functions were used in this study to represent three different ways of approaching errors in hydrograph timing. The Nash-Sutcliffe model efficiency only compares simulated and observed values from the same point in time:

\[
NSE = 1 - \frac{\sum_i (S_i - O_i)^2}{\sum_i (O_i - \bar{O})^2}
\]

Table 1. Calibration parameters.

| Description                                      | Symbol | Min   | Max   | Unit  |
|--------------------------------------------------|--------|-------|-------|-------|
| Manning’s number, impervious surfaces            | n.IMP  | 0.010 | 0.025 | m/s   |
| Manning’s number, green areas                    | n.GR   | 0.030 | 0.5   | m/s   |
| Manning’s number, sewer pipes                    | n.pipe | 0.005 | 0.020 | m/s   |
| Depression storage, impervious surfaces          | s.IMP  | 0     | 2.5   | mm    |
| Depression storage, green areas                  | s.GR   | 0     | 20    | mm    |
| Hydraulic conductivity, green areas              | ksat.GR| 0.1   | 200   | mm/hr |
| Rainfall multiplier                              | rm     | 0.5   | 2.5   |       |

*The green areas include a number of drainage swales that may have relatively large storage, hence the high maximum value for this parameter.
Where $S_i$ and $O_i$ denote the simulated resp. observed value at
time $i$. A value of 1 indicates perfect agreement, and a value of 0
indicates performance as good as when using the mean of the
observations as the simulation.

The relative volume error was used as a traditional objective
function that is completely insensitive to timing:

$$VE = \frac{\sum_i (S_i - O_i)}{\sum_i O_i}$$  \hspace{1cm} (2)

Note that a value of 0 is ideal: positive values indicate over-
estimation of the total flow volume and vice versa.

For a timing-sensitive objective function the Hydrograph
Matching Algorithm (HMA; Ewen 2011) was selected, since it
is relatively easy to implement and it has been found to have
good performance compared to expert judgment (Crochemore
et al. 2015). The method first calculates an error value for each
‘ray’ connecting a simulated and observed point:

$$e = r^2 + b^2 \tau^2$$  \hspace{1cm} (3)

Where $r$ is the difference in magnitude (e.g. in l/s), $\tau$ is the
difference in time (e.g. in minutes), and $b$ is a weighting factor
(with unit l/s/minute) applied to the time error. Any units can
be used for time and flow rates, as long as an appropriate value
of $b$ is set. From the calculated errors the HMA then finds the
optimal path (i.e. the collection of rays with the lowest sum of
e), with the following constraints:

1. One ray attaches to each observed point.
2. A simulated point may be connected to zero, one or two
observed points, but no more than one consecutive
simulated point may be skipped.
3. No ‘crossing’ of rays in time, e.g. if there is a ray from $S_5$
to $O_n$, there cannot be a ray from $S_4$ to $O_b$.
4. $-w_1 \leq \tau \leq w_2$, where $w_1$ and $w_2$ are user-controlled
parameters to impose a maximum allowed lag ($w_1$) and
lead ($w_2$). Note that in this paper only cases where $w_1$
and $w_2$ are used.

These limitations mean that a simple algorithm (give in 13 lines
of pseudocode by Ewen (2011)) suffices to find this optimal
path. Next, an objective value with the same interpretation as
NSE can be obtained using:

$$H = 1 - \frac{\sum_i e_{opt,i}}{\sum_i (O_i - \bar{O})^2}$$  \hspace{1cm} (4)

Where $e_{opt,i}$ is the value of $e$ associated with the $i$-th observa-
tion-simulation ray on the optimal path. This form facilitates
comparison with regular NSE, as the interpretation (a value of 1
is perfect, the mean of all observations is 0) is the same. Note
that the errors of the benchmark model (i.e. the denominator in
Equation (4)) would also have to be calculated accounting for
timing errors using Equation (3). However, since the benchmark
model (mean of all observations) is simply a horizontal line, the
error rays would be vertical lines without any timing compo-
nent, so the expression in Equation (4) suffices.

To show the effect of using different values for the timing-
error weight $b$, two different values were used in this paper: 0.5
and 5. These two cases are denoted $H_{0.5}$ and $H_5$ respectively.

