hydrology

MDPI

Article

Quality Assessment of Small Urban Catchments Stormwater Models: A New Approach Using Old Metrics

Luís Mesquita David [1,2,*] and Tiago Martins Mota [1]

[1] LNEC—Laboratório Nacional de Engenharia Civil, 1700-066 Lisboa, Portugal; tfmartinsmota@gmail.com
[2] ISEL—Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, 1959-007 Lisboa, Portugal
* Correspondence: ldavid@lnec.pt

Abstract: Small urban catchments pose challenges in applying performance metrics when comparing measured and simulated hydrographs. Indeed, results are hampered by the short peak flows, due to rainfall variability and measurement synchronization errors, and it can be both difficult and inconvenient to remove base flows from the analysis, given their influence on combined sewer overflow (CSO) performance. A new approach, based on the application of metrics to peak flows for a selected set of different durations, is proposed and tested to support model quality assessment and calibration. Its advantages are: avoiding inconveniences arising from lags in peak flows and subjectivity of possible adjustments; favouring the assessment of the influence of base flow variability and flow lamination by CSOs; promoting integrated analysis for a wide range of rainfall events; facilitating bias identification and also guiding calibration. However, this new approach tends to provide results (e.g., for NSE, r^2 and PBIAS) closer to optimal values than when applying metrics to compare the measured and simulated values of hydrographs, so the comparison of results with thresholds widely used in the literature should be done with caution. The various case study examples highlight the importance of using a judicious set of different metrics and graphical analyses.

Keywords: urban drainage; stormwater; combined sewer overflows (CSO); uncertainty; model calibration and verification; performance ratings; Nash–Sutcliffe Efficiency (NSE); Kling-Gupta Efficiency (KGE)

Citation: David, L.M.; Mota, T.M. Quality Assessment of Small Urban Catchments Stormwater Models: A New Approach Using Old Metrics. *Hydrology* **2022**, *9*, 87. https://doi.org/10.3390/hydrology9050087

Academic Editors: Shirley Gato-Trinidad and Pingping Luo

Received: 3 March 2022
Accepted: 9 May 2022
Published: 12 May 2022

1. Introduction

Urban drainage models usually aim at supporting decisions on flooding mitigation, pollution control, sewer system management and, increasingly, city planning and regeneration [1–3]. Depending on the objectives of the work and the availability of hydrological and asset data, the models can be either distributed and physically based or more aggregated and conceptual [4–7]. The increasing implementation of decentralized measures and nature-based solutions, i.e., solutions that seek to replicate natural phenomena across the catchment, often requires distributed models and, in some cases, more detailed modelling of infiltration, evapotranspiration and water quality [8,9].

Calibration requirements also depend on the objectives of the study, the availability of data and the time and resources needed to collect data [10]. Calibration tends to be more demanding and conservative for the variables and scales most relevant for the study purpose. For example, in a model developed for flood risk assessment, emphasis is put on calibrating peak flows of the most severe storms [11,12], while in a model designed to manage combined sewer overflows (CSO), greater accuracy is required in calibrating the volumes of hydrographs for a much wider range of events, which includes medium and small magnitude rainfall [13–15]. In the latter case in particular, daily and seasonal variations in dry weather flow [16] and rainfall derived infiltration and inflow (RDII) may play a very important role [17–21].

The number of rain events recommended for calibration and verification and the representativeness of the monitored events also depend on the objectives for which the

model was designed [22]. However, it is common that models developed for a particular purpose are later used for other analyses with a broader scope than the one for which they were initially validated.

Uncertainty in hydrological modelling derives from several sources: observed data used to force and calibrate the model (measurement); model structure, which includes errors in the mathematical formulation that represents the hydrological phenomena; parameterization; and initial and boundary conditions. Parameterization uncertainty results from both structural and measurement uncertainties, the need to conceptually simplify processes, and the inability to estimate and measure the temporal and spatial variability of the effective parameters, among others [23–27].

Several methods have been used to analyse the uncertainty of hydrological models, although only a few have the ability to address, in an explicit and cohesive way, the three critical aspects of uncertainty analysis: understanding, quantification and uncertainty reduction [24]. One of the widely used methods is the generalized likelihood uncertainty estimation (GLUE), although it depends on subjective choices and lacks a formal statistical basis [25]. Of the various developments in formal Bayesian approaches, the most current that has been widely used is the differential evolution adaptive metropolis (DREAM), which merges the differential evolution algorithm with the adaptive Markov Chain Monte Carlo (MCMC) approach [25]. Machine learning and polynomial chaos expansion are other approaches based on data-driven uncertainty quantification that have recently experienced increasing development [16,28–30]. These methods make it possible to quantify uncertainty, determine error confidence intervals and, in the case of formal Bayesian approaches, reduce the parameter uncertainty by the inclusion of prior knowledge. Despite their advantages, these methods are complex, require expert-knowledge and a high computational effort, some have convergence issues and others depend on subjective decisions [24,25].

For most cases, the most practical and viable way to assess uncertainty is to quantify the goodness-of-fit or accuracy of model results in measured data, for which several statistical and graphical techniques have been used, developed and discussed [31–43]. An open source library with over 60 metrics recently implemented in Python and MATLAB® is presented in [44].

In practice, only about a dozen statistical metrics have been commonly used in hydrology, which can be classified into three main categories [36]: dimensionless (e.g., the widely used Nash–Sutcliffe Efficiency coefficient [31]); error index (e.g., the root mean square error); standard regression. It is consensual that both graphical techniques and more than one quantitative statistics should be used in model evaluation [33,36,39]. The Nash–Sutcliffe Efficiency coefficient has been by far the most used.

The most widely applied metrics do not take into account measurement errors, although modifications to some of them have been proposed to account for measurement uncertainty [35] and, later, for both measurement and simulation uncertainties [38].

Based on results reported in the literature, from watershed-scale models and for annual, monthly and daily temporal scales, Moriasi et al. (2007) [36] proposed thresholds for various dimensionless performance metrics according to qualitative ratings: "Very Good", "Good", "Satisfactory" and "Unsatisfactory". These thresholds were reviewed in Moriasi et al. (2015) [39], although the previous thresholds continue to be widely cited in the literature.

Despite the obvious advantages of this benchmarking, some authors call the attention to the drawbacks of the ad hoc use of aggregated efficiency metrics and to the greater need for modellers to understand the suitability of each metric and how it should be interpreted for the purposes of each model [45].

Obviously, in urban drainage, graphical and statistical techniques have always been applied to model calibration and evaluation, although the generalized use of aggregated efficiency metrics, such as the Nash–Sutcliffe coefficient and the Kling-Gupta coefficient [37], is more recent (e.g., [46–55]).

However, the temporal and spatial scales of urban catchments and the issues related to wet weather discharges (from both CSO and Sanitary Sewer Overflows, SSO) pose challenges that have not yet been sufficiently discussed, in particular when adopting thresholds for performance ratings proposed on the basis of other realities. The variability of the dry weather flow and the RDII can substantially contribute to the uncertainty of the model results concerning CSO and SSO discharges (CSO structures are usually designed to carry 3 to 6 times the average daily dry weather flow to the wastewater treatment plant). In these cases, the traditional approach of removing base flows to calibrate or assess the accuracy of the hydrological model becomes particularly difficult and debatable.

This article presents a brief description and discussion in the light of urban drainage of the most used metrics for model calibration and quality assessment, which are then applied to small urban catchment models with different levels of performance. An innovative approach more suited to the challenges of these catchments is proposed and the results are discussed in detail.

2. Main Metrics Used in Calibration and Assessment of Hydrological Models

2.1. Nash–Sutcliffe Efficiency Coefficient (NSE)

The Nash–Sutcliffe Efficiency coefficient (NSE, Equation (1)) [31] is probably the most used dimensionless coefficient to evaluate the performance of hydrological models [33,35]. NSE is a normalized statistic that compares the mean squared error (MSE) with the variance of the observed data ($\sigma_o{}^2$):

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2} = 1 - \frac{\mathrm{MSE}}{\sigma_o{}^2} \quad (1)$$

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{n} \quad (2)$$

$$\overline{O} = \frac{\sum_{i=1}^{n} O_i}{n} \quad (3)$$

$$\sigma_o{}^2 = \frac{\sum_{i=1}^{n}(O_i - \overline{O})^2}{n} \quad (4)$$

where o_i is the observed value at timestep i or at the ith preselected time "such as peaks or troughs in the hydrograph" [31]; P_i is the predicted value at timestep i or at the ith preselected time; $\overline{O}$ is the mean of the measured values; n is the total number of observations. In the literature, the population variance has been used instead of sample variance ($s^2 = \left(\sum_{i=1}^{n}(O_i - \overline{O})^2\right)/(n-1)$), since the sample comprises all timesteps of each hydrograph, i.e., it is equal to the population.

NSE varies between $-\infty$ and 1, with higher values indicating a better fit of the model to the observed data. NSE values lower than zero indicate that the model is a worse predictor than the mean of the observations.

The NSE value has been widely interpreted as a classic skill score of models, for various hydraulic and water quality variables [37,45]. The performance ratings proposed by Moriasi et al. (2007) [36] for a watershed scale have been widely used in the literature. According to [36], the model simulation can be judged as "satisfactory" for NSE > 0.50, "Good" for NSE > 0.60 and "Very good" for NSE > 0.75. In Moriasi et al. (2015) [39], the thresholds for the ratings of "Good" and "Very good" were raised to 0.70 and 0.80, for the annual, monthly and daily temporal scales, respectively.

In urban catchments, important rainfall events last a few hours or even minutes, and therefore the time step of measurement records is usually only of a few minutes. The urban drainage literature reports NSE values from the calibration of hydrologic or hydraulic models, normally between 0.5 and 0.9 [2], most of them being greater than 0.7 [7,10,20,22,46–48,50–53]. However, in some cases, NSE values above 0.95 are

reported [1,53,55], while in other specific cases, usually for water quality parameters, very low values, close to zero, are considered as acceptable [9].

Despite NSE being widely used, there has been lengthy discussion about its limitations and suitability, and, as such, several modifications and more refined criteria have been proposed.

2.2. Kling-Gupta Efficiency Coefficient (KGE)

The Kling-Gupta Efficiency (KGE, Equation (5) or Equation (6)) [37] is one of the criterion proposed to overcome some limitations of the NSE, which has been increasingly used in recent years, in urban drainage also (e.g., [9,48,51,54]). The KGE results from the decomposition of the NSE into three distinctive components representing the correlation, the bias, and a measure of the relative variability between predicted and observed values.

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{5}$$

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_p}{\sigma_o} - 1\right)^2 + \left(\frac{\mu_p}{\mu_o} - 1\right)^2} \tag{6}$$

where r is the linear correlation coefficient between predicted and observed values; α is the ratio between the standard deviations of predicted and observed values (σ_m/σ_o); and β is the ratio between the mean predicted and mean observed flows (μ_p/μ_o), representing the bias.

Like the NSE, KGE varies between $-\infty$ and 1, where KGE = 1 corresponds to an absolute agreement between the model and the observations. Although KGE thresholds are often considered the same as NSE ones in model classification, ref. [45] demonstrates that a model is a better predictor than the mean of the observed data if KGE $> 1 - \sqrt{2} = -0.41$, whereas with NSE, this only happens for NSE > 0.

2.3. Percent Bias (PBIAS)

The Percent bias, or Percent deviation (PBIAS, Equation (7)) is used to assess the average tendency of the predicted data to be smaller or greater than the observed data. Positive values of PBIAS indicate underestimation, and negative values indicate overestimation.

$$\mathrm{PBIAS} = \frac{\sum_{i=1}^{n}(O_i - P_i)}{\sum_{i=1}^{n} O_i} \cdot 100\% \tag{7}$$

where the variables have the same meaning as those in Equation (1).

Other names have been given to indicators calculated in a similar manner to PBIAS [36]: Percent streamflow volume error (PVE), prediction error (PE), and percent deviation of streamflow volume (Dv). Absolute values of PBIAS < 10% have been usually classified as very good [36,49]. However, in Moriasi et al. (2015) [39], the PBIAS thresholds for the ratings of "Satisfactory", "Good" and "Very good" were reduced to <±15%, <±10% and <±5%, for all daily, monthly and annual scales, respectively.

2.4. Root Mean Square Error (RMSE)

Both the root mean square error (RMSE) and the Mean Absolute Error (MAE) describe the difference between model simulations and observations in the units of the variable. The RMSE (Equation (8)) is less intuitive than the MAE, but tends to be more used in hydrological studies because it penalizes the largest errors more severely [32,33,35,36]. Reduced RMSE values are associated with smaller errors, with the null value corresponding to a model that perfectly fits the measured data.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(O_i - P_i)^2}{n}} \tag{8}$$

where the variables have the same meaning as those in Equation (1).

Another advantage of RMSE is that, by assuming that the errors are unbiased and the error mean is zero, the RMSE is the standard deviation of the error [56]. Therefore, if the errors are normally distributed, 68% of errors will lie within one standard deviation of the mean, and 95% of the errors will lie within 1.96 (≈2) standard deviations of the mean [8].

Hence, the estimated interval of errors at a 95% probability (which we will call I95) can be expressed as in Equation (9):

$$\text{I95} = \pm 1.96 \cdot \text{RMSE} \approx \pm 2 \cdot \text{RMSE} \tag{9}$$

For the normalization of the RMSE to a dimensionless coefficient, different methods are found in the literature. The most common ones consist of dividing the RMSE by the following parameter of the observed values: mean; standard deviation; difference between the maximum and the minimum; and interquartile range, i.e., the difference between 25th and 75th percentile. In this work we will use the first two.

The first one is known as the Coefficient of Variation of the RMSE (CVRMSE or CV(RMSE), Equation (10)).

$$\text{CVRMSE} = \frac{\text{RMSE}}{\overline{O}} \tag{10}$$

The second one, named by [36] as the RMSE-observations standard deviation ratio (RSR), is expressed as in Equation (11).

