Stormwater pollutant loads modelling: epistemological aspects and case studies on the influence of field data sets on calibration and verification

Jean-Luc Bertrand-Krajewski
Laboratoire URGC, INSA de Lyon, 34 avenue des Arts, 69621 Villