Sensitivity to experimental data of pollutant site mean concentration in stormwater runoff

M. Mourad*, J.-L. Bertrand-Krajewski* and G. Chebbo**

* URGC Hydrologie Urbaine, INSA de Lyon, 34 avenue des Arts, 69621 Villeurbanne cedex, France (E-mail: mohammad.mourad@insa-lyon.fr; jean-luc.bertrand-krajewski@insa-lyon.fr)
** Faculté de Génie III, Université Libanaise, Beyrouth, Liban; Cereve – ENPC, France (E-mail: chebbo@cereve.enpc.fr)

Abstract Urban wet weather discharges are known to be a great source of pollutants for receiving waters, which protection requires the estimation of long-term discharged pollutant loads. Pollutant loads can be estimated by multiplying a site mean concentration (SMC) by the total runoff volume during a given period of time. The estimation of the SMC value as a weighted mean value with event runoff volumes as weights is affected by uncertainties due to the variability of event mean concentrations and to the number of events used. This study carried out on 13 catchments gives orders of magnitude of these uncertainties and shows the limitations of usual practices using few measured events. The results obtained show that it is not possible to propose a standard minimal number of events to be measured on any catchment in order to evaluate the SMC value with a given uncertainty.

Keywords Concentration; sample size; uncertainty; variability; urban wet weather discharges

Introduction

Urban Wet Weather Discharges (UWWD) from combined and separate sewer systems are considered as an important source of pollution for receiving waters. In almost all developed countries, the environmental legislation requires or recommends the estimation of annual discharged loads in order to evaluate long-term impacts of UWWD. Evaluation methods are numerous in the literature. They are based on different approaches with different levels of complexity. After three decades of research, sophisticated and complex models are not necessarily able to provide much better results at large catchment scale than simpler ones regarding the greater effort and expense they need. It is known that the more the model is complex, the more data and effort is needed for its application. From an operational point of view, many constraints lead practitioners to use simplistic approaches. One of the catchment-based approaches consists to use a Site-specific Mean Concentration (SMC). The long-term discharged load is then calculated by multiplying the total runoff volume by the SMC value. As an example, this approach is used in 43% of the studies carried out in France by public and private consulting companies (Gromaire et al., 2002).

The SMC value should be representative of the distribution of the Event Mean Concentrations (EMCs) to be encountered during a given period of time. It is estimated as a central value of the distribution. It is also very important to evaluate the uncertainty in the estimated SMC value from the available data. This uncertainty is necessary to have an overview of the range of possible values of the estimated load.

As in practice, the available data are usually limited due to time and cost constraints, the problem can be formulated as follows. How many events are required to estimate the SMC value with a given level of uncertainty? Reciprocally, what is the uncertainty in the SMC value estimated from a given set of measured events? Regarding the variability of the
SMC value, how the SMC value may vary with different sets of measured events having the same size? From the literature, it also appears that SMC values are not always estimated in the same way: this point will be first discussed in the next paragraph, before investigating the above questions.

**SMC estimation**

The most commonly used approach to estimate the SMC value is the simple arithmetic mean. This value is well discussed in the literature and can be easily calculated (Eq. (1)) as well as its uncertainty (Eq. (2)).

$$\text{SMC} = \frac{\sum_{i=1}^{n} \text{EMC}_i}{n}$$  \hspace{1cm} (1)

$$\sigma_{\text{SMC}} = \frac{\sigma_{\text{EMC}}}{\sqrt{n}}$$  \hspace{1cm} (2)

where $n$ is the number of events used to estimate the SMC value and $\sigma_{\text{EMC}}$ is the standard deviation of the EMC values. In Eq. (2), the uncertainties in each EMC value are neglected. The 95% confidence interval bounds are located at $\pm 1.96 \sigma_{\text{SMC}}$ around the SMC value.

One can then compute the required number of measured EMC values to estimate the SMC value with a given acceptable difference between the estimated and the true mean values.

The simple arithmetic mean can be a good estimate if the EMC values are normally distributed. But it is well known that EMC values usually follow a lognormal distribution (e.g. Athayde et al., 1983; Brizio, 1988; Driscoll et al., 1990; Rossi, 1998). Hence, it is more logic to use the lognormal mean. The lognormal mean and its standard deviation are given respectively by Eq. (3) and Eq. (4).

$$\text{SMC} = \exp \left( Mu + \frac{\sigma^2}{2} \right)$$  \hspace{1cm} (3)

$$\sigma_{\text{EMC}} = \exp (Mu + \sigma^2) - \exp \left( Mu + \frac{\sigma^2}{2} \right)$$  \hspace{1cm} (4)

where $Mu$ and $\sigma$ are respectively the mean and the standard deviation of the logarithmic transform of the EMC values used to calculate the SMC value.