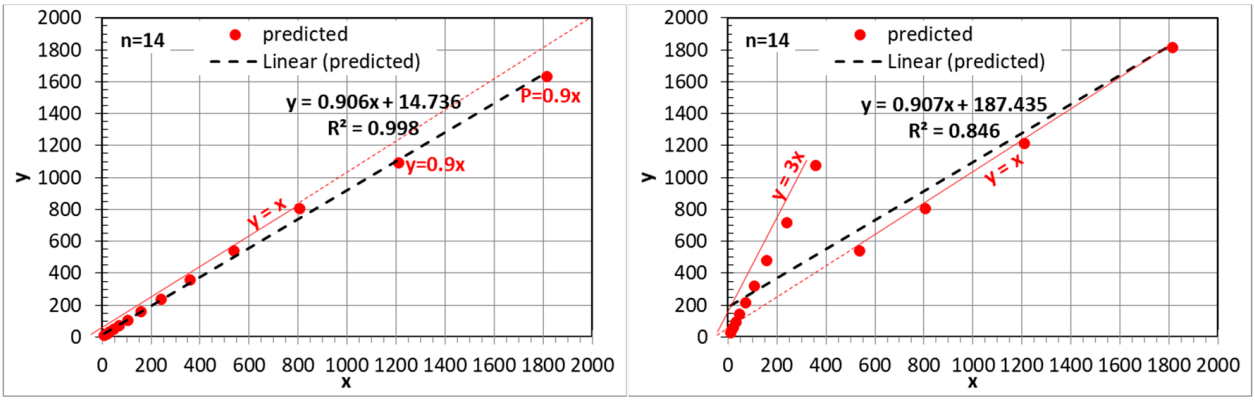
$$\text{RSR} = \frac{\text{RMSE}}{\sigma_o} \tag{11}$$

2.5. Linear Regression Coefficients and Graph

The slope, the y-intercept and the coefficient of determination (r^2) of the linear regression that best fits the simulation results to the observed values can provide important information about the model's quality. While the slope reflects the relative relationship between observed and predicted data, the y-intercept indicates a bias [36]. The coefficient of determination describes the amount of observed dispersion that is explained by the regression. It varies between 0 and 1, where 1 is for perfect alignment [34].

The coefficient of determination is also one of the most recommended criteria for calibrating and evaluating the accuracy of hydrological models. In Moriasi et al. (2015) [39], the r^2 thresholds for the ratings of "Satisfactory", "Good" and "Very good" are, respectively, 0.60, 0.75 and 0.85, for the daily flow at catchment scale.

However, there is also a consensus on the need to carefully interpret the linear regression parameters, due to the great weight in the results of the highest values, which can be significantly assisted by the analysis of linear regression plots. See the example in Figure 1, in which two series of 14 dots each have the same linear regression slope (slope = 0.906), but different behaviours: in (a), twelve values correspond to y = x and the two largest values to y = 0.9x; in (b), the four highest values correspond to y = x, but the other ten values correspond to y = 3x. In (a), the coefficient of determination is practically equal to unity, showing an excellent correlation between x and y for almost all values, but the slope of 0.9 indicates that a reduced number of intense events is underestimated by 10%. In (b), the coefficient of determination is still quite high from a statistical point of view (r^2 = 0.85) and the model is excellent for the most intense events; however, the slope of 0.9 and a high y-intercept value compared to the smallest events result from the model being too poor for most events.



(**a**) (**b**)

Figure 1. Example of two series with the same linear regression slope and good coefficients of determination, but with very different behaviours.

3. Challenges from Short-Duration Peak Flows in the Application of Metrics

The NSE, KGE and error statistics are commonly used to compare time series between modelled and observed values. However, in small urban and natural catchments, including ephemeral streams [57,58], many significant peak flows occur during very short periods of time. The spatial variability of rainfall and small desynchronizations between rain and flow measurement equipment can lead to delays or advances of the modelled series in relation to the observed series of only a few minutes, but with a significant impact on the statistical results.

Figure 2 shows the measured and modelled hydrographs of the most intense storm monitored in the case study presented below. Table 1 compares the results of the statistics described in the previous section for four scenarios: (a) the case represented in Figure 2; (b) the measured flow rate advanced by 2 min; (c) the measured flow rate delayed by 4 min; and (d) measured flow delayed by 6 min.

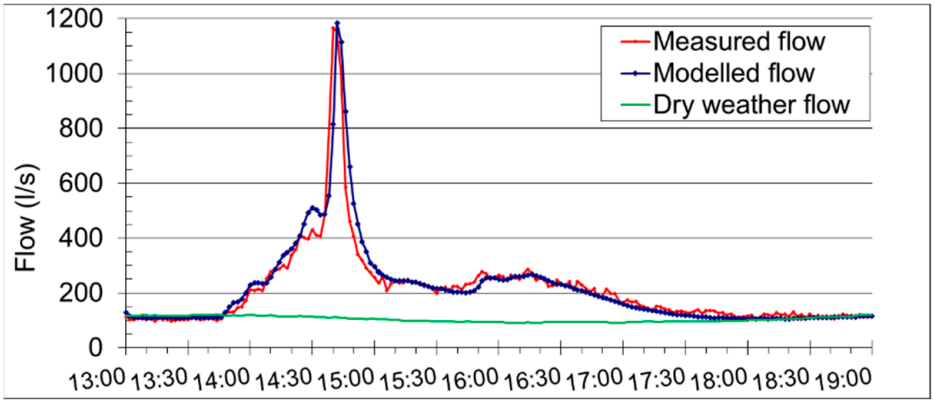


Figure 2. Measured and modelled hydrographs during a rainfall event.

Table 1. Statistical results for the four scenarios of synchronization of the time series in Figure 2.

Measured Flow	#	Mean	RMSE	I95	CV_{RMSE}	RSR	NSE	KGE	PBIAS	Slope	y-Interc.	r^2
		(L/s)	(L/s)	(L/s)	(L/s)	(%)	(-)	(-)	(%)	(-)	(L/s)	(-)
As recorded	181	211	49	97	23%	32%	0.90	0.94	−1.2%	0.99	4.9	0.91
Advanced in 2-min	181	211	28	55	13%	18%	0.97	0.96	−1.2%	1.02	−2.2	0.97
Delayed by 4-min	181	211	111	221	52%	72%	0.48	0.75	−1.1%	0.78	49.0	0.56
Delayed by 6-min	181	211	126	251	60%	82%	0.33	0.67	−1.1%	0.70	64.9	0.46

According to the results in Table 1, scenario (b) is the one that leads to the best results, with values of NSE, KGE, slope and coefficient of determination very close to the unity. However, NSE is less than 0.5 and 0.35 for scenarios (c) and (d), respectively, with only 4 and 6 min of rainfall delay. The error values also rise significantly for scenarios (c) and (d).

These results highlight the significant impact that small time deviations between measured and simulated series can have on the results of various metrics.

4. Materials and Methods

4.1. The Proposed New Approach

In a context of an increasingly widespread adoption of decentralized and nature-based solutions, the measures to be modelled will influence the entire urban water cycle, covering small to heavy rainfall. Therefore, assessing the shape of hydrographs for a wide range of rainfall events is of great importance. In order to strengthen the assessment of the shape of hydrographs and reduce the inconveniences of the event-by-event analysis described above, a new approach is proposed to assess the quality of hydrological models.

Rather than the performance metrics being applied to compare measured and simulated values within each hydrograph (and/or the measured and simulated peak flows of the various hydrographs), they could be applied to compare measured and simulated maximum flows for various durations. For each duration, the measured and simulated maximum flow series can be easily calculated by applying a rolling-window search routine to each hydrograph.

Hence, the assessment of model results is performed simultaneously for a pre-selected set of durations from all hydrographs. To avoid excessive complexity in the analysis, it is important to use a limited but representative number of durations, so we recommend selecting five to eight durations with increasing intervals between them. For the case study presented below, the maximum flow rates associated with the following durations will be assessed: 2, 6, 16, 30, 60, 104 and 150 min.

Table 2 presents an example of the application of the metrics described in Section 2 to the durations selected in the case study, according to the proposed approach. In addition to numerical metrics, graphical techniques must also be applied, as mentioned above and will be presented in the case study.

Table 2. Example of the application of the metrics described in Section 2 according to the proposed new approach.

Variable	Units	*n*	Mean	RMSE	I95	CV_{RMSE}	RSR	NSE	KGE	PBIAS	Slope	y-Interc.	r^2
		(#)	(Units)	(Units)	(Units)	(%)	(%)	(-)	(-)	(%)	(-)	(Units)	(-)
Volume	(m^3)	26	2509	321	643	13%	19%	0.96	0.85	7.5%	0.87	137.7	0.99
2-min peak	(L/s)	26	228	37	74	16%	58%	0.66	0.84	9.5%	0.97	−14.5	0.81
6-min peak	(L/s)	26	224	34	69	15%	55%	0.70	0.85	8.2%	0.98	−13.7	0.82
16-min peak	(L/s)	26	217	31	63	14%	51%	0.74	0.88	7.4%	0.94	−2.2	0.82
30-min peak	(L/s)	26	209	28	56	13%	46%	0.79	0.86	7.5%	0.84	17.3	0.85
60-min peak	(L/s)	25	197	29	59	15%	47%	0.78	0.76	8.5%	0.74	34.9	0.87
104-m peak	(L/s)	25	182	30	60	16%	50%	0.75	0.71	9.6%	0.69	39.3	0.88
150-m peak	(L/s)	24	126	26	51	20%	46%	0.79	0.72	9.3%	0.71	33.6	0.92

By evaluating peak flows for a wide range of durations, this new approach favours the assessment of the shape of hydrographs, as well as the effect of the uncertainty of both base flows and CSO discharges.

This new approach will be applied to assess eight different models (or modelling conditions) of the case study.

4.2. Case Study

The study area is located at Odivelas, a 26.5 km^2 municipality of the Lisbon Metropolitan Area, Portugal, and is 102 ha in size. It consists of two distinct catchments: a combined catchment, with 22 ha, of which the main sewer receives the foul flow from a 400 mm

upstream interceptor sewer; a mixed and partially separate catchment upstream, with about 80 ha, served by the mentioned interceptor sewer. For wet weather, the interceptor sewer transports a mixture of wastewater and stormwater, sometimes under pressure.

The two catchments have mainly a residential occupation with some commerce. In the downstream combined catchment, most sewers are built of concrete (Manning coefficient of 0.014 $s \cdot m^{-1/3}$) and have a circular cross-section of 300 and 400 mm, increasing up to 1000 mm downstream. The sewer slopes are close to those of the terrain, ranging from 0.3 to 11% and with a mean and median of 2.8% and 2.3%, respectively. The percentages of paved, roofed and green areas are 46%, 35%, and 19%, respectively. However, only about 70% of the runoff from impermeable areas drains into the sewer network, due to drainage to backyards and the insufficient number of inlet devices.

Over a decade ago, the downstream combined catchment was modelled in detail using the Stormwater Management Model (SWMM) [59], with 86 sub-catchments, 145 nodes and 153 sewer branches. However, the drainage system of the upstream mixed catchment is complex and is not known in detail, so it was modelled in an aggregated way using SWMM, considering only the two main combined sewer overflows (CSO) and two sub-catchments, with 68 ha and 12 ha (Figure 3).

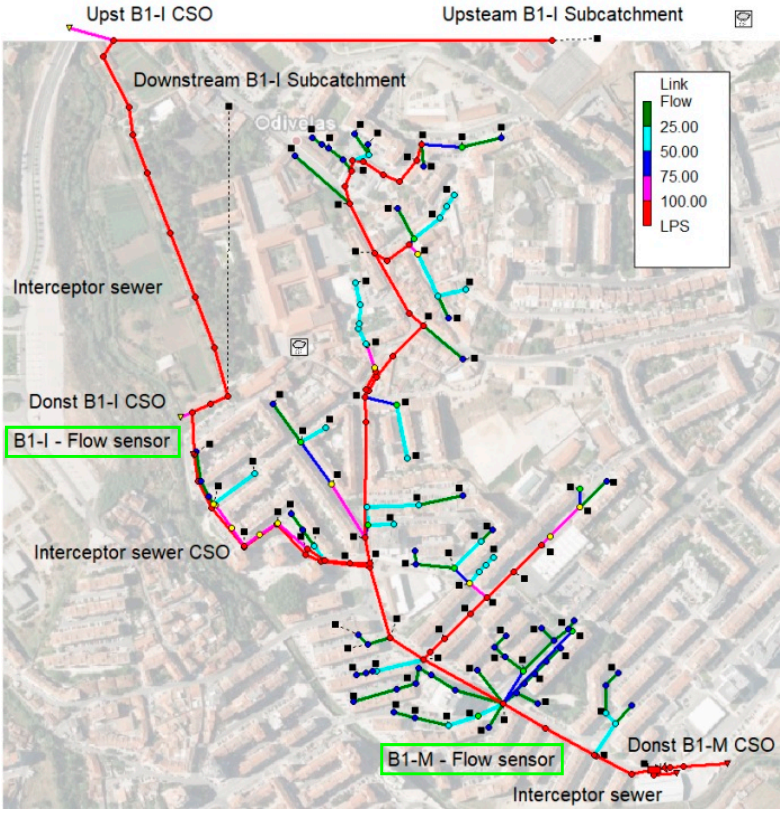


Figure 3. SWMM mathematical model and location of measurement sections.

For both catchments, the model was calibrated and verified on basis of data from a 4-month monitoring survey, in which 26 rainfall events were recorded by two rain gauges and two flowmeters. One flow meter was installed in the interceptor sewer, a few meters upstream of the entrance to the combined catchment (section B1-I), and the other was installed downstream from the combined catchment (section B1-M).

The peak flow of the most intense monitored event reached 80% of the maximum capacity of the combined system, of 1530 L/s, in a sewer downstream B1-M (and 60% of the capacity in B1-M, which already has a maximum diameter of 1 m). Downstream B1-M, the sewer is under pressure for return periods greater than 2–5 years and flooding occurs for return periods greater than 5–10 years.

As the complexity of the upstream catchment behavior did not allow us to obtain good calibration and verification results in the section B1-I, part of the underestimation of the flows in the interceptor sewer was compensated by some overestimation of the flows in the downstream combined catchment, allowing us to obtain very good results in B1-M. Thus, the model was left with a "black box" component inside, but it was quite adequate for the purpose of the study at the time. The model was used to evaluate the CSO discharges from the downstream catchment, both by event-by-event analyses [60,61] and using a 19-year rainfall historical series [62,63].

Currently, the model is intended to be used to study stormwater management measures distributed within the combined catchment, with a view to reduce CSO discharges, mitigate floods and improve the urban water cycle. Therefore, it is important to improve the calibration of the upstream aggregated model (modelled with only the two main CSO and two sub-catchments) and, hence, to model more accurately the flows generated in the combined catchment downstream. Between sections B1-I and B1-M there is a small CSO structure that shaves off the highest flows measured in B1-I, which could not be monitored and has not been modelled in the past. This CSO adds complexity to the model and its calibration requires a detailed quantification of the shape of the hydrographs in B1-I and B1-M.

The variability of base flows during rainfall events also plays an important role both in the estimation of the wet weather overflow discharges and in the calibration of this CSO structure. Figure 4 shows the hourly variation of the average and median flow (Qav and Qmedian) for the weekdays, as well as the 10th, 25th, 75th and 90th percentiles (Q10, Q25, Q75 and Q90). The horizontal lines represent the same statistics for the daily flows.