The value of 0.5 represents a case where timing errors are
penalized lightly, but still considerably more than when using
the volume error; the value of 5 represents a case where timing
errors are penalized more strictly, but still considerably less
than when using NSE. In real applications the choice of $b$
should depend on the purpose of the study; visualizing the
resulting observational uncertainty as described below in §2.3
and shown in e.g. Figure 2 can be helpful. The value of $w_1$ and
$w_2$ was fixed at 60 minutes. Theoretically, this allows for con-
necting rays between simulated and observed points that are
one hour apart; however, such large timing errors would lead to
large error values (Equation (3)) and in practice they do not
actually occur.

### 2.3. Calibration approach

The Generalized Likelihood Uncertainty Estimation method
(GLUE; Beven and Binley 1992) was used to generate uncertain
estimates for model parameters and model outputs (in this case
catchment outflow). First, 10 000 parameter samples were
randomly generated using a Monte Carlo method and the
model run for each of these. Runs where the objective function
value falls below a user-determined threshold are discarded.
From the ensemble of remaining parameter sets and the corre-
sponding simulated outflows it is then possible to construct
uncertain estimates (probability distributions) of parameter
values and simulated outflow. Since four objective functions
were used it was not possible to directly determine what
threshold value to use while maintaining comparability
between the different objective functions. Instead, the method
proposed by Vezzaro et al. (2013) was used. For HMA-based
objective functions this method was extended to explicitly
consider timing errors as follows:

1. Generate N parameter samples and run the model and
calculate the objective function value(s) $F$ for each of
these.
2. For each observation $i$, let $q_i$ be the nominal value of the
observation, $u_{qi}$ be the (estimated) uncertainty of $q_i$ (e.g.
as described in §2.1), and $t_i$ be the time of the observa-
tion. According to Vezzaro et al., an observation is con-
sidered to be ‘covered’ if the predicted outflow
(uncertainty band) overlaps with the interval $[q_i - u_{qi}, q_i +
u_{qi}]$. Extending this to include the time dimension
according to Equation (3), the simulation uncertainty
and observational uncertainty overlap if (see Figure 1):

$$u_{qi}^2 \leq (q - q_i)^2 + b^2 (t - t_i)^2$$  \hspace{1cm} (5)

which describes an ellipse centred on the nominal
observation where the HMA parameter $b$ controls the
eccentricity of the ellipse, i.e. the weight of time errors
relative to magnitude errors. The ellipse shape follows
from allowing a trade-off between magnitude and tim-
ing errors according to Equation (3) (i.e. using squared
errors); different formulations of Equation (3) would lead
to different shapes, e.g. if absolute errors were used the
shape would be a diamond instead. The maximum
Figure 1. Principle sketch of the method for determining observational cover. In this example, the observation is considered covered by the model prediction (since there is overlap between the simulation and observation uncertainty), while it would not have been considered as covered if only the vertical dimension was considered.

Timing error (i.e., if there is no magnitude error) is then given by:

\[ u_t = |t - t_o| = \sqrt{\frac{u_{Q_t}^2}{b^2}} \quad (6) \]

(3) Take the 2 parameter sets with the best value for \( F \), calculate the upper limit \( S_U \) and lower limit \( S_L \) of the simulation uncertainty bound (e.g., the central 95% interval as used in this paper). Then count the number of observations for which there is an overlap between this bound and the ellipse from Equation (5), see Figure 2. This can be operationalized as follows:

a. For each observation time step \( t_o \) uniformly sample a number of points \( t_p \) along the time axis between \( t_o - u_t \) and \( t_o + u_t \).

b. Observation \( i \) is considered ‘covered’ by the simulation uncertainty bound if there is:

i. At least one point \( t_p \) for which

\[ S_U(t_p) \geq q_i - \sqrt{u_{Q_t}^2 - b^2(t_p - t_o)^2} \quad (7) \]

ii. And at least one point \( t_p \) for which

\[ S_L(t_p) \leq q_i + \sqrt{u_{Q_t}^2 - b^2(t_p - t_o)^2} \quad (8) \]

Equation 7a describes the lower half of the ellipse in Figure 1, equation 7b the upper half. Note that, if simulations and observations are available at e.g., 1-minute time intervals, the time step between points \( t_o \) should be smaller than 1 minute (e.g., 15 s) for the count to be accurate and \( S_U \) and \( S_L \) are interpolated to this time step.