The confidence interval of the lognormal mean is not symmetrical and less easy to calculate than for the normal distribution. Various methods like the H-statistic based procedure, the Jackknife procedure, the Bootstrap procedure, the central limit theorem and the Chebychev theorem allow computing the confidence interval around the mean. The first method is not recommended when $n < 30$, while the Bootstrap and Jackknife procedures are recommended when the distribution turns out to be neither normal nor lognormal (Singh et al., 1997). The computation method of the number of samples required to estimate the lognormal mean with a given acceptable difference between the estimated and the true mean values can be found in Perez and Lefante (1997).

As the simple mean and the lognormal mean values are very sensitive to extreme values, their estimates, using limited datasets, can be significantly biased. Another central value of the distribution is the median value, which is less sensitive to extreme scores than the simple and lognormal mean values. It is then a better estimate than the means for highly skewed distributions. The standard deviation of the median value for very large samples and normal distributions is about 25% larger than for the simple mean value. This may lead to strong biases for extremely non-normal distributions and limited size sample sets. Moreover, for non-normal distributions, the standard deviation of the median value is difficult to compute.
Since the main objective of our study is to estimate long-term pollutant loads by multiplying the SMC value by runoff volumes, it is more appropriate to take runoff volume into account. Hence, the weighted mean, with event volumes as weights (Eq. (5)), appears as a more pertinent estimate of the SMC value. This approach is less sensitive to high concentrations associated to low runoff volumes.

\[
SMC = \frac{\sum_{i=1}^{n} EMC_i V_i}{\sum_{i=1}^{n} V_i}
\]

There is no universal method to compute the standard deviation of the weighted mean value. Various expressions and methods are given in the literature (Cochran, 1977; Galloway et al., 1984). A comparison of these methods to bootstrapping (Efron and Tibshirani, 1986) for ions concentration in precipitations is presented in Gatz and Smith (1995a). Gatz and Smith (1995b) compared also two methods to calculate confidence intervals for the weighted mean value. The first one assumes a normal distribution of the weighted mean value and the other one is the bias-corrected percentile method applied to bootstrapped simulation of the distribution of the weighted mean value.

Two comparisons of the four above calculation methods have been carried out on three catchments for three pollutants (BOD, COD and TSS). In the first comparison, the SMC value has been calculated starting with \( n = 1 \) event and updated with the next measured event by respecting the chronological order. Then the total load for all available events is calculated for each computed SMC value by multiplying the SMC value by the total runoff volume of all events. The results of the four methods are compared to the measured total load. Figure 1 shows the case of one catchment (Les Ulis Nord) and one pollutant (TSS). For example, when the first \( n = 40 \) events are used to estimate the SMC value, the estimated total load for 75 events range from 30,000 to 50,000 kg TSS, depending on the approach used.

**Figure 1** Comparison of the four methods to estimate the SMC value (75 events)
The normal and lognormal mean values give in general very close results but the disparities with the median and the weighted mean values are shown to be significant. It appears from the results that, for a given catchment, the four methods lead to similar conclusions for the three pollutants. But results differ from one catchment to another and depend on the data used. As expected, the median value is less sensitive and less significant jumps or falls in its curve are observed. Even by using all measured events, the result obtained with the simple mean value has a very significant bias compared to the measured total load.

In the second comparison, subsets of events of different sizes are resampled from the available data set (details in the next section). For each size a density of probability of the SMC value is obtained for each of the four calculation methods. A density of probability of the estimated total load of all events is derived by multipliying the SMC values by the total volume of all events. From these densities of probability we can obtain the probability to have the estimated and measured total load of all events equal (with a $\pm 20\%$ tolerance). The results are shown in Figure 2. The median and the weighted mean showed convergence of the estimated total load to the measured total load as the number of events used to calculate the SMC value increased. For normal and lognormal mean an obvious bias is introduced.

According to these results and to our objectives, the weighted mean value has been selected for further calculations and results presented in the next paragraphs to investigate the variability of the weighted mean SMC value estimated from limited datasets.

**Method**

**Virtual measurement campaigns**

In order to evaluate the effect of using limited datasets to estimate the SMC value and to take into account the variability of the EMC values, subsets of measured events have been withdrawn from all available measured events. Assuming $N$ is the number of available data on a given catchment for a given pollutant, a large number $N_S$ of subsets, each one representing a virtual measurement campaign and having a size equal to $n$, are formed. Different

![Figure 2](https://iwaponline.com/wst/article-pdf/51/2/155/434801/155.pdf)
datasets sizes are used, from \( n = 1 \) to \( N - 1 \). Each subset is sampled without replacement, which prevents the use of an event twice or more in the same subset, as, in reality, the same event can not be measured more than once. By increasing \( n \), the number of possible combinations increases rapidly and starts to decrease when \( n \) becomes greater than \( N/2 \). When the number of possible combinations is greater than 1,000, the number of withdrawn subsets is limited to 1,000. For each subset, the SMC value is estimated. Consequently, for each value of \( n \), the distribution of the SMC value is obtained, which represents the variability that can be encountered for a given catchment based on the available data.

**Data**

The data series used in this study is taken from the French database QASTOR (Saget and Chebbo, 1996) and from a study on Le Marais catchment in Paris (Gromaire, 1998). Six catchments with separate sewer systems and seven catchments with combined sewer systems have been selected. The number of measured events ranged between 16 and 121 events per catchment. The study dealt with three pollutants: BOD, COD and TSS. Available data are shown in Table 1.