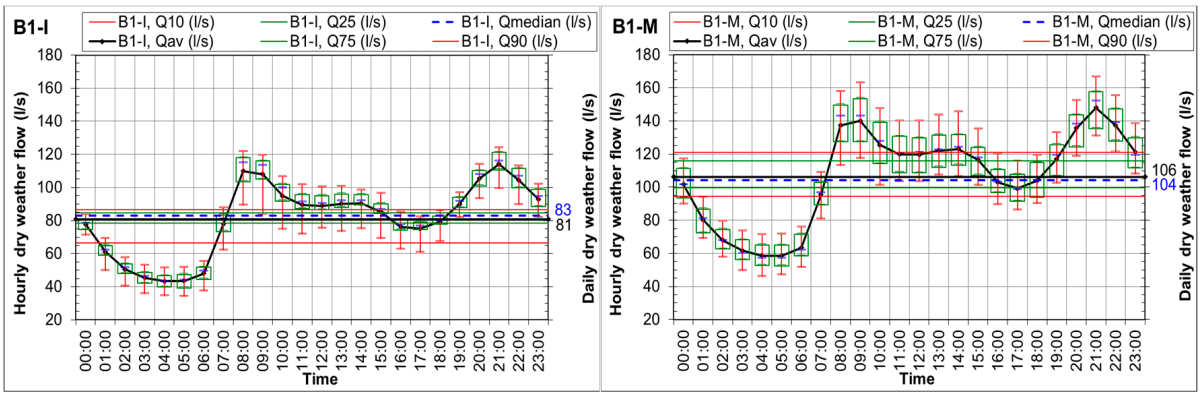


Figure 4. Hourly and daily dry weather flow statistics for weekdays in section B1-I and B1-M.

As Figure 4 shows, there is substantial variability in the daily dry weather flows, particularly in the downstream section. This variability is attributed to three main factors: the activities in the catchment; the groundwater infiltration into the sewers, although there is not a sufficiently long series to make it possible to model the RDII component; as well as to the measurement error, which significantly depends on the cleanliness and the accumulation of debris on the submerged pressure and velocity gauges.

4.3. Model Recalibration and Verification

Nine from the 26 events were selected for the sensitivity analysis of the parameters and for the recalibration of the models: events 1, 4, 7, 9, 13, 16, 19, 21 and 24. The selected events include five of the eight events that led to flow rates above the discharge threshold of the CSO structure located between sections B1-I and B1-M.

The other 17 events were used to verify the models.

The recalibration of the partially separate upstream catchment consisted mainly of adjusting the contributing areas of the two sub-catchments and the flow capacity of the respective CSO structures. The recalibration of the downstream catchment consisted mainly of calibrating the discharge capacity of the CSO structure between B1-I and B1-M, adjusting

the impermeable area and improving the shape of the hydrographs, through the slope and width of the sub-catchments.

Calibration was carried out manually based on volumes, peak flows and the shape of the hydrographs.

All 26 events were used to assess the quality of the models based on the proposed new approach described in Section 4.1, given that only five calibration and three verification events led to discharges in all CSO structures.

However, during the quality assessment of the global model in B1-M by the proposed approach and using the 26 events, it was determined that a small correction to the shape of the hydrographs should be done by increasing the slope and the width of the catchments and slightly decreasing the contribution of the impervious area. Hence, the set of 26 events was initially used to assess the quality of the model and later to enhance the calibration.

If the monitored stormwater event set were large enough that it could be split into two representative subsets of at least 20 events each, it would be recommended to split it into two subsets, one for model calibration and one for verification.

Performance metrics were applied to all rainfall events, but the results were substantially influenced by the time lags between the measured and modelled hydrographs, as described in Section 3. No attempt was made to synchronize the simulated and measured hydrographs, due to the subjectivity this would introduce. These results are presented in Tables A1 and A2 of Appendix A and will be discussed in Section 5.4.

4.4. Application of the Proposed New Approach

The new approach described in Section 4.1 was applied to eight quality assessments of case study models.

For both monitoring sections B1-I and B1-M, three assessments were carried out, two to evaluate the quality of the initial and recalibrated models, in which the dry weather flow (DWF) is adjusted event by event, and the third to evaluate the accuracy of the recalibrated model results without DWF adjustment.

For the downstream monitoring section B1-M, two additional assessments were carried out: one with a recalibrated model of the upstream interceptor sewer catchment, but still with the initial model of the downstream combined catchment (with adjustment of DWF); and another with the model recalibrated downstream, but receiving from the interceptor sewer the inflows measured in B1-I (with the DWF adjusted only in B1-M).

Table 3 lists the order in which the results of the eight assessments will be presented and discussed in the next section.

Table 3. Assessments carried out.

Assessments in B1-I	
C1	Results in **B1-I** of the **initial model with DWF adjustment** per event in B1-I
C2	Results in **B1-I** of the **recalibrated model with DWF adjustment** per event in B1-I
C3	Results in **B1-I** of the **recalibrated model without DWF adjustment** (to assess the accuracy of the results without any measurement information)
Assessments in B1-M	
C4	Results in **B1-M** of the **initial model with DWF adjustment** per event in B1-I and B1-M
C5	Results in **B1-M** of the **initial downstream model**, **but** with the **recalibrated interceptor sewer model** (with **DWF adjustment** per event in B1-I and B1-M)
C6	Results in **B1-M** of the **recalibrated model with DWF adjustment** per event in B1-I and B1-M
C7	Results in **B1-M** of the **recalibrated model without DWF adjustment** (to assess the accuracy of the results without any measurement information)
C8	Results in **B1-M** of the **recalibrated downstream model**, but receiving from the interceptor sewer the **inflows measured in B1-I** (**DWF adjustment** only for B1-M)

5. Results and Discussion

5.1. Results of the Upstream Model in B1-I

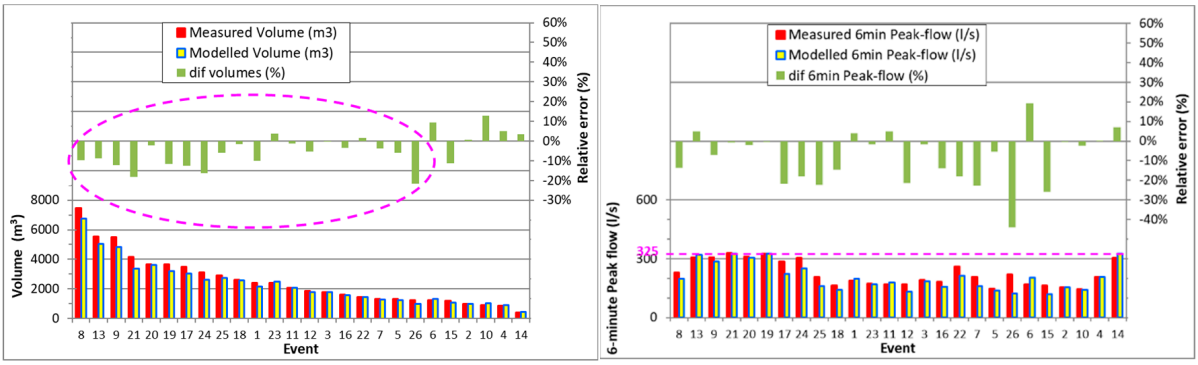
Table 4 presents the results obtained in section B1-I, for the first three assessments described in Table 3. The coefficients and statistics are those described in Section 2. The variables analysed are the volume of the hydrograph (Volume, in m^3) and the maximum flows (*i*th-min peak, in L/s) associated with the following durations (*i*th): 2, 6, 16, 30, 60, 104 and 150 min. The 2 min duration corresponds to the recording time interval of both the monitored data and the model results.

Table 4. Statistical results in the upstream monitoring section B1-I.

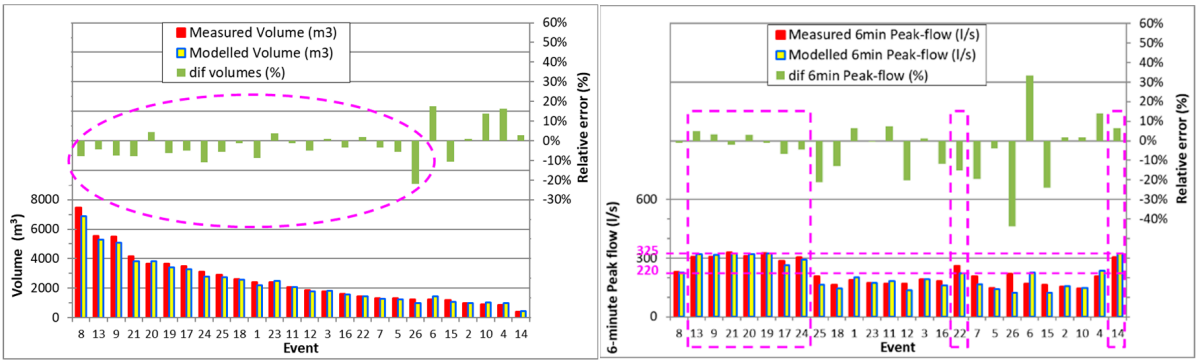
C1. Initial Upstream Model with Dry Weather Flow Adjustment per Event													
Variable	**Units**	***n***	**Mean**	**RMSE**	**I95**	**CV_{RMSE}**	**RSR**	**NSE**	**KGE**	**PBIAS**	**Slope**	**y-Interc.**	**r^2**
in B1-I		**(#)**	**(units)**	**(units)**	**(units)**	**(%)**	**(%)**	**(-)**	**(-)**	**(%)**	**(-)**	**(units)**	**(-)**
Volume	(m^3)	26	2509	321	643	13%	19%	0.96	0.85	7.5%	0.87	137.7	0.99
2-min peak	(L/s)	26	228	37	74	16%	58%	0.66	0.84	9.5%	0.97	−14.5	0.81
6-min peak	(L/s)	26	224	34	69	15%	55%	0.70	0.85	8.2%	0.98	−13.7	0.82
16-min peak	(L/s)	26	217	31	63	14%	51%	0.74	0.88	7.4%	0.94	−2.2	0.82
30-min peak	(L/s)	26	209	28	56	13%	46%	0.79	0.86	7.5%	0.84	17.3	0.85
60-min peak	(L/s)	25	197	29	59	15%	47%	0.78	0.76	8.5%	0.74	34.9	0.87
104-m peak	(L/s)	25	182	30	60	16%	50%	0.75	0.71	9.6%	0.69	39.3	0.88
150-m peak	(L/s)	24	126	26	51	20%	46%	0.79	0.72	9.3%	0.71	33.6	0.92
C2. Recalibrated upstream model with dry weather flow adjustment per event													
Variable	**Units**	***n***	**Mean**	**RMSE**	**I95**	**CV_{RMSE}**	**RSR**	**NSE**	**KGE**	**PBIAS**	**Slope**	**y-Interc.**	**r^2**
in B1-I		**(#)**	**(units)**	**(units)**	**(units)**	**(%)**	**(%)**	**(-)**	**(-)**	**(%)**	**(-)**	**(units)**	**(-)**
Volume	(m^3)	26	2509	214	427	9%	13%	0.98	0.91	4.0%	0.92	104.6	0.99
2-min peak	(L/s)	26	228	32	64	14%	51%	0.74	0.84	5.3%	1.02	−15.6	0.82
6-min peak	(L/s)	26	224	30	60	14%	48%	0.77	0.84	3.7%	1.03	−15.4	0.83
16-min peak	(L/s)	26	217	28	55	13%	45%	0.80	0.85	1.8%	1.04	−12.0	0.85
30-min peak	(L/s)	26	209	25	49	12%	40%	0.84	0.89	0.8%	1.00	−1.8	0.86
60-min peak	(L/s)	25	197	23	46	12%	37%	0.87	0.93	1.6%	0.93	11.0	0.87
104-m peak	(L/s)	25	182	21	42	12%	35%	0.88	0.92	3.2%	0.90	12.9	0.89
150-m peak	(L/s)	24	126	17	35	14%	31%	0.90	0.91	4.1%	0.89	12.0	0.92
C3. Recalibrated upstream model without dry weather flow adjustment													
Variable	**Units**	***n***	**Mean**	**RMSE**	**I95**	**CV_{RMSE}**	**RSR**	**NSE**	**KGE**	**PBIAS**	**Slope**	**y-Interc.**	**r^2**
in B1-I		**(#)**	**(units)**	**(units)**	**(units)**	**(%)**	**(%)**	**(-)**	**(-)**	**(%)**	**(-)**	**(units)**	**(-)**
Volume	(m^3)	26	2509	294	589	12%	18%	0.97	0.93	2.5%	0.92	125.1	0.97
2-min peak	(L/s)	26	228	37	74	16%	59%	0.65	0.81	4.2%	0.96	−0.5	0.74
6-min peak	(L/s)	26	224	36	71	16%	57%	0.67	0.81	2.6%	0.98	−0.6	0.75
16-min peak	(L/s)	26	217	34	67	15%	54%	0.70	0.83	0.8%	0.98	3.2	0.76
30-min peak	(L/s)	26	209	31	61	15%	50%	0.75	0.87	0.0%	0.94	12.5	0.78
60-min peak	(L/s)	25	197	29	58	15%	46%	0.79	0.89	0.8%	0.88	22.9	0.80
104-m peak	(L/s)	25	182	28	55	15%	46%	0.79	0.87	2.6%	0.83	25.4	0.80
150-m peak	(L/s)	24	126	24	48	19%	43%	0.82	0.86	3.4%	0.82	24.3	0.83
value	Value better than that of the previous analysis												
value	Value worse than that of the previous analysis												

Figure 5 presents two graphs for each of the assessments carried out in B1-I. The graphs compare the measured and modelled values and show the relative errors of the "volume" and "6-min peak flow" variables, for each of the 26 rainfalls arranged by descending order of the volume.

C1. Initial upstream model with dry weather flow adjustment per event.



C2. Recalibrated upstream model with dry weather flow adjustment per event.



C3. Recalibrated upstream model without dry weather flow adjustment.

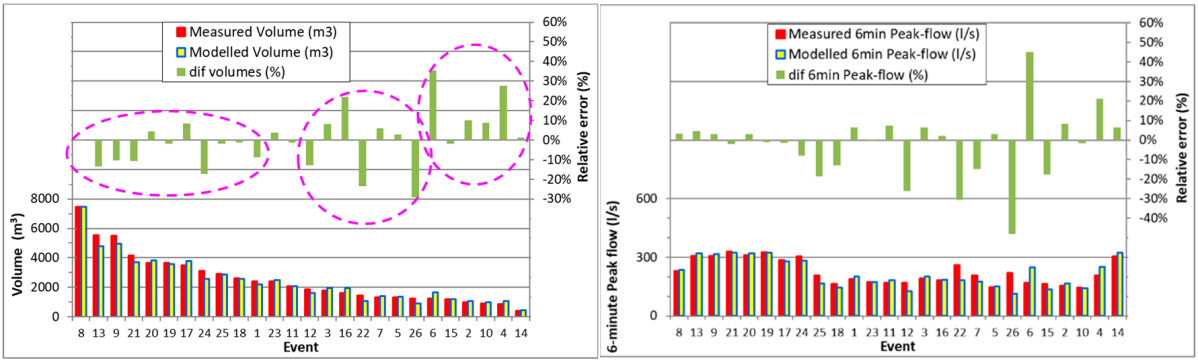


Figure 5. Graphical results for volume and 6-min peak flow in monitoring section B1-I.