(4) Repeat step 3 with the best 3, 4, …, \( N \) parameter sets, counting for each number the number of observations that is covered by the simulation uncertainty bound. (A larger step size may be used to lower computational demands.)

(5) A plot of the number of observations covered vs. number of parameter sets included can be used to identify a point where adding further parameter sets does not further improve model performance. Preliminary results for this paper showed that observation coverage (calculated over all observations > 1 l/s) kept slowly increasing throughout the entire parameter range, so no clear cut-off point was visible, and the results could be very sensitive to the somewhat arbitrary choice of the cutoff point. Instead, as an alternative that was objectively applicable to all twelve calibrations (four objective functions in three time drift scenarios) the cut-off point was set at the first point where at least 90% of flow peak observations were covered. A flow peak was defined as any observed value where no higher values existed within the preceding or following 15 minutes.

For non-HMA objective functions the same method is applied without the time-dependent components in Equations (5)–(7) and without the need for interpolating to multiple points in step 5, so that observation \( i \) is considered ‘covered’ by the simulation uncertainty if:

Figure 2. Example of using estimated observation uncertainty (blue ellipses) to count how many observations the model prediction captures.
Figure 3. Effect of using HMA on the ranking of candidate parameter sets, with just the (unknown) timing errors present in the data as observed, and with additional time drift added. The cutoff points (see Table 2) are indicated.
Table 2. Summary of predictive uncertainty for different likelihoods and time drift scenarios. Coverage of peaks in the calibration phase is not shown, since it is equal to 0.90 for all rows.

| Likelihood | Time drift | # accepted par sets | ARIL | Cover, all obs | ARIL | Cover, all obs | Cover, peaks |
|------------|------------|---------------------|------|----------------|------|----------------|--------------|
| NSE        | as observed| 3445                | 0.91 | 0.89           | 0.98 | 0.82          | 0.92         |
| VE         | as observed| 2000                | 0.67 | 0.85           | 0.80 | 0.80          | 0.92         |
| H_{0.5}    | as observed| 986                 | 0.52 | 0.86           | 0.63 | 0.80          | 0.90         |
| H_{5}      | as observed| 2420                | 0.74 | 0.88           | 0.83 | 0.82          | 0.92         |
| NSE        | +0.14min4W | 3805                | 0.97 | 0.88           | 1.05 | 0.82          | 0.88         |
| VE         | +0.14min4W | 2840                | 0.81 | 0.85           | 0.94 | 0.80          | 0.90         |
| H_{0.5}    | +0.14min4W | 1015                | 0.53 | 0.85           | 0.64 | 0.80          | 0.86         |
| H_{5}      | +0.14min4W | 3263                | 0.88 | 0.88           | 0.97 | 0.82          | 0.88         |
| NSE        | −0.14min4W | 3395                | 0.89 | 0.89           | 0.96 | 0.83          | 0.93         |
| VE         | −0.14min4W | 2230                | 0.71 | 0.87           | 0.84 | 0.81          | 0.93         |
| H_{0.5}    | −0.14min4W | 652                 | 0.48 | 0.87           | 0.58 | 0.80          | 0.91         |
| H_{5}      | −0.14min4W | 2350                | 0.70 | 0.88           | 0.79 | 0.82          | 0.93         |

\[ S_{ij} \geq q_i - u_{qi} \land S_{ij} \geq q_i + u_{qi} \]  \hspace{1cm} (9)

In this case the method reduces to the original method proposed by Vezzaro et al. (2013).