**Results**

As mentioned in the introduction, the main objective is to assess uncertainty in SMC values. Hence, graphical and analytical tests have been performed to fit a probability function to the SMC distributions. Two functions have been compared: the normal and lognormal probability functions. It was found that in most cases, for \( n < 10–15 \), the SMC distributions are more likely lognormal and tend to be symmetrical when \( n \) increases. As an example, Table 2 shows the \( \chi^2 \) tests for Le Marais catchment.

Assuming that the SMC distributions are lognormal, 95% confidence intervals have been calculated. In order to have a global view of the results, the width of the confidence interval (i.e. width = 3.92 times the standard deviation) is shown on Figure 3 and Figure 4 for COD,

<table>
<thead>
<tr>
<th>Number of events</th>
<th>( \chi^2 ) normal distribution</th>
<th>( \chi^2 ) lognormal distribution</th>
<th>( \chi^2 ) theoretical (0.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n )</td>
<td>SS COD</td>
<td>SS COD</td>
<td>SS COD</td>
</tr>
<tr>
<td>1</td>
<td>1 1 19.98 6.57</td>
<td>2.63 1.69 11.07</td>
<td>5.99</td>
</tr>
<tr>
<td>10</td>
<td>10 10 106.42 10.18</td>
<td>16.57 1.06 11.07</td>
<td>7.81</td>
</tr>
<tr>
<td>20</td>
<td>20 20 92.79 9.13</td>
<td>26.51 0.97 15.60</td>
<td>7.81</td>
</tr>
<tr>
<td>40</td>
<td>40 40 28.60 13.46</td>
<td>11.83 7.93 16.91</td>
<td>7.81</td>
</tr>
<tr>
<td>67</td>
<td>64 64 21.85 1.81</td>
<td>1.89 1.79 12.59</td>
<td>5.99</td>
</tr>
</tbody>
</table>
respectively for separate and combined sewer systems. As expected, catchments with higher variability of EMC values have larger confidence intervals. Catchments with similar characteristics behave similarly (see e.g. the case of the five La Briche catchments which are close from each others).

More significant disparities can be observed for the separate sewer systems, due to: i) the variability of EMC values between sites; ii) the effect of the number of events in the original dataset that includes too the variability of the EMC values on the same site. It was also found from a previous data analysis that in most cases the mean value and the standard deviation are well correlated for EMC values: higher variability of the EMC values can be expected for more polluted catchments.

It is obvious that the results shown in Figure 3 and Figure 4 depend strongly on the initial available dataset, the effect of which was investigated for TSS for the Maurepas catchment that has the larger number of measured events and the lowest variability in EMC values. The question to be answered is: how the width of the confidence intervals shown in Figure 3 and

![Figure 3](image1.png)

**Figure 3** Width of the 95% confidence interval of the SMC value for separate sewer systems

![Figure 4](image2.png)

**Figure 4** Width of the 95% confidence interval of the SMC value for combined sewer systems
Figure 4 can change when having less data? Twelve datasets with 30 events each were withdrawn among the 121 available measured events. Each dataset was considered as independent and processed with the method described above.

The results are shown in Figure 5 with \( n \) ranging from 3 to 27. The dark line represents the confidence interval width calculated with the 121 events and the lighter ones represent the confidence interval widths calculated using the 12 datasets. Depending on the datasets used, i) the confidence interval width systematically decreases with increasing \( n \), which confirms that SMC values based on few events are affected by very significant uncertainties: from \( n = 5 \) to \( n = 25 \) among \( N = 30 \), the confidence interval width is approximately reduced by 50%; ii) the SMC values strongly vary from one dataset to another one: for \( n = 10 \) among \( N = 30 \), the confidence interval width ranges from approximately 180 to 470 mg TSS/L.

Conclusion
This paper showed that the four methods (simple arithmetic mean value, lognormal mean value, median value and weighted mean value) used to estimate the SMC value and its uncertainty give dissimilar results. If the weighted mean value is preferable for long term evaluation when large datasets are available, this preference is not obvious when few events are used to estimate the SMC value. When dealing with very limited size datasets, it is very difficult to evaluate the variability of EMC values and its consequence on the estimation of the SMC value. The estimated SMC value can be biased. The uncertainty in the SMC value may vary very significantly, as well as the annual pollutant loads estimated from the SMC value. Based on the study carried out on 13 catchments with different sizes of datasets, orders of magnitude of the uncertainties in SMC values have been evaluated, which shows clearly the limitations of usual practices using few measured events. The results obtained show that, due to the variability of the EMC values from catchment to catchment, it is not possible to propose a standard minimal number of events to be measured on any catchment in order to evaluate the SMC value with a given uncertainty.

Acknowledgements
The authors gratefully acknowledge the financial support of the RGCU “Réseau Génie Civil et Urbain”.

Figure 5 Effect of the initial dataset
References