5.1.1. Initial Upstream Model with Dry Weather Flow Adjustment per Event (C1)

As written in the description of the case study in Section 4.2, the initial model significantly underestimates the results in section B1-I, in particular for some events. Based on a qualitative appreciation of hydrographs in B1-I, the authors of this work would classify the initial model as providing a useful but limited and underestimated approximation. However, in a blind evaluation using the thresholds from Moriasi et al. (2007) [36] and Moriasi et al. (2015) [39], the model would be classified as "good" and "very good" for all PBIAS, NSE and r^2.

On closer analysis to Table 4, the model's shortcomings are mainly reflected on:

- In accordance with the model's underestimation, the PBIAS values are positive for all durations. However, given that the interceptor capacity on B1-I is limited to about 320 L/s and that CSO discharges occur upstream, the variation in flows at the interceptor sewer is limited when compared to the base flows and, therefore, PBIAS values do not exceed 9.6%.

- RSR error indicators are greater than 46% for all peak flow durations, reaching 58% for the shortest duration. However, RSR is only 19% for the volume variable, probably because the duration of events is variable, causing the average of measured volumes to be much higher than the model errors for the smallest events.
- Linear regression slopes are less than 0.84 for durations greater than 30 min, reflecting an increasing underestimation bias with the hydrograph duration. For the volume, the linear slope increases to 0.87 (with r^2 of 0.99) probably due to the greater influence of base flows and to the explanation given above.
- The coefficients of determination are less than 0.9 for almost all durations, although they increase with duration and reach 0.99 for the volume. They are close to 0.8 for the shortest durations, showing some dispersion of results. This dispersion is attributed to the difficulty in the aggregated model covering the variety of situations that occur in the partially separate upstream system. Although these values have good statistical significance and are also well classified according to Moriasi et al. (2015) [39], they should be interpreted with caution due to the great weight of base flows relative to wet weather flows.
- Except for the volume, where NSE is 0.96, the NSE values are below 0.8 for all durations, reaching 0.70 and 0.66 for, respectively, 6 and 2 min peak flows. Although these NSE values are classified as "good" according to [39] (except NSE = 0.66), they reflect the influence of the base flow and they are much lower than those obtained in the downstream section (B1-M), as will be seen below. These results indicate that within the scope of this new approach, in which NSE is not used to analyze errors in each hydrograph, but errors in pre-selected parts of the various hydrographs, NSE values below 0.8 should not be considered as "good", but simply as "satisfactory".
- KGE values are between 0.71 and 0.88. Unlike the NSE, the lowest KGE values occur for the shortest durations.

5.1.2. Recalibrated Upstream Model with Dry Weather Flow Adjustment per Event (C2)

In Table 4, the results from the recalibrated model that are better than the results from the original model are shaded in green. The results that are worse are shaded in orange. The recalibrated model provides better results on B1-I than the initial model in virtually all statistics and for all durations. A substantial improvement stands out in the less satisfactory statistics of the initial model.

However, the recalibrated model continues to show an underestimation trend due to:

- The increase in the base flow occurring during and after major rainfall events, which is attributed to the groundwater infiltration into the sewer network. Although the model acceptably represents the tail of some hydrographs, there are not enough events to model the RDII component.
- The interceptor sewer capacity being limited to roughly 320 L/s, and, therefore, the model deviations for the most intense peak flows also tend to be limited (they can be slightly positive only in the cases where the model results extend over time with this threshold value) (see Figure 5).

Based on a qualitative assessment of the hydrographs in B1-I, the authors would classify the recalibrated model as providing results that tend to be good, but with deviations and limitations for some events.

A more detailed analysis of the results included in Table 4 shows that:

- The PBIAS values are less than 4% for all durations, except for the 2 min one (which is 5.3%), evidencing the much smaller underestimation of the model. However, the PBIAS values provided by this new approach cannot be compared with the thresholds in [39] (where the rating would be "very good"), because the underestimation of the largest events is quite muffled in the set of all events.
- For durations of up to 30 min, the slopes of the regression line are between 1.00 and 1.04 and the interceptions remain reduced. For the maximum flows over 60 min and

for the volume, the slope became greater than 0.9, reflecting improvements over the initial model.

- The coefficients of determination improved only slightly compared to the initial model, remaining below 0.9 for almost all durations and increasing with duration.
- RSR error indicators remain relatively high, but significantly lower than for the initial model, particularly for longer durations.
- Only for the 2 and 6 min durations did the NSE values remain below 0.8. For the volume variable, the NSE increased to 0.98. However, based on the hydrograph analysis, it would be abusive to classify these results as "very good" according to [39].
- KGE values have increased to the range between 0.84 and 0.93. While in the initial model the lowest KGE values occurred for the longest durations, in the recalibrated model, KGE values above 0.9 occurred for the longest durations, highlighting the effects of the model recalibration. However, it is for longer durations that the model continues to behave worse (due to not modelling the RDII component), which highlights some limitations of these aggregate metrics and the misinterpretation that can result if they are used alone.

5.1.3. Recalibrated Upstream Model without Dry Weather Flow Adjustment (C3)

The use of the model without base flow adjustment corresponds to the standard situation of its use and, therefore, it is the situation for which the precision of the results should be quantified.

As might be expected, not adjusting the dry weather flows leads to worsening of virtually all statistics (shaded in orange in Table 4), with great significance for some. Except for the volume variable, almost all NSE values are now less than 0.8, with some being less than 0.7, and almost all coefficients of determination are less than 0.8. For durations greater than 60 min, the slopes of the regression line are less than 0.9. Interestingly, PBIAS improves slightly for all durations, indicating a lower underestimation.

The CVRMSE values are between 15% and 19% for all durations and are 12% for the volume. As the estimated interval of errors at a 95% probability is I95 $\approx$ 2·RMSE, this means that I95 is between 30% and 38% of the mean of the flows associated with each duration and is 24% of the mean of the measured volumes.

A significant increase in the RMSE (and I95) values is noteworthy, in relation to the model with base flow adjustments. For the Volume variable, RMSE rises from 214 m^3 to 294 m^3, corresponding to a 38% increase in the error. The relative error of RMSE increases with the duration of the hydrograph tip, increasing from 16% for the 2 min duration (from 32 L/s to 37 L/s) to 41% for the 150 min duration (from 17 L/s to 24 L/s).

These results indicate that an important component of the uncertainty of the model results in B1-I comes from the variability of the base flows.

5.2. Results of the Global Model in B1-M

Table 5 presents the statistical results obtained for section B1-M, for the assessments C4 to C8 described in Table 3.

Figure 6 compares the measured and modelled values for both the volume and the 6 min peak flow, for the 26 precipitation events of the recalibrated model. Results are presented for both assessments with and without dry weather flow adjustment (C6 and C8). The graphs of the other assessments are not presented as they do not add value to this discussion.

Table 5. Statistical results in the downstream monitoring section B1-M.

C4. Initial Global Model with Dry Weather Flow Adjustment per Event

Variable	Units	*n*	Mean	RMSE	I95	CV_{RMSE}	RSR	NSE	KGE	PBIAS	Slope	y-Interc.	r^2
in B1-M		**(#)**	**(units)**	**(units)**	**(units)**	**(%)**	**(%)**	**(-)**	**(-)**	**(%)**	**(-)**	**(units)**	**(-)**
Volume	(m^3)	26	3350	251	503	8%	12%	0.99	0.93	1.7%	0.93	171.6	0.99
2-min peak	(L/s)	26	383	42	83	11%	16%	0.98	0.98	1.5%	0.98	2.7	0.98
6-min peak	(L/s)	26	362	33	65	9%	13%	0.98	0.98	−0.6%	0.97	12.6	0.98
16-min peak	(L/s)	26	327	27	55	8%	15%	0.98	0.96	−1.7%	1.02	−1.0	0.98
30-min peak	(L/s)	26	301	24	48	8%	17%	0.97	0.96	−1.7%	1.03	−2.5	0.98
60-min peak	(L/s)	26	268	21	41	8%	19%	0.96	0.95	−2.0%	1.03	−3.5	0.97
104-m peak	(L/s)	26	238	16	31	7%	19%	0.97	0.97	−0.7%	1.01	0.2	0.97
150-m peak	(L/s)	25	165	13	26	8%	17%	0.97	0.98	−0.2%	0.97	7.2	0.97

C5. Initial downstream model, but with the recalibrated interceptor sewer model (with DWF adjustment)

Variable	Units	*n*	Mean	RMSE	I95	CV_{RMSE}	RSR	NSE	KGE	PBIAS	Slope	y-Interc.	r^2
in B1-M		**(#)**	**(units)**	**(units)**	**(units)**	**(%)**	**(%)**	**(-)**	**(-)**	**(%)**	**(-)**	**(units)**	**(-)**
Volume	(m^3)	26	3350	238	477	7%	11%	0.99	0.97	−0.9%	0.96	148.9	0.99
2-min peak	(L/s)	26	383	43	87	11%	16%	0.97	0.97	−0.2%	0.96	14.7	0.97
6-min peak	(L/s)	26	362	37	75	10%	15%	0.98	0.96	−2.5%	0.96	24.2	0.98
16-min peak	(L/s)	26	327	34	67	10%	18%	0.97	0.95	−4.3%	1.02	8.3	0.97
30-min peak	(L/s)	26	301	32	64	11%	22%	0.95	0.92	−5.1%	1.05	1.7	0.97
60-min peak	(L/s)	26	268	31	62	12%	29%	0.92	0.86	−6.4%	1.11	−11.4	0.96
104-m peak	(L/s)	26	238	24	48	10%	28%	0.92	0.84	−5.3%	1.13	−17.9	0.97
150-m peak	(L/s)	25	165	18	36	11%	23%	0.95	0.89	−4.2%	1.09	−9.7	0.97

C6. Recalibrated global model with dry weather flow adjustment per event

Variable	Units	*n*	Mean	RMSE	I95	CV_{RMSE}	RSR	NSE	KGE	PBIAS	Slope	y-Interc.	r^2
in B1-M		**(#)**	**(units)**	**(units)**	**(units)**	**(%)**	**(%)**	**(-)**	**(-)**	**(%)**	**(-)**	**(units)**	**(-)**
Volume	(m^3)	26	3350	**228**	**456**	**7%**	**11%**	0.99	**0.95**	**0.2%**	**0.94**	184.1	0.99
2-min peak	(L/s)	26	383	55	109	14%	20%	0.96	0.94	−3.4%	1.03	**2.2**	0.96
6-min peak	(L/s)	26	362	40	79	11%	16%	0.97	0.96	−4.3%	**0.99**	20.2	0.98
16-min peak	(L/s)	26	327	34	68	10%	19%	0.97	0.95	−4.0%	**1.00**	13.3	0.97
30-min peak	(L/s)	26	301	29	58	10%	21%	0.96	0.96	−3.0%	0.97	19.4	0.96
60-min peak	(L/s)	26	268	25	50	9%	23%	0.95	**0.96**	−3.3%	0.96	18.9	0.95
104-m peak	(L/s)	26	238	18	36	8%	22%	0.95	0.97	−2.5%	0.97	13.0	0.96
150-m peak	(L/s)	25	165	15	30	9%	19%	0.96	0.96	−2.0%	0.95	15.0	0.97

C7. Recalibrated global model without dry weather flow adjustment

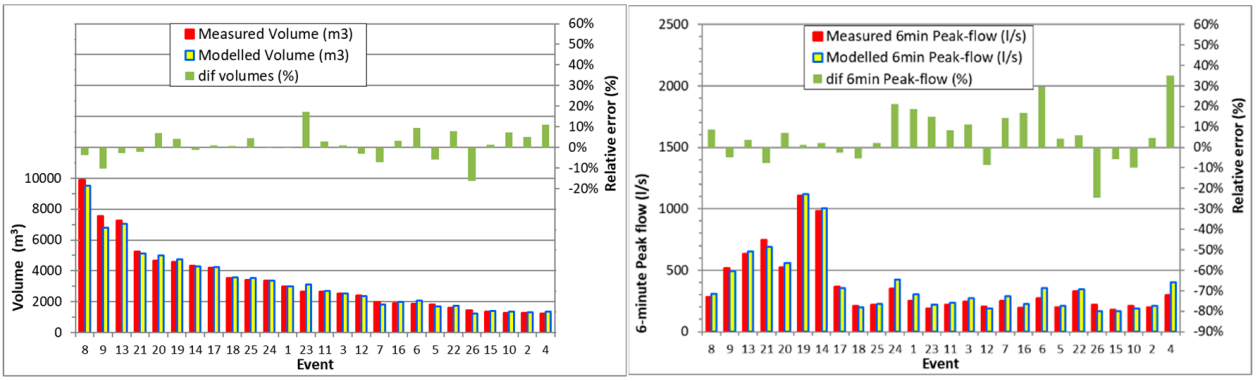
Variable	Units	*n*	Mean	RMSE	I95	CV_{RMSE}	RSR	NSE	KGE	PBIAS	Slope	y-Interc.	r^2
in B1-M		**(#)**	**(units)**	**(units)**	**(units)**	**(%)**	**(%)**	**(-)**	**(-)**	**(%)**	**(-)**	**(units)**	**(-)**
Volume	(m^3)	26	3350	421	842	13%	20%	0.96	0.95	−2.7%	0.95	273.8	0.96
2-min peak	(L/s)	26	383	59	119	16%	22%	0.95	0.93	−4.8%	1.02	10.0	0.96
6-min peak	(L/s)	26	362	46	92	13%	19%	0.96	0.94	−5.7%	0.98	27.5	0.97
16-min peak	(L/s)	26	327	40	80	12%	22%	0.95	0.94	−5.5%	0.99	21.4	0.96
30-min peak	(L/s)	26	301	35	71	12%	25%	0.94	0.94	−4.7%	0.95	27.8	0.95
60-min peak	(L/s)	26	268	31	62	12%	29%	0.92	0.94	−5.2%	0.95	26.9	0.93
104-m peak	(L/s)	26	238	26	51	11%	30%	0.91	0.94	−4.6%	0.96	21.2	0.93
150-m peak	(L/s)	25	165	23	47	14%	30%	0.91	0.94	−4.2%	0.93	23.6	0.92

C8. Recalibrated downstream model, but receiving the inflows measured in B1-I (DWF adjustment in B1-M)

Variable	Units	*n*	Mean	RMSE	I95	CV_{RMSE}	RSR	NSE	KGE	PBIAS	Slope	y-Interc.	r^2
in B1-M		**(#)**	**(units)**	**(units)**	**(units)**	**(%)**	**(%)**	**(-)**	**(-)**	**(%)**	**(-)**	**(units)**	**(-)**
Volume	(m^3)	26	3350	175	351	5%	8%	0.99	0.98	−2.0%	0.99	85.6	0.99
2-min peak	(L/s)	26	383	53	106	14%	20%	0.96	0.94	−5.7%	1.01	18.5	0.97
6-min peak	(L/s)	26	362	38	77	11%	16%	0.98	0.93	−6.6%	0.97	36.4	0.99
16-min peak	(L/s)	26	327	34	68	10%	19%	0.97	0.93	−6.6%	0.97	30.0	0.98
30-min peak	(L/s)	26	301	26	53	9%	19%	0.97	0.93	−4.8%	0.94	32.1	0.98
60-min peak	(L/s)	26	268	18	37	7%	17%	0.97	0.95	−3.6%	0.95	22.5	0.98
104-m peak	(L/s)	26	238	15	29	6%	17%	0.97	0.96	−3.1%	0.97	14.9	0.98
150-m peak	(L/s)	25	165	13	27	8%	17%	0.97	0.96	−3.1%	0.96	14.7	0.98

value	For C5, C6 and C7, the value is better than that of the previous analysis. For C8, the value is better than that of C6.
value	For C5, C6 and C7, the value is worse than that of the previous analysis. For C8, the value is worse than that of C6.
value	**For C6, the value is better than that of the initial model**
value	For C6, the value is worse than that of the initial model

C6. Recalibrated global model with dry weather flow adjustment per event.