To assess the precision of the model predictions, the Average Relative Interval Length was used:

\[ \text{ARIL} = \frac{\sum_{i=1}^{n} (S_{ij} - S_{ij})}{\sum_{i=1}^{n} O_i} \]  \hspace{1cm} (10)

Lower values indicate a more precise prediction, i.e. a narrower uncertainty bound as shown e.g. in Figure 2. Other authors (Dotto et al. 2012; Jin et al. 2010) used a different form of the ARIL where the interval length was divided by the observed flow at each time step before taking the average. However, this can result in very high values (or even undefined ones) when near-zero (or equal to zero) flow observations are present. This is avoided by Equation (9) (a similar approach was used by Evin et al. 2014). Furthermore, considering only the width of the confidence interval for each observation time step does not fit with the timing-sensitive approach being tested: a visually narrow band in a rising or falling hydrograph section may be quite wide along the vertical dimension.

3. Results

3.1. Ranking of parameter sets

Step 3 in §2.3 uses the rank of the candidate parameter sets according to the objective to determine the order in which they get added to the ensemble, so the choice of objective function may affect the outcomes. Figure 3 shows the effect of using HMA on this ranking. In this case, it is mainly the best half of the parameter sets whose rank is affected, and the changes in rank become larger if an additional time drift is added to the observations (right panel). The effect will also depend on the value chosen for the HMA parameter b, the weight assigned to timing errors. For the worst half of the parameter sets, using HMA had little effect on their ranking, showing that the effect of HMA is fine-tuning between relatively good parameter sets rather than ‘fixing’ relatively bad parameter sets.

3.2. Precision and reliability of model predictions

Table 2 summarizes the precision (ARIL) and reliability (observation cover) for the four different likelihoods under different scenarios for time drift (see §2.1). The reliability was similar in all cases, which was not surprising since the same level of reliability (for peak flows) was used as the target for calibration. However, as the number of included parameter sets varied by a factor of up to 6 between the different likelihoods, the precision of the model predictions (i.e. the width of the uncertainty band) varied. During calibration, ARIL for H_{0.5} (i.e. with a small penalty for timing errors) was approx. 0.4 lower than for regular NSE. When a heavier weight was attached to timing errors (H_{5})
the difference was smaller (0.1–0.2) but still present. In the validation phase the coverage was about the same in all cases, both when looking at all observations and when considering just flow peaks. Note that in the validation phase the observation cover was counted without allowing for timing errors, in order to create an even comparison between all likelihoods (otherwise the HMA-based likelihoods could theoretically get an advantage from covering additional observations by considering the timing, even if their predictions were the same as e.g. regular NSE). Observation cover was also counted while allowing for timing error as in the calibration phase, but since this had negligible influence on the results (maximum one percent point change) this conceptually less satisfying metric is not shown. As in the calibration phase, the HMA-based likelihoods again resulted in more precise predictions. An example of this can be seen in Figure 4. Note that the reliability (hit/miss of observations) is the same for the different likelihoods, but the precision (width of the simulated uncertainty band) varies.

### 3.3. Parameter identification

The identifiability of the different model parameters is summarized in Table 3 by means of the Kolmogorov-Smirnov distance metric between the accepted and the rejected parameter samples. If the distance is significantly different from zero (according to the Holm-Bonferroni method (Holm 1979), with a maximum 5% chance of at least one false positive in the entire table) are shown in bold, italic font.

| likelihood | Time drift | npars | n_IMP | s_IMP | n_GR | s_GR | ksat_GR | n_pipe | rm |
|------------|------------|-------|-------|-------|------|------|---------|--------|----|
| nse        | as observed| 3445  | 0.03  | 0.03  | 0.02 | 0.02 | 0.06    | 0.09   | 0.88|
| ve         | as observed| 2000  | 0.04  | 0.02  | 0.02 | 0.02 | 0.04    | 0.03   | 0.74|
| H0.5       | as observed| 986   | 0.06  | 0.13  | 0.05 | 0.05 | 0.06    | 0.02   | 0.77|
| H5         | as observed| 2420  | 0.05  | 0.02  | 0.02 | 0.03 | 0.04    | 0.13   | 0.78|
| nse        | +0.14min4W | 3805  | 0.03  | 0.04  | 0.03 | 0.02 | 0.06    | 0.09   | 0.90|
| ve         | +0.14min4W | 2840  | 0.03  | 0.02  | 0.02 | 0.03 | 0.06    | 0.02   | 0.77|
| H0.5       | +0.14min4W | 1015  | 0.10  | 0.12  | 0.04 | 0.05 | 0.04    | 0.26   | 0.71|
| H5         | +0.14min4W | 3265  | 0.03  | 0.02  | 0.02 | 0.03 | 0.05    | 0.10   | 0.82|
| nse        | −0.14min4W | 3395  | 0.02  | 0.02  | 0.02 | 0.03 | 0.07    | 0.07   | 0.87|
| ve         | −0.14min4W | 2230  | 0.04  | 0.02  | 0.02 | 0.03 | 0.04    | 0.02   | 0.75|
| H0.5       | −0.14min4W | 652   | 0.11  | 0.22  | 0.04 | 0.06 | 0.03    | 0.27   | 0.71|
| H5         | −0.14min4W | 2350  | 0.02  | 0.02  | 0.03 | 0.03 | 0.06    | 0.08   | 0.79|