C7. Recalibrated global model without dry weather flow adjustment.

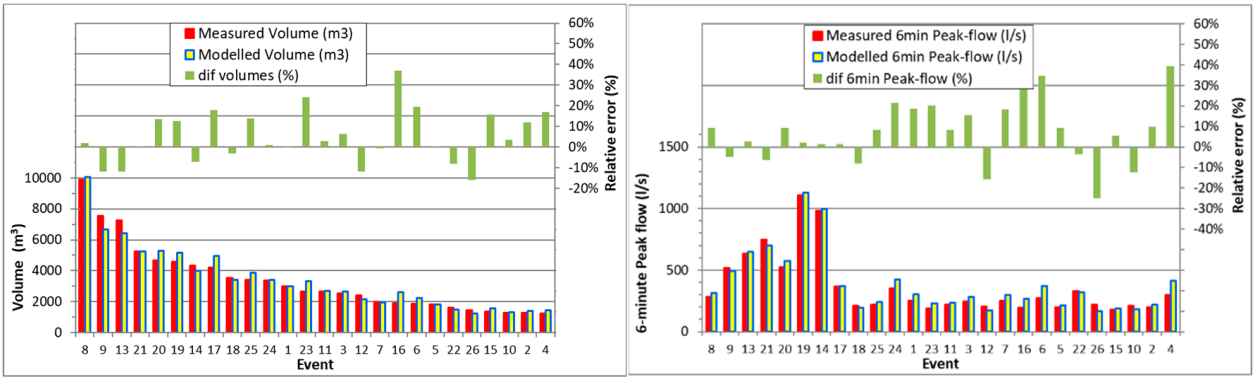


Figure 6. Graphical results for volume and 6-min peak flow of the recalibrated model in monitoring section B1-M.

5.2.1. Initial Global Model with Dry Weather Flow Adjustment per Event (C4)

Based on a qualitative but in-depth appreciation of the hydrographs in B1-M, the authors of this work would classify the initial model as providing very good results. The modelled flows are fairly coincident with the measured flows for many events and show minor deviations in smaller events. Most deviations result from the RDII component in the upstream catchment not being modelled.

The quality of the results of this model is confirmed by the very good values of all statistics for all durations: NSE between 0.96 and 0.99, KGE between 0.95 and 0.98, PBIAS from −1.7% to 2.0%, linear regression slope between 0.97 and 1.03, coefficients of determination not less than 0.97 and CVRMSE values between 8% and 11% (corresponding to I95 between 13% and 21% of the means).

For the volume variable, NSE is 0.99, although KGE is "only" 0.93. PBIAS is only 1.7%, but the slope of the linear regression is 0.93 (with r^2 = 0.99), reflecting the underestimation due to the RDII component not being modelled.

Both slope and KGE for volume are the only metrics that point to a slight bias that would not allow classifying the model as excellent and the volume is the only variable accumulated over different durations. This result seems to show the advantages of the new approach in including a variable that accumulates over the different durations of the events.

These results also call attention to the importance of the modeler's knowledge in assessing the model quality. Despite having more than 100 nodes and providing very accurate results in B1-M, the initial model has some usability limitations due to the black box component left during calibration (it provides underestimated results for the upstream catchment and compensates most of this deviation in the downstream combined catchment, as described in Section 4.2).

It is important to note that the proximity of the values of NSE, KGE and r^2 to optimal values is higher than in most cases in the literature and may derive from the approach applied in this work. As mentioned in Section 2.1, most of the NSE values reported in the calibration of hydrological or hydraulic urban drainage models range from 0.5 to 0.9, with most being greater than 0.7 [2,7,10,20,22,46–48,50–53]. However, NSE values above 0.95 are also reported [1,53,55].

5.2.2. Initial Downstream Model, but with the Recalibrated Interceptor Sewer Model (with DWF Adjustment) (C5)

With the recalibration of the upstream catchment, the accuracy of the initial model in the downstream section B1-M is significantly impaired (shaded in orange in the results of the C5 analysis of Table 5), except for the volume variable (shaded in green) due to the increase in the upstream flows.

As shown in the C5 results of Table 5, the worst results occur for the 60 and 104 min durations (with NSE of 0.92, KGE of 0.86 and 0.84, linear regression slopes greater than 1.11, and negative PBIES values with absolute values greater than 5%). The overestimation for these durations results from the model not taking into account that, during some events, part of the flows from the interceptor sewer is overflowed between B1-I and B1-M.

In fact, downstream of B1-I, there is a by-pass (in a manhole of the interceptor sewer) that allows overflowing to the stream when the interceptor sewer is under pressure. In fact, a detailed analysis of the linear slope and PBIAS results of the initial model (C4) already indicates a slight overestimation of the peak flows with durations between 16 and 60 min, in relation to the other durations.

Figure 7 compares the hydrographs in B1-M of two rainfall events for the following three cases, all with dry weather flow adjustment: C4) the initial model; C5) the initial model downstream, but with the upstream catchment recalibrated; C6) model with the new CSO in the interceptor sewer downstream B1-I and recalibrated. The hydrographs in Figure 7 highlight the excess of inflows to B1-M in the C5 case and the good results of the C6 case, after shaving the excess flow in the new CSO between B1-I and B1-M.

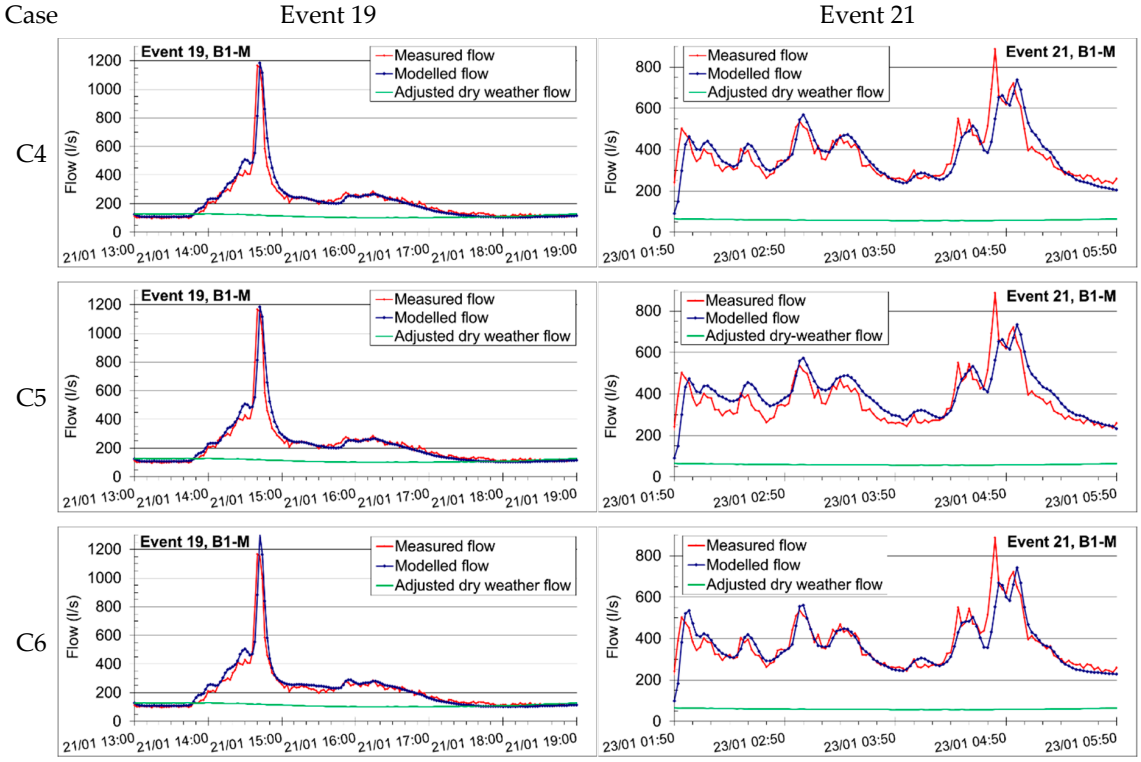


Figure 7. Hydrographs of two rainfall events for assessments C4, C5 and C6.

5.2.3. Recalibrated Global Model with Dry Weather Flow Adjustment per Event (C6)

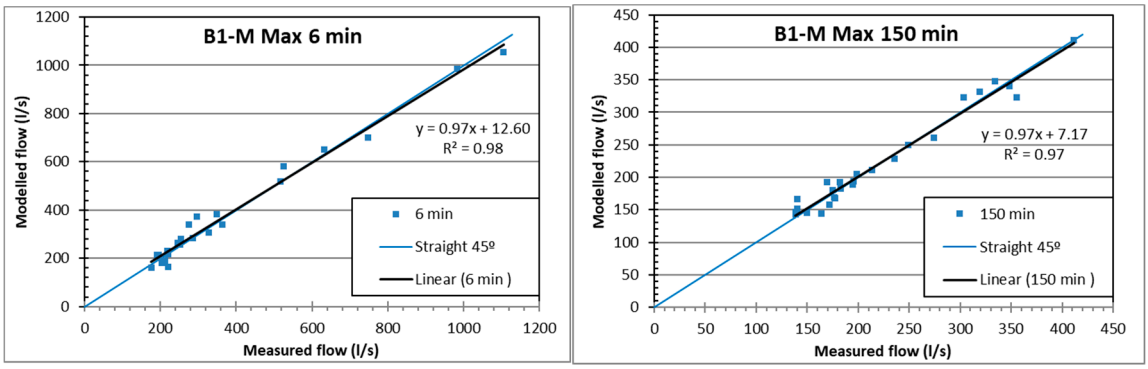
Based on the slopes of the interceptor sewer downstream B1-I and considering the Manning–Strickler coefficient for concrete $Ks = 1/0.014 = 71\ m^{1/3}{\cdot}s^{-1}$, the flow capacity of the interceptor sewer downstream B1-I was estimated to be about 220 L/s (about 100 L/s less than in B1-I). The downstream model was then recalibrated, obtaining less accurate results than the initial model, but also quite good.

In the C6 results of Table 5 (recalibrated model with dry weather flow adjustment), the results that are better than the C5 results (initial model downstream, but with the upstream catchment recalibrated) are shaded in green. The worst results are shaded in orange.

The new CSO structure and the recalibration of the downstream model led to a significant improvement in virtually all the statistics for both the volume and the flows lasting longer than 30 min. The slope has also improved for the 2, 6 and 16 min peak flows, but the other statistics are now worse for these durations. This is because a detailed analysis of the scatterplots between the measured and modelled values led to the decision to slightly increase the 6 min peak flow, compared to the initial model.

Figure 8 shows the scattergraphs and linear regression lines for the 6 and 150 min maximum flows, for both the initial model and the recalibrated model. For the 6 min peak flows, in the initial model the regression line has a slope of 0.97 and a slight underestimation of the highest values is observed. In the recalibrated model, the slight increase in the highest values of the 6 min peak flow is confirmed and the slope of the linear regression increased to 0.99. The y-intercept increased slightly to 20 L/s, mostly due to a slight overestimation of some lower magnitude 6 min peak flows.

C4. Initial model with dry weather flow adjustment per event.



C6. Recalibrated model with dry weather flow adjustment per event.

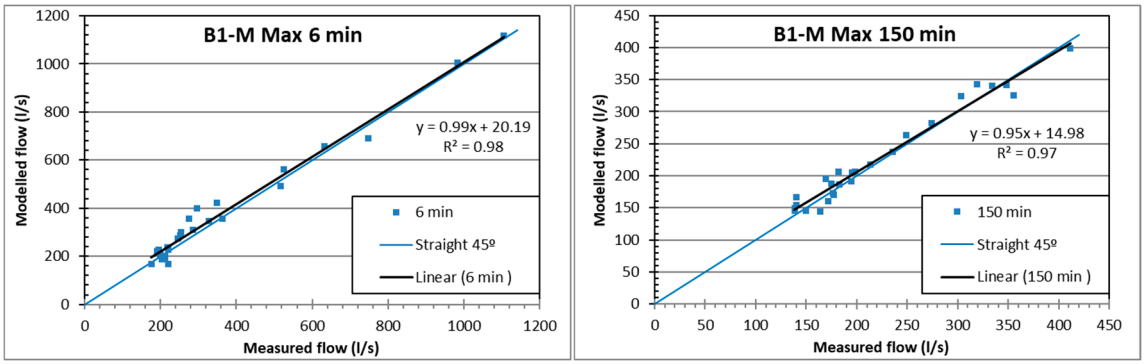


Figure 8. Scattergraphs of the 6 and 150 min maximum flows in B1-M from the initial and recalibrated models.

In Table 5, the results of C6 written in purple are worse than those of the initial model and the results written in bold are better than those of the initial model. Except for most of the volume variable values and the 6 min and 16 min peak flow slopes, almost all other statistics in the recalibrated model are worse (albeit slightly) than the initial model. This results from the increase in determinism, and consequent uncertainty, introduced by the CSO structure in the recalibrated model.

5.2.4. Recalibrated Global Model without Dry Weather Flow Adjustment (C7)

As expected, the lack of adjustment of the dry weather flow leads to a worsening of virtually all C7 results (shaded in orange in Table 5), with some significance for some. However, all statistics maintain values that should be classified as good. (All NSE values are greater than 0.91, decreasing with duration, KGE is between 0.93 and 0.95, slopes are equal to or greater than 0.95 (except for 150 min) and the regression coefficients range from 0.92 to 0.97. PBIAS values are between −2.7% and −5.7%.)