Figure 5. Kernel density estimates (probability distributions) for estimated values for selected parameters. The boundaries of the x-axis are set to the calibration ranges of parameters from Table 1.
in any scenario, which may be attributed to the fact that green areas generate runoff in only a subset of events (Broekhuizen et al. 2020). The parameter probability densities (i.e. how likely different values for each parameter are) in Figure 5 shows that the parameter estimates for different likelihoods and time drift scenarios overlap with each other, but H0.5 shows a clearer peak for all parameters.

4. Discussion

There are several benefits to the application of HMA-based objective functions. First, as shown in Figure 3, the HMA affects the relative order of different candidate parameter sets, and thereby the selection of which parameter sets are accepted and rejected in the GLUE procedure. This appears to only have a noticeable effect on the parameter sets that have relatively good performance, i.e. the sets that provide a reasonable approximation of the actual system response. Worse performing parameter sets are not affected much, so the HMA is a polishing or fine-tuning step, rather than a fix for poor performance. For use as an objective function, this is a positive, since the goal of any objective function is to be able to distinguish between parameter sets with different performance. Reducing the differences between the performance of different parameter sets too much would be unhelpful for model calibration.

The model predictions that were made using HMA proved to be equally reliable, but more precise than predictions made using traditional objective functions. The extent of this depends on the weight assigned to timing errors, but even the case with a heavier penalty for timing errors (H0) still improves the precision of the model predictions. Improved precision, even without improved reliability, is a positive property in a modelling context, since more precise answers are useful in both scientific and practical model applications. Additional time drift may increase the difference between the regular and the HMA objective functions to some degree (see Table 2): adding a time drift of +0.14 min/week increased the ARIL for NSE by 0.06, but for H0.5 by only 0.01. The case with heavier timing error penalties (H0) saw an increase in ARIL of 0.14, but the ARIL was still lower than for regular NSE. Applying time drift in the other direction reduced the ARIL for both regular NSE and for H0.5 and H0, so this time drift scenario may have reduced the amount of time drift present in the observed data.

The reduced predictive uncertainty is the result of fewer parameter sets being needed to obtain a reliable model prediction. To the extent that model parameters reflect physical catchment properties, this improves the information about the catchment that is inferred in the calibration process. Additionally, the lower number of members of the final model ensemble means that fewer runs are needed later when e.g. making forecasts, thus reducing computational demands.

Since HMA dynamically searches for an optimal pairing of observed points to simulated points, it only allows for timing errors when it is beneficial (for the objective function value) to do so. In other words, the ‘worst-case’ scenario (where there is no advantage to allowing for timing errors) simply means that the HMA measures reduce to the ordinary NSE. There is thus no risk of worsening the objective function value, the only downside is that additional computational effort is required for no gain.

There are also several downsides to the application of the HMA. The first is that the computational effort can be considerable, depending on the number of observation points that is being used. In some cases, the available computer memory may also be a limiting factor. For one example dataset with 10,000 observation points, HMA (following the pseudo-code from Ewen (2011)) took 5.7 s, which is a considerable additional burden when thousands of model runs are required. However, faster implementations of HMA were developed as part of this study (github.com/icobro/PyHMA that help remedy this issue: the fastest implementation takes only 0.7 s, i.e. 88% less than the naive implementation, while also using less memory.

The HMA introduces two parameters (timing error weight b and maximum timing error w) in the objective function, for which the user has to choose values. This is a subjective choice that should be made based on the objectives of the study. For example, in real-time control applications timing errors will be more critical, so the value of b should be relatively high, while if the goal is to dimension pipes and storage volumes the timing is less important and the value for b should be lower. Although such subjectivity is not ideal, it can also be considered that using traditional objective functions also implicitly makes this choice: standard NSE is the same as using HMA with b = ∞ and using b = 0 results in an approximation of the volume error. Making this choice an explicit one is not necessarily a negative factor. The choice of w may be less critical, depending on the value used for b: unless b is very small, there will always be a noticeable penalty for pairing simulation and observations that are far apart in time, so such pairings will be selected only in extreme cases, if ever.

There are some limitations to the current study that should be pointed out. First, the single case study used here is to some extent dominated by one model parameter (the rainfall multiplier that controls overall volumes), which may have hidden to some extent how identifiable other parameters were. The use of observed data helps demonstrate that the proposed method works in practice, but it also means that other uncertainties (e.g. in rainfall and flow measurements, deficiencies in the model structure (e.g. Deletic et al. 2012)) also influenced the modelling process and may make the differences between the different objective functions less visible. Second, HMA is only one of a number of approaches to allowing for timing errors. Alternatives that are available (Crochemore et al. 2015) include the Series Distance approach (Seibert, Ehret, and Zehe 2016), the approach by Lerat and Andersen (2015) and Dynamic Time Warping, and these may give different results. One advantage of HMA is that it is relatively easy to implement and may also be easier to explain to stakeholders in the modelling process that are not familiar with it.

5. Conclusions

The objective of this study was to investigate the effects that an objective function that allows for a limited degree of timing errors between simulated and observed hydrographs (HMA) had on model calibration. Model predictions made using HMA were as reliable as predictions made using
ordinary Nash-Sutcliffe model efficiency (NSE), but more precise, i.e. the uncertainty in the modelled outflow was lower. The higher precision is underpinned by more precise estimates for some model parameters. Since HMA will reduce to the normal NSE if there is no benefit to allowing timing errors, there is no risk of it giving worse results. HMA could also be seen as a way to find a middle ground between NSE and VE: the user-specified weight and limits of these timing errors allow the modeller to adjust the method to the objectives of their study. The main downside of HMA is the additional computational burden, which may be reduced with the optimized implementations developed in this study.

Acknowledgements

This study was funded by The Swedish Research Council Formas (grant nr. 2015-121 and 2015-778) and by VINNOVA as part of DRIZZLE – Centre for Stormwater Management (grant nr. 2016-05176). We gratefully acknowledge technical expertise provided by the Stormwater&Sewers network and would particularly like to thank Helen Galf, Ralf Rentz and Karolina Berggren for their work in setting up and maintaining the field measurements. The authors would like to thank CHL/HydroPraxis for providing a license for PCSWMM.

Data availability statement

The data and model used in this study are available online (https://doi.org/10.5281/zenodo.3925024). The implementations of the Hydrograph Matching Algorithm that were written for this paper are available at http://doi.org/10.5281/zenodo.3923792.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Svenska Forskningsrådet Formas (2015-2015-778); VINNOVA (2015-05176).

ORCID

Ico Broekhuizen
http://orcid.org/0000-0002-6907-8127

Günther Leonhardt
http://orcid.org/0000-0003-0367-3449

Maria Viklander
http://orcid.org/0000-0003-1725-6478

References

AgUILAR, M. F., W. M. McDonALD, and R. L. Dymond. 2016. “Benchmarking Laboratory Observation Uncertainty for In-Pipe Storm Sewer Discharge Measurements.” Journal of Hydrology 534 (March): 73–86. doi:10.1016/j.jhydrol.2015.12.052.

Barco, J., K. M. Wong, and M. K. Stenstrom. 2008. “Automatic Calibration of the U.S. EPA SWMM Model for a Large Urban Catchment.” Journal of Hydraulic Engineering 134 (4): 466–474. doi:10.1061/(ASCE)0733-9429-(2008)134:4(466).