The CVRMSE values are between 11% and 16% for all variables, which means that I95 ≈ 2·RMSE is between 22% and 32% of the mean of the measured values for all variables.

A significant increase in the values of RMSE (and I95) is noteworthy in relation to the model with the base flow adjustments. For the volume variable, RMSE rises from 228 m^3 to 421 m^3, corresponding to an 85% increase in the error. The relative error of RMSE increases with the duration of peak flow, growing from 7% for the 2-min duration (from 55 L/s to 59 L/s) to 53% for the 150-min duration (from 15 L/s to 23 L/s).

These results show that a significant percentage of the errors in B1-M result from the variability in the base flows.

In spite of the values of several metrics decreasing with some significance in relation to assessment C6, the values of NSE, KGE, r^2 and probably the PBIAS remain relatively close to optimal values, being above most of the values in the literature.

The analysis of the set of results from Tables 4 and 5 and from Tables A1 and A2 of Appendix A shows the potential of the proposed approach, but indicates that it delivers results closer to optimal values than when applying metrics to each rainfall event.

5.2.5. Recalibrated Downstream Model, but Receiving the Inflows Measured in B1-I (DWF Adjustment in B1-M) (C8)

If the measurement errors in B1-I and B1-M were null, the results of the C8 analysis would deliver the errors of the downstream model, and the differences in relation to the C6 case (recalibrated model with DWF adjustment) would reflect the effect of the errors of the upstream catchment model on B1-M. In Table 5, the results of the C8 analysis that are better and worse than the results of C6 are shaded, respectively, in green and orange.

The results tend to be slightly better than the C6 ones for most statistics. However, they are worse for PBIAS, showing an overestimation trend for all durations of the downstream recalibrated model. As such, both the slope and the y-intercept are worse for practically all durations, except for 150 min and for volume, which benefit from the contribution of the infiltration flows and the RDII component. The coefficients of determination are 0.98 and 0.99 (except for the 2-min duration only, which is 0.97).

Similarly, KGE values are also better for volume, in principle due to the RDII component, and are worse for practically all durations, due to the model's tendency to overestimation.

Table 6 compares the RMSE values obtained for all variables between assessment C6 (recalibrated model with DWF adjustment) and the following two cases: C7 (model without DWF adjustment); C8 (model using the values measured in B1-I and adjusting the DWF only for B1-M).

Table 6. RMSE values in B1-M for assessment C6 and relative variations of RMSE for both assessment C7 and assessment C8.

Parameter in B1-M	Units	Mean	RMSE for C6	RMSE for C7		RMSE for C8	
		(Units)	(Units)	(Units)	(%)	(Units)	(%)
Volume	(m^3)	3350	228	421	(+85%)	175	(−23%)
2-min peak	(L/s)	383	55	59	(+7%)	53	(−4%)
6-min peak	(L/s)	362	40	46	(+15%)	38	(−5%)
16-min peak	(L/s)	327	34	40	(+18%)	34	(0%)
30-min peak	(L/s)	301	29	35	(+21%)	26	(−10%)
60-min peak	(L/s)	268	25	31	(+24%)	18	(−28%)
104-m peak	(L/s)	238	18	26	(+44%)	15	(−17%)
150-m peak	(L/s)	165	15	23	(+53%)	13	(−13%)

As expected, the RMSE values for C8 are very close to those obtained for C6 for the shortest durations (2, 6 and 16 min) and are smaller for the other durations. The RMSE reduction is maximum for the 60-min duration (28%), probably due to the combined effect of the RDII component and the improvement in the permeable areas infiltration component throughout the event.

However, the RMSE reductions in assessment C8 are much smaller than the RMSE increases when the dry weather flow adjustment is not carried out (C7). These results highlight the role of the base flow variability in model errors.

Part of the variability of the base flows, as well as the errors obtained in C8, results from measurement errors in B1-I and B1-M. Nor can it be excluded that part of the overestimation trend of the downstream model results from some overestimation of the average DWF of the combined catchment in relation to the average DWF coming from the upstream catchment.

5.3. Quality of Monitored Data

Figure 9 shows the dry weather flow adjustments considered for each event in sections B1-I and B1-M, as well as the adjustment difference between the two sections. Although the flow adjustments vary with some significance for some events (between ±40 L/s, i.e., around 50% and 40% of the dry weather average flow in B1-I and B1-M, respectively), the adjustment difference between sections B1-I and B1-M is relatively constant for all events: this difference is null until event 11; and ranges from −15 L/s to −10 L/s between events 15 and 26 (except for event 18, where it is 5 L/s, probably due to the accumulation of debris or grease on the sensors). The difference in results between these two different periods indicates a systematic error in measurements. This error is between 9 and 14% of the average dry weather flow in B1-M.

This systematic error was initially attributed to the fact of the ring that fixed the pressure, ultrasonic and velocity sensors to the sewer in B1-I being dragged by the flow during event 14, forcing it to be reinstalled under possibly slightly different conditions. However, the analysis of the records of water depth, pressure, velocity and flow in section B1-I, as well as the scattergraphs between these variables, did not allow identifying any bias between those two time periods. On the contrary, the analysis of the scattergraphs at B1-M suggests that the systematic error is likely due to some changes in measurement conditions at B1-M.

The new approach described above could also be used to assess the influence of systematic errors in measurements on the accuracy of the model results. This analysis is not presented to avoid overloading.

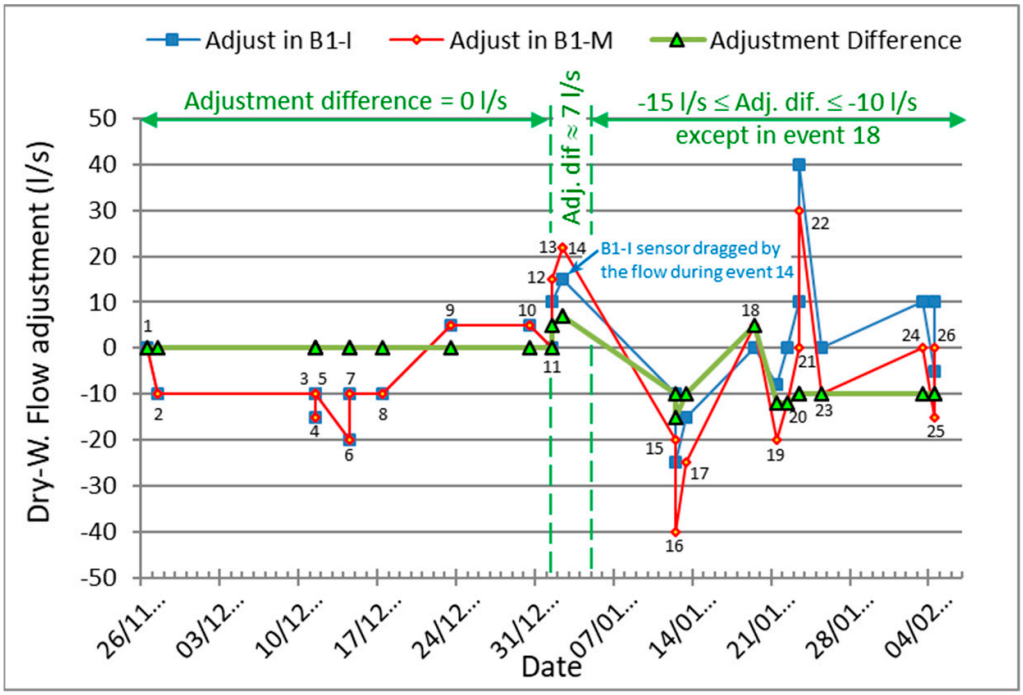


Figure 9. Values of the dry weather flow adjustment per modelled event.

5.4. Results of the Metrics Applied to Each Rainfall Event

Appendix A presents two tables containing the results of the performance metrics applied event by event to four cases. Table A1 presents and compares the results for all events in B1-I, for both the initial and the recalibrated model, the two with DWF adjustment (C1 and C2 analysis). Table A2 compares the results for all events in B1-M for the recalibrated model, with and without DWF adjustment (C6 and C7 assessments).

As described in Section 3, in small urban catchments, these results depend significantly on both time lags between rainfall and flow measurements and the temporal and spatial variability of rainfall within the catchment. The values presented in Tables A1 and A2 were not the subject of any attempt to synchronize the times between rainfall and flow measurements, although a short 2–4-min adjustment could significantly improve the results for some events. In cases like Figure 4, the lack of synchronism seems obvious, but in other cases this attempt at synchronization would be subjective and debatable due to the intrinsic variability in rainfall.

The qualitative ratings proposed by Moriasi et al. (2015) [39] are represented in the color of the NSE, PBIAS and r^2 results in Tables A1 and A2. Both B1-I and B1-M have a very wide range of results for most metrics, with many events rated as "very good" and many as "unsatisfactory".

However, the analysis of the hydrographs shows that the events classified as "unsatisfactory" usually correspond to small rainfalls, in which small errors of the flow or even of the base flow lead to significant relative errors. Despite such unsatisfactory results in an event-by-event analysis, the model even has a reasonable performance for the complexity and non-linearity of the simulated phenomena and associated uncertainty, as can be observed in Figure 10. This Figure compares the measured and the simulated hydrographs of the two events with the worst metrics (events 26 and 4). While in Table A2 these events make the analysis confusing and even biased or erroneous, in the proposed new approach they are all taken into account, but with the due importance (see Figures 6 and 8).

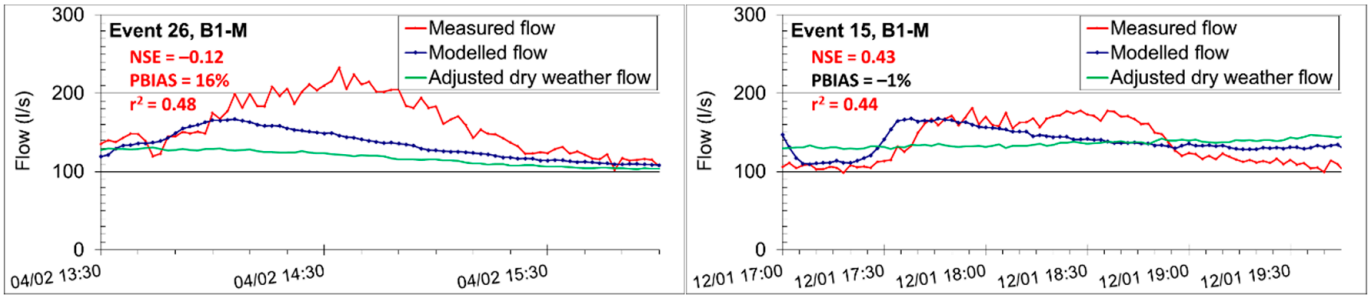


Figure 10. Measured and simulated hydrographs of the two events with the worst metrics.

From the predominance of green shading in Table A1, we can see that the recalibrated model is better than the initial model in B1-I, but interestingly, there are more events classified as "unsatisfactory" for NSE with the recalibrated model (10 events) than with the initial one (8 events). The results in Table A1 hardly guide the modeler on how to improve the model during calibration, contrary to the new approach presented here using the same metrics.

From the predominance of orange shading in Table A2, we can also see that the non-adjustment of the base flow impairs the quality of the results in B1-M. However, the results are scattered and do not allow us to give a quantitative or even qualitative idea of how much the quality is impaired, contrary to the proposed approach (see Sections 5.2.4 and 5.2.5).

6. Conclusions

Small urban catchments pose challenges in applying performance metrics when comparing measured and simulated hydrographs. Indeed, results are hampered by the short peak flows, due to rainfall variability and measurement synchronization errors, and it can be both difficult and inconvenient to remove base flows from the analysis, given their influence on the performance of CSO structures. In addition, base flows are an important source of uncertainty in modelling small rainfall events, which must be taken into account when assessing the quality of the models.

A new approach was proposed and tested to assess the quality of models of small combined catchments, which proved to be quite suitable not only for the assessment of the quality of the models, but also to support calibration. In the proposed approach, rather than the performance metrics being applied to compare the measured and simulated values within each hydrograph and/or the measured and simulated peak flows of the various events, they are applied to compare measured and simulated maximum flows for a set of different durations. For each duration, the measured and simulated maximum flow series can be easily calculated by applying a rolling-window search routine to each hydrograph. To keep the assessment simple, five to eight different durations with increasing intervals should be analysed.

This new approach presents the following advantages: (a) being simple; (b) avoiding the inconveniences arising from the time lags of very short peaks (described in Section 3) and the subjectivity of possible adjustments; (c) favouring the assessment of the influence of base flow and RDII variability (see assessments C2, C3 and C7) and the influence of peak flow shaving by upstream CSOs (assessment C5); (d) promoting and facilitating an integrated analysis for a wide range of rainfall events; (e) avoiding subjectivity in interpreting different results for the various events; and (f) making it possible to identify biases in simulated hydrographs that would otherwise be difficult to detect, also guiding calibration.

However, it has the disadvantage of requiring a sufficiently large and representative set of rainfall events to ensure statistical significance, which, in principle, should not be less than 20 events for evaluating the quality of model results or twice that number for a complete model calibration and verification.

In addition, the results delivered by this new approach should not be compared with the thresholds proposed in Moriasi et al. (2015) [39] without careful consideration, as the values of NSE, r^2 and PBIAS tend to be closer to optimal values than when applying metrics to compare measured and simulated values within each hydrograph.

This recommendation is extended to all modelers who apply performance metrics to peak flows in urban drainage systems.

In the application of the described new approach to a model classified as providing a useful but limited and underestimated approximation (assessment C1), the NSE values were below 0.8 for all durations and, for some, below 0.7. KGE values ranged from 0.7 to 0.85. PBIAS values were less than $\pm$10%. For the recalibrated model, providing results that tend to be good but with deviations and limitations for some events (assessment C2), the NSE values tended to be greater than 0.8, the KGE values tended to approach 0.9 and PBIAS were within $\pm$5%. Finally, for models of which the simulated and measured hydrographs are very coincident for some events and show small deviations in other events (assessments C4 and C6), both the NSE and the KGE values were higher than 0.95 for all durations, reaching 0.98 and 0.99 in some cases.

During normal use of the model, base flows are unknown. Without adjusting base flows (assessments C3 and C7), NSE values tended to fall by up to 0.1, depending on duration, while KGE values tended to vary much less. For the "very good" quality model (assessments C7 compared with C6), the RMSE values increased between 20% and 50% as the analysed duration increased to 150 min, due to the influence of the unmodelled RDII.

The various examples of the case study highlight the importance of using different metrics and graphical analyses and the pertinence of the proposed approach.

Author Contributions: Conceptualization, L.M.D. and T.M.M.; methodology, L.M.D. and T.M.M.; supervision, L.M.D.; validation, L.M.D. and T.M.M.; writing—original draft, L.M.D. and T.M.M.; writing—review & editing, L.M.D. All authors have read and agreed to the published version of the manuscript.