Bennett, N. D., B. F. W. Croke, G. Guariro, J. H. A. Guillaume, S. H. Hamilton, A. J. Jakeman, S. Marsili-Libelli, et al. 2013. “Characterising Performance of Environmental Models.” Environmental Modelling & Software 40 (February): 1–20. doi:10.1016/j.envsoft.2012.09.011.

Bertrand-Krajewski, J.-L., and J.-P. Bardin. 2002. “Evaluation of Uncertainties in Urban Hydrology: Application to Volumes and Pollutant Loads in a Storage and Settling Tank.” Water Science and Technology 45 (4–5): 437–444. doi:10.2166/wst.2002.0645.

Beven, K., and A. Binley. 1992. “The Future of Distributed Models: Model Calibration and Uncertainty Prediction.” Hydrological Processes 6 (3): 279–298. doi:10.1002/hyp.3360060305.

Broekhuizen, I. 2019. “Uncertainties in Rainfall-Runoff Modelling of Green Urban Drainage Systems: Measurements, Data Selection and Model Structure.” Licentiate thesis. Luleå University of Technology, Luleå. http://urn.kb.se/resolver/urn:nbn:se:ltu:diva-73367

Broekhuizen, I., G. Leonhardt, J. Marsalek, and M. Viklander. 2020. “Event Selection and Two-Stage Approach for Calibrating Models of Green Urban Drainage Systems.” Hydrology and Earth System Sciences 24 (2): 869–885. doi:10.5194/hess-24-869-2020.

Crochemore, L., C. Perrin, V. Andréassian, U. Ehret, S. P. Seibert, S. Grimaldi, H. Gupta, and J.-E. Faturel. 2015. “Comparing Expert Judgement and Numerical Criteria for Hydrograph Evaluation.” Hydrological Sciences Journal 60 (3): 402–423. doi:10.1080/02666667.2014.903331.

Deletic, A., C. B. S. Dotto, D. T. McCarthy, M. Kleidorfer, G. Freni, G. Mannina, M. Uhl, et al. 2012. “Assessing Uncertainties in Urban Drainage Models.” Physics and Chemistry of the Earth, Parts A/B/C 44–42 (January): 3–10. doi:10.1016/j.pce.2011.04.007.

Dotto, C., M. Kleidorfer, A. Deletic, W. Rauch, and D. T. McCarthy. 2014. “Impacts of Measured Data Uncertainty on Urban Stormwater Models.” Journal of Hydrology 508 (January): 28–42. doi:10.1016/j.jhydrol.2013.10.025.

Dotto, C., G. Mannina, M. Kleidorfer, L. Vezzaro, M. Henrichs, D. T. McCarthy, G. Freni, W. Rauch, and A. Deletic. 2012. “Comparison of Different Uncertainty Techniques in Urban Stormwater Quality and Modelling.” Water Research 46 (8): 2545–2558. doi:10.1016/j.watres.2012.02.009.

Easton, Z. M., P. Gérard-Marchant, M. T. Walter, A. M. Petrovic, and T. S. Steenhuys. 2007. “Hydrologic Assessment of an Urban Variable Source Watershed in the Northeast United States.” Water Resources Research 43 (3): 3. doi:10.1029/2006WR005076.

Ehret, U., and E. Zehe. 2011. “Series Distance – An Intuitive Metric to Quantify Hydrograph Similarity in Terms of Occurrence, Amplitude and Timing of Hydrological Events.” Hydrology and Earth System Sciences 15 (3): 877–896. doi:10.5194/hess-15-877-2011.

Evín, G., M. Thyer, D. Kavetski, D. McNerney, and G. Kuczera. 2014. “Comparison of Joint versus Postprocessor Approaches for Hydrological Uncertainty Estimation Accounting for Error autocorrelation and Heteroscedasticity.” Water Resources Research 50 (3): 2350–2375. doi:10.1002/2013WR014185.

Evensen, J. 2011. “Hydrograph Matching Method for Measuring Model Performance.” Journal of Hydrology 408 (1–2): 178–187. doi:10.1016/j.jhydrol.2011.07.038.