Funding: Part of this work was co-funded by the European Regional Development Fund (FEDER), under programs POR Lisboa2020 and CrescAlgarve2020, through Project SINERGEA (ANI 33595).

Acknowledgments: Thanks are due to Graça Tomé for her collaboration in the English revision.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Statistical results for all events in B1-I, for the initial and the recalibrated models, both with DWF adjustment.

C1. Initial Model with Dry Weather Flow Adjustment per Event at B1-I

Event in B1-I	*n* (#)	Mean (L/s)	Peak (L/s)	RMSE (L/s)	I95 (L/s)	CV_{RMSE} (%)	RSR (%)	NSE (-)	KGE (-)	PBIAS (%)	Slope (-)	y-Interc. (L/s)	r^2 (-)
8	481	130	232	18	36	14%	40%	0.84	0.89	10%	0.97	−8	0.92
13	286	161	312	30	59	18%	41%	0.83	0.82	9%	0.81	17	0.87
9	271	169	312	30	61	18%	43%	0.82	0.78	12%	0.79	15	0.92
21	114	303	335	62	124	21%	211%	**−3.43**	0.52	**18%**	0.90	−26	**0.47**
20	211	145	312	23	45	16%	32%	0.90	0.85	2%	0.82	23	0.91
19	181	168	327	32	63	19%	42%	0.82	0.79	12%	0.79	15	0.91
17	264	110	289	30	60	27%	36%	0.87	0.69	13%	0.70	19	0.98
24	166	157	313	41	82	26%	52%	0.73	0.65	**16%**	0.66	27	0.90
25	181	134	206	25	50	19%	89%	**0.21**	0.40	6%	0.34	81	**0.29**
18	241	91	168	10	20	11%	34%	0.89	0.92	2%	0.89	8	0.89
1	166	121	191	20	40	17%	69%	0.53	0.77	10%	1.04	−18	0.79
23	166	121	179	15	30	12%	52%	0.73	0.87	−4%	0.86	21	0.76
11	121	144	172	18	37	13%	58%	0.66	0.74	1%	0.66	46	0.66
12	121	129	172	12	24	9%	78%	**0.39**	0.55	5%	0.48	60	0.63
3	106	140	196	12	25	9%	48%	0.77	0.89	0%	0.90	13	0.79
16	136	99	189	9	18	9%	29%	0.92	0.91	3%	0.89	8	0.93
22	76	156	268	30	59	19%	78%	**0.39**	0.55	−1%	0.47	85	**0.40**
7	106	105	210	15	29	14%	50%	0.75	0.81	4%	0.76	21	0.77
5	91	120	151	12	24	10%	61%	0.63	0.87	6%	0.89	6	0.78
26	76	137	227	42	85	31%	110%	**−0.22**	0.22	**22%**	0.23	76	0.70
6	91	112	173	17	35	16%	61%	0.62	0.63	−10%	1.29	−22	0.92
15	83	119	167	24	48	20%	91%	**0.17**	0.32	11%	0.28	73	**0.59**
2	76	109	157	9	19	9%	28%	0.92	0.88	−1%	1.08	−8	0.94
10	76	101	149	27	54	27%	138%	**−0.91**	0.37	−13%	0.80	34	**0.31**
4	54	133	208	13	26	10%	34%	0.88	0.92	−5%	1.00	6	0.92
14	15	270	312	42	84	16%	120%	**−0.43**	−0.08	−3%	1.98	−255	0.90

C2. Recalibrated model with dry weather flow adjustment per event at B1-I

	Event in B1-I	*n* (#)	Mean (L/s)	Peak (L/s)	RMSE (L/s)	I95 (L/s)	CV_{RMSE} (%)	RSR (%)	NSE (-)	KGE (-)	PBIAS (%)	Slope (-)	y-Interc. (L/s)	r^2 (-)
V	8	481	130	232	18	35	14%	39%	0.85	0.87	8%	1.05	−16	0.92
Cal	13	286	161	312	24	48	15%	33%	0.89	0.93	4%	0.97	−2	0.90
Cal	9	271	169	312	24	48	14%	34%	0.89	0.91	7%	0.95	−4	0.92
Cal	21	114	303	335	33	66	11%	111%	**−0.23**	0.64	8%	1.04	−37	0.65
V	20	211	145	312	20	41	14%	28%	0.92	0.92	−4%	1.03	2	0.94
Cal	19	181	168	327	22	45	13%	30%	0.91	0.92	6%	0.93	1	0.93
V	17	264	110	289	16	32	15%	19%	0.96	0.88	5%	0.88	8	0.98
Cal	24	166	157	313	30	60	19%	38%	0.86	0.82	11%	0.83	9	0.91
V	25	181	134	206	25	49	18%	87%	**0.24**	0.45	6%	0.38	76	**0.32**
V	18	241	91	168	9	17	10%	30%	0.91	0.93	1%	0.91	7	0.91
Cal	1	166	121	191	20	41	17%	70%	0.51	0.70	9%	1.13	−26	0.79
V	23	166	121	179	14	29	12%	50%	0.75	0.88	−4%	0.90	17	0.79
V	11	121	144	172	19	38	13%	61%	0.63	0.74	1%	0.66	47	0.63
V	12	121	129	172	11	23	9%	74%	**0.45**	0.59	5%	0.52	55	0.68
V	3	106	140	196	14	29	10%	55%	0.69	0.82	−1%	0.98	4	0.76
Cal	16	136	99	189	8	17	8%	26%	0.93	0.94	3%	0.94	3	0.94
V	22	76	156	268	29	57	18%	75%	**0.43**	0.63	−2%	0.55	73	**0.46**
Cal	7	106	105	210	14	29	14%	48%	0.77	0.86	4%	0.83	14	0.78
V	5	91	120	151	12	24	10%	60%	0.64	0.85	6%	0.95	−1	0.79
V	26	76	137	227	42	84	31%	110%	**−0.21**	0.24	**22%**	0.24	74	0.70
V	6	91	112	173	33	65	29%	115%	**−0.33**	0.19	**−17%**	1.66	−55	0.87
V	15	83	119	167	23	45	19%	87%	**0.25**	0.36	11%	0.32	69	0.65
V	2	76	109	157	8	16	8%	25%	0.94	0.86	−1%	1.11	−11	0.96
V	10	76	101	149	27	54	27%	138%	**−0.89**	0.38	−14%	0.87	27	**0.36**
Cal	4	54	133	208	29	58	22%	77%	**0.41**	0.58	**−16%**	1.33	−23	0.92
V	14	15	270	312	40	80	15%	114%	**−0.29**	−0.02	−3%	1.92	−240	0.90

Cal = recalibration event; V = verification event; Peak = measured 2-min peak flow; Event duration = 2.*n* minutes

Better than in C1 / Worsen than in C1; Very Good according to [39]; Good according to [39]; Satisfactory according to [39]; **Not satisfactory according to** [39]

Table A2. Statistical results for all events in B1-M, for the recalibrated model with and without DWF adjustment.

C6. Recalibrated Global Model with Dry Weather Flow Adjustment

	Event in B1-M	*n* (#)	Mean (L/s)	Peak (L/s)	RMSE (L/s)	I95 (L/s)	CV_{RMSE} (%)	RSR (%)	NSE (-)	KGE (-)	PBIAS (%)	Slope (-)	y-Interc. (L/s)	r^2 (-)
V	8	481	171	299	24	49	14%	45%	0.79	0.74	4%	1.19	−38.7	0.90
Cal	9	271	233	532	35	71	15%	34%	0.88	0.87	10%	0.90	−1.1	0.94
Cal	13	286	212	692	28	56	13%	29%	0.91	0.92	3%	1.03	−12.0	0.93
Cal	21	114	384	887	61	122	16%	51%	0.74	0.85	2%	0.81	66.0	0.75
V	20	211	184	550	27	55	15%	26%	0.93	0.90	−7%	1.04	6.1	0.95
Cal	19	181	211	1167	45	91	22%	30%	0.91	0.91	−4%	1.03	1.6	0.93
V	14	129	281	1053	47	95	17%	29%	0.92	0.94	1%	1.01	−5.3	0.92
V	17	264	133	376	12	25	9%	13%	0.98	0.96	−1%	1.03	−2.3	0.99
V	18	241	123	223	10	20	8%	27%	0.93	0.96	−1%	0.95	6.5	0.93
V	25	181	157	226	21	43	14%	69%	0.53	0.77	−4%	0.73	48.9	0.61
Cal	24	166	169	367	23	47	14%	29%	0.92	0.89	0%	1.07	−11.9	0.94
Cal	1	166	150	259	24	48	16%	60%	0.64	0.66	0%	1.20	−29.9	0.82
V	23	166	134	198	25	51	19%	70%	0.51	0.82	**−17%**	1.01	21.3	0.92
V	11	121	182	221	22	44	12%	57%	0.67	0.82	−3%	0.77	46.9	0.70
V	3	106	198	260	24	48	12%	79%	**0.37**	0.63	−1%	1.04	−6.3	0.64
V	12	121	167	206	12	23	7%	72%	**0.48**	0.69	3%	0.61	59.9	**0.59**
Cal	7	106	155	259	21	43	14%	53%	0.72	0.82	7%	1.05	−19.6	0.85
Cal	16	136	117	199	14	28	12%	41%	0.83	0.72	−3%	1.24	−23.9	0.94
V	6	91	172	282	34	67	19%	71%	**0.49**	0.44	−9%	1.51	−71.2	0.95
V	5	91	166	207	16	32	10%	59%	0.65	0.76	6%	1.15	−34.8	0.88
V	22	76	177	343	32	64	18%	66%	0.56	0.80	−8%	0.92	28.0	0.71
V	26	76	160	233	37	74	23%	106%	**−0.12**	0.41	**16%**	0.36	76.1	**0.48**
V	15	83	136	181	21	42	15%	76%	**0.43**	0.45	−1%	0.37	87.5	**0.44**
V	10	76	138	214	21	41	15%	66%	0.56	0.80	−7%	0.95	16.7	0.73
V	2	76	138	208	14	29	10%	30%	0.91	0.91	−5%	0.90	20.2	0.93
Cal	4	54	195	310	40	79	20%	74%	**0.46**	0.46	−11%	1.48	−72.9	0.94

C7. Recalibrated global model without dry weather flow adjustment

Event in B1-M	*n* (#)	Mean (L/s)	Peak (L/s)	RMSE (L/s)	I95 (L/s)	CV_{RMSE} (%)	RSR (%)	NSE (-)	KGE (-)	PBIAS (%)	Slope (-)	y-Interc. (L/s)	r^2 (-)
8	481	171	299	23	46	13%	43%	0.81	0.75	−2%	1.18	−27.4	0.90
9	271	233	532	38	76	16%	37%	0.86	0.86	12%	0.91	−7.7	0.94
13	286	212	692	39	77	18%	41%	0.84	0.83	12%	1.07	−40.5	0.93
21	114	384	887	60	120	16%	50%	0.75	0.85	0%	0.81	74.9	0.75
20	211	184	550	35	69	19%	33%	0.89	0.85	−14%	1.04	18.3	0.95
19	181	211	1167	51	103	24%	34%	0.89	0.86	−13%	1.02	22.4	0.93
14	129	281	1053	53	106	19%	32%	0.90	0.89	7%	1.03	−28.5	0.92
17	264	133	376	27	54	20%	28%	0.92	0.82	**−18%**	0.98	26.6	0.98
18	241	123	223	10	20	8%	28%	0.92	0.95	3%	0.95	2.5	0.93
25	181	157	226	30	59	19%	95%	**0.10**	0.74	−14%	0.74	62.3	0.61
24	166	169	367	25	49	15%	30%	0.91	0.86	−1%	1.10	−15.1	0.94
1	166	150	259	24	48	16%	60%	0.64	0.66	0%	1.20	−29.9	0.82
23	166	134	198	34	68	25%	94%	**0.12**	0.75	**−24%**	1.03	28.3	0.93
11	121	182	221	22	44	12%	57%	0.67	0.82	−3%	0.78	45.7	0.70
3	106	198	260	27	54	13%	88%	**0.22**	0.63	−6%	1.05	3.1	0.64
12	121	167	206	22	45	13%	138%	**−0.91**	0.66	12%	0.60	46.1	0.61
7	106	155	259	18	37	12%	46%	0.79	0.84	1%	1.05	−8.5	0.84
16	136	117	199	46	91	39%	132%	**−0.75**	0.53	**−37%**	1.24	15.4	0.94
6	91	172	282	43	85	25%	91%	**0.17**	0.48	**−19%**	1.43	−41.5	0.94
5	91	166	207	12	25	7%	46%	0.79	0.76	0%	1.15	−24.7	0.87
22	76	177	343	31	62	18%	64%	0.59	0.81	8%	0.93	−1.9	0.73
26	76	160	233	37	74	23%	106%	**−0.11**	0.40	**16%**	0.35	77.8	**0.46**
15	83	136	181	29	59	22%	106%	**−0.13**	0.45	**−16%**	0.40	103.3	**0.46**
10	76	138	214	19	37	13%	60%	0.64	0.81	−4%	0.95	11.5	0.73
2	76	138	208	21	42	15%	44%	0.80	0.86	−12%	0.91	28.5	0.93
4	54	195	310	47	93	24%	87%	**0.25**	0.48	**−17%**	1.43	−50.7	0.92

Cal = recalibration event; V = verification event

Peak = measured 2-min peak flow

Event duration = 2.*n* minutes

Better than in C6 | Worse than in C6

Very Good according to [39]

Good according to [39]

Satisfactory according to [39]

Not satisfactory according to [39]

References

1. Gallo, E.M.; Bell, C.D.; Panos, C.L.; Smith, S.M.; Hogue, T.S. Investigating Tradeoffs of Green to Grey Stormwater Infrastructure Using a Planning-Level Decision Support Tool. *Water* **2020**, *12*, 2005. [CrossRef]
2. Hou, X.; Qin, L.; Xue, X.; Xu, X.S.; Yang, Y.; Liu, X.; Li, M. A city-scale fully controlled system for stormwater management: Consideration of flooding, non-point source pollution and sewer overflow pollution. *J. Hydrol.* **2021**, *603 Pt D*, 127155. [CrossRef]
3. David, L.M.; Carvalho, R.F.D. Designing for People's Safety on Flooded Streets: Uncertainties and the Influence of the Cross-Section Shape, Roughness and Slopes on Hazard Criteria. *Water* **2021**, *13*, 2119. [CrossRef]
4. Refsgaard, J.C. Parameterisation, calibration and validation of distributed hydrological models. *J. Hydrol.* **1996**, *198*, 69–97. [CrossRef]
5. David, L.M.; Matos, J.S. Combined sewer overflow emissions to bathing waters in Portugal. How to reduce in densely urbanised areas? *Water Sci. Technol.* **2005**, *52*, 183–190. [CrossRef] [PubMed]
6. Obropta, C.; Kardos, J. Review of Urban Stormwater Quality Models: Deterministic, Stochastic, and Hybrid Approaches. *JAWRA J. Am. Water Resour. Assoc.* **2007**, *43*, 1508–1523. [CrossRef]
7. Melo, P.A.; Alvarenga, L.A.; Tomasella, J.; Santos, A.C.N.; Mello, C.R.; Colombo, A. On the performance of conceptual and physically based modelling approach to simulate a headwater catchment in Brazil. *J. S. Am. Earth Sci.* **2022**, *114*, 103683, ISSN 0895-9811. [CrossRef]
8. Revitt, D.M.; Ellis, J.B.; Lundy, L. Assessing the impact of swales on receiving water quality. *Urban Water J.* **2017**, *14*, 839–845. [CrossRef]
9. Gorgoglione, A.; Bombardelli, F.A.; Pitton, B.J.L.; Oki, L.R.; Haver, D.L.; Young, T.M. Role of Sediments in Insecticide Runoff from Urban Surfaces: Analysis and Modeling. *Int. J. Environ. Res. Public Health* **2018**, *15*, 1464. [CrossRef]
10. Vonach, T.; Kleidorfer, M.; Rauch, W.; Tscheikner-Gratl, F. An Insight to the Cornucopia of Possibilities in Calibration Data Collection. *Water Resour. Manag.* **2019**, *33*, 1629–1645. [CrossRef]
11. Rodrigues, M.; Guerreiro, M.; David, L.M.; Oliveira, A.; Menaia, J.; Jacob, J. Role of environmental forcings on fecal contamination behavior in a small, intermittent coastal stream: An integrated modelling approach. *J. Environ. Eng.* **2016**, *142*, 05016001. [CrossRef]
12. Mizukami, N.; Rakovec, O.; Newman, A.J.; Clark, M.P.; Wood, A.W.; Gupta, H.V.; Kumar, R. On the choice of calibration metrics for "high-flow" estimation using hydrologic models. *Hydrol. Earth Syst. Sci.* **2019**, *23*, 2601–2614. [CrossRef]
13. Yu, Y.; Kojima, K.; An, K.; Furumai, H. Cluster analysis for characterization of rainfalls and CSO behaviours in an urban drainage area of Tokyo. *Water Sci. Technol.* **2013**, *68*, 544–551. [CrossRef]
14. Montserrat, A.; Hofer, T.; Poch, M.; Muschalla, D.; Corominas, L. Using the duration of combined sewer overflow events for the calibration of sewer hydrodynamic models. *Urban Water J.* **2017**, *14*, 782–788. [CrossRef]
15. Carvalho, R.F.; Lopes, P.; Leandro, J.; David, L.M. Numerical Research of Flows into Gullies with Different Outlet Locations. *Water* **2019**, *11*, 794. [CrossRef]
16. Troutman, S.C.; Schambach, N.; Love, N.G.; Kerkez, B. An automated toolchain for the data-driven and dynamical modeling of combined sewer systems. *Water Res.* **2017**, *126*, 88–100. [CrossRef]
17. An, W.W.; Gianvito, J.M. Kiski Valley WPCA Combined Sewer System Long Term Model Study. *J. Water Manag. Modeling* **2011**, R241-17. [CrossRef]
18. Peche, A.; Graf, T.; Fuchs, L.; Neuweiler, I. Physically based modeling of stormwater pipe leakage in an urban catchment. *J. Hydrol.* **2019**, *573*, 778–793, ISSN 0022-1694. [CrossRef]
19. Nasrin, T.; Tran, H.D.; Muttil, N. Modelling Impact of Extreme Rainfall on Sanitary Sewer System by Predicting Rainfall Derived Infiltration/Inflow. In Proceedings of the 20th International Congress on Modelling and Simulation (MODSIM2013), Adelaide, Australia, 1–6 December 2013.
20. Nasrin, T.; Sharma, A.K.; Muttil, N. Impact of Short Duration Intense Rainfall Events on Sanitary Sewer Network Performance. *Water* **2017**, *9*, 225. [CrossRef]
21. Wang, M.; Zhang, M.; Shi, H.; Huang, X.; Liu, Y. Uncertainty analysis of a pollutant-hydrograph model in assessing inflow and infiltration of sanitary sewer systems. *J. Hydrol.* **2019**, *574*, 64–74. [CrossRef]
22. Ferreira, P.M.D.L.; Paz, A.R.D.; Bravo, J.M. Objective functions used as performance metrics for hydrological models: State-of-the-art and critical analysis. *RBRH Braz. J. Water Resour.* **2020**, *25*, e42. [CrossRef]
23. Beven, K.J.; Binley, A.M. The future of distributed models: Model calibration and uncertainty prediction. *Hydrol. Processes* **1992**, *6*, 279–298. [CrossRef]
24. Liu, Y.; Gupta, H.V. Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. *Water Resour. Res.* **2007**, *43*, W07401. [CrossRef]
25. Moges, E.; Demissie, Y.; Larsen, L.; Yassin, F. Review: Sources of Hydrological Model Uncertainties and Advances in Their Analysis. *Water* **2021**, *13*, 28. [CrossRef]
26. ASME PTC Committee. Standard for Verification and Validation in Computational Fluid Dynamics and Heat Transfer. *Am. Soc. Mech. Eng.* **2009**, *20*, 1–42.
27. Gorgoglione, A.; Bombardelli, F.A.; Pitton, B.J.; Oki, L.R.; Haver, D.L.; Young, T.M. Uncertainty in the parameterization of sediment build-up and wash-off processes in the simulation of sediment transport in urban areas. *Environ. Model. Softw.* **2019**, *111*, 170–181. [CrossRef]

28. Oladyshkin, S.; Nowak, W. Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion. *Reliab. Eng. Syst. Saf.* **2012**, *106*, 179–190. [CrossRef]
29. Ghaith, M.; Siam, A.; Li, Z.; El-Dakhakhni, W. Hybrid Hydrological Data-Driven Approach for Daily Streamflow Forecasting. *J. Hydrol. Eng.* **2019**, *25*, 04019063. [CrossRef]
30. Zhou, P.; Li, C.; Li, Z.; Cai, Y. Assessing uncertainty propagation in hybrid models for daily streamflow simulation based on arbitrary polynomial chaos expansion. *Adv. Water Resour.* **2022**, *160*, 104110. [CrossRef]
31. Nash, J.E.; Sutcliffe, J.V. River flow forecasting through conceptual models: Part 1. A discussion of principles. *J. Hydrol.* **1970**, *10*, 282–290. [CrossRef]
32. Willmot, C.J. On the validation of models. *Phys. Geogr.* **1981**, *2*, 184–194. [CrossRef]
33. Legates, D.R.; McCabe, G.J. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resour. Res.* **1999**, *35*, 233–241. [CrossRef]
34. Krause, P.; Boyle, D.P.; Base, F. Comparison of different efficiency criteria for hydrological model assessment. *Adv. Geosci.* **2005**, *5*, 89–97. [CrossRef]
35. Harmel, R.D.; Smith, P.K. Consideration of measurement uncertainty in the evaluation of goodness-of-fit in hydrologic and water quality modelling. *J. Hydrol.* **2007**, *337*, 326–336. [CrossRef]
36. Moriasi, D.N.; Arnold, J.G.; Van Liew, M.W.; Bingner, R.L.; Harmel, R.D.; Veith, T.L. Model evaluations guidelines for systematic quantification of accuracy in watershed simulations. *Trans. ASABE* **2007**, *50*, 885–900. [CrossRef]
37. Gupta, H.V.; Kling, H.; Yilmaz, K.K.; Martinez, G.F. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol.* **2009**, *377*, 80–91. [CrossRef]
38. Harmel, R.D.; Smith, P.K.; Migliaccio, K.W. Modifying Goodness-of-Fit Indicators to Incorporate Both Measurement and Model Uncertainty in Model Calibration and Validation. *Trans. ASABE* **2010**, *53*, 55–63. [CrossRef]
39. Moriasi, D.N.; Gitau, M.W.; Pai, N.; Daggupati, P. Hydrologic and Water Quality Models: Performance Measures and Evaluation Criteria. *Trans. ASABE* **2015**, *58*, 1763–1785. [CrossRef]
40. Chen, H.; Luo, Y.; Potter, C.; Moran, P.J.; Grieneisen, M.L.; Zhang, M. Modeling pesticide diuron loading from the San Joaquin watershed into the Sacramento-San Joaquin Delta using SWAT. *Water Res.* **2017**, *121*, 374–385. [CrossRef]
41. Rodríguez, R.; Pastorini, M.; Etcheverry, L.; Chreties, C.; Fossati, M.; Castro, A.; Gorgoglione, A. Water-Quality Data Imputation with a High Percentage of Missing Values: A Machine Learning Approach. *Sustainability* **2021**, *13*, 6318. [CrossRef]
42. Kastridis, A.; Theodosiou, G.; Fotiadis, G. Investigation of Flood Management and Mitigation Measures in Ungauged NATURA Protected Watersheds. *Hydrology* **2021**, *8*, 170. [CrossRef]
43. Segura-Beltrán, F.; Sanchis-Ibor, C.; Morales-Hernández, M.; González-Sanchis, M.; Bussi, G.; Ortiz, E. Using post-flood surveys and geomorphologic mapping to evaluate hydrological and hydraulic models: The flash flood of the Girona River (Spain) in 2007. *J. Hydrol.* **2016**, *541*, 310–329. [CrossRef]
44. Jackson, E.K.; Roberts, W.; Nelsen, B.; Williams, G.P.; Nelson, E.J.; Ames, D.P. Introductory overview: Error metrics for hydrologic modelling—A review of common practices and an open source library to facilitate use and adoption. *Environ. Model. Softw.* **2019**, *119*, 32–48, ISSN 1364-8152. [CrossRef]
45. Knoben, W.J.; Freer, J.E.; Woods, R.A. Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrol. Earth Syst. Sci.* **2019**, *23*, 4323–4331. [CrossRef]
46. Randall, M.; Perera, N.; Gupta, N.; Ahmad, M. Development and Calibration of a Dual Drainage Model for the Cooksville Creek Watershed, Canada. *J. Water Manag. Modeling* **2017**, *25*, C419. [CrossRef]
47. Rujner, H.; Leonhardt, G.; Marsalek, J.; Viklander, M. High-resolution modelling of the grass swale response to runoff inflows with Mike SHE. *J. Hydrol.* **2018**, *562*, 411–422, ISSN 0022-1694. [CrossRef]
48. Hossain, S.; Hewa, G.A.; Wella-Hewage, S. A Comparison of Continuous and Event-Based Rainfall–Runoff (RR) Modelling Using EPA-SWMM. *Water* **2019**, *11*, 611. [CrossRef]
49. González-Álvarez, Á.; Molina-Pérez, J.; Meza-Zúñiga, B.; Viloria-Marimón, O.M.; Tesfagiorgis, K.; Mouthón-Bello, J.A. Assessing the Performance of Different Time of Concentration Equations in Urban Ungauged Watersheds: Case Study of Cartagena de Indias, Colombia. *Hydrology* **2020**, *7*, 47. [CrossRef]
50. Rosa, D.W.B.; Nascimento, N.O.; Moura, P.M.; Macedo, G.D. Assessment of the hydrological response of an urban watershed to rainfall-runoff events in different land use scenarios—Belo Horizonte, MG, Brazil. *Water Sci. Technol.* **2020**, *81*, 679–693. [CrossRef]
51. Iffland, R.; Förster, K.; Westerholt, D.; Pesci, M.H.; Lösken, G. Robust Vegetation Parameterization for Green Roofs in the EPA Stormwater Management Model (SWMM). *Hydrology* **2021**, *8*, 12. [CrossRef]
52. Radinja, M.; Škerjanec, M.; Džeroski, S.; Todorovski, L.; Atanasova, N. Design and Simulation of Stormwater Control Measures Using Automated Modeling. *Water* **2021**, *13*, 2268. [CrossRef]
53. Rohith, A.N.; Gitau, M.W.; Chaubey, I.; Sudheer, K.P. A multistate first-order Markov model for modeling time distribution of extreme rainfall events. *Stoch. Environ. Res. Risk Assess.* **2021**, *35*, 1205–1221. [CrossRef]
54. Saadi, M.; Oudin, L.; Ribstein, P. Physically consistent conceptual rainfall–runoff model for urbanized catchments. *J. Hydrol.* **2021**, *599*, 126394, ISSN 0022-1694. [CrossRef]
55. Wu, W.; Lu, L.; Huang, X.; Shangguan, H.; Wei, Z. An automatic calibration framework based on the InfoWorks ICM model: The effect of multiple objectives during multiple water pollutant modeling. *Environ. Sci. Pollut. Res.* **2021**, *28*, 31814–31830. [CrossRef]

56. Chai, T.; Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* **2014**, *7*, 1247–1250. [CrossRef]
57. Sapountzis, M.; Kastridis, A.; Kazamias, A.P.; Karagiannidis, A.; Nikopoulos, P.; Lagouvardos, K. Utilization and uncertainties of satellite precipitation data in flash flood hydrological analysis in ungauged watersheds. *Glob. NEST J.* **2021**, *23*, 388–399.
58. Von Schiller, D.; Datry, T.; Corti, R.; Foulquier, A.; Tockner, K.; Marcé, R.; Zoppini, A. Sediment respiration pulses in Intermittent Rivers and ephemeral streams. *Glob. Biogeochem. Cycles* **2019**, *33*, 1251–1263. [CrossRef]
59. Rossman, L.A. *Storm Water Management Model User's Manual Version 5.1*; U.S. Environmental Protection Agency: Washington, DC, USA, 2015.
60. David, L.M.; Matos, R.S. Wet weather water quality modelling of a Portuguese urban catchment: Difficulties and benefits. *Water Sci. Technol.* **2002**, *45*, 131–140. [CrossRef]
61. David, L.M. Water quality in Portuguese pseudo-separate and combined systems: A conceptual modelling approach for data comparison. In *Global Solutions for Urban Drainage*; Strecker, E.W., Huber, W.C., Eds.; American Society of Civil Engineers: Reston, VA, USA, 2002; pp. 1–11. ISBN 0-7844-0644-8. [CrossRef]
62. David, L.M.; Matos, J.S. Wet-weather urban discharges: Implications from adopting the revised European Directive concerning the quality of bathing water. *Water Sci. Technol.* **2005**, *52*, 9–17. [CrossRef]
63. Cambez, M.J.; Pinho, J.; David, L.M. Using SWMM 5 in the continuous modelling of stormwater hydraulics and quality. In Proceedings of the 11th International Conference on Urban Drainage, Edinburgh, UK, 31 August–5 September 2008; ISBN 9781899796212.