Hemri, S., and B. Klein. 2017. “Analog-Based Postprocessing of Navigation-Related Hydrological Ensemble Forecasts.” Water Resources Research 53 (11): 9059–9077. doi:10.1002/2017WR020684.

Holm, S. 1979. “A Simple Sequentially Rejective Multiple Test Procedure.” Scandinavian Journal of Statistics 6 (2): 7.

Jin, X., C.-Y. Xu, Q. Zhang, and V. P. Singh. 2010. “Parameter and Modeling Uncertainty Simulated by GLUE and a Formal Bayesian Method for a Conceptual Hydrological Model.” Journal of Hydrology 383 (3): 147–153. doi:10.1016/j.jhydrol.2009.12.028.

Kleidorfer, M., M. Möderl, S. Fach, and W. Rauch. 2009. “Optimization of Measurement Campaigns for Calibration of a Conceptual Sewer Model.” Water Science and Technology 59 (8): 1523–1530. doi:10.2166/wst.2009.154.

Krebs, G., T. Kokkonen, M. Valtanen, H. Setälä, and H. Koivusalo. 2014. “Spatial Resolution Considerations for Urban Hydrological Modelling.” Journal of Hydrology 512 (May): 482–497. doi:10.1016/j.jhydrol.2014.03.013.

Lerat, J., and R. S. Anderssen. 2015. “Calibration of Hydrological Models Allowing for Timing Offsets.” In MODSIM2015, 21st International Congress on Modelling and Simulation, edited by T. Weber, M. J. McPhee, and R. S. Anderssen. Modelling and Simulation Society of Australia and New Zealand. doi:10.36334/MODSIM2015.A2.Lerat.

Liu, Y., J. Brown, J. Demarge, and D.-J. Seo. 2011. “A Wavelet-Based Approach to Assessing Timing Errors in Hydrologic Predictions.” Journal of Hydrology 397 (3): 210–224. doi:10.1016/j.jhydrol.2010.11.040.

Petrucci, G., and C. Bonhomme. 2014. “The Dilemma of Spatial Representation for Urban Hydrology Semi-Distributed Modelling: Trade-Offs among Complexity, Calibration and Geographical Data.”
Seibert, S. P., U. Ehret, and E. Zehe. 2016. “Disentangling Timing and Amplitude Errors in Streamflow Simulations.” Hydrology and Earth System Sciences 20 (9): 3745–3763. doi:10.5194/hess-20-3745-2016.

Thorndahl, S., K. J. Beven, J. B. Jensen, and K. Schaarup-Jensen. 2008. “Event Based Uncertainty Assessment in Urban Drainage Modelling, Applying the GLUE Methodology.” Journal of Hydrology 357 (3–4): 421–437. doi:10.1016/j.jhydrol.2008.05.027.

Towler, E., and J. L. McCreight. 2020. “A Wavelet-Based Approach to Streamflow Event Identification and Modeled Timing Error Evaluation.” Hydrology and Earth System Sciences Discussions, September. Copernicus GmbH, 1–37. doi:10.5194/hess-2020-323.

Tscheikner-Gratl, F., P. Zeisl, C. Kinzel, J. Leimgruber, T. Ertl, W. Rauch, and M. Kleidorfer. 2016. “Lost in Calibration: Why People Still Do Not Calibrate Their Models, and Why They Still Should – A Case Study from Urban Drainage Modelling.” Water Science and Technology 74 (10): 2337–2348. doi:10.2166/wst.2016.395.

Vezzaro, L., P. S. Mikkelsen, A. Deletic, and M. David. 2013. “Urban Drainage Models – Simplifying Uncertainty Analysis for Practitioners.” Water Science and Technology 68 (10): 2136–2143. doi:10.2166/wst.2013.460.

Zhang, W., and T. Li. 2015. “The Influence of Objective Function and Acceptability Threshold on Uncertainty Assessment of an Urban Drainage Hydraulic Model with Generalized Likelihood Uncertainty Estimation Methodology.” Water Resources Management 29 (6): 2059–2072. doi:10.1007/s11269-015-0928-8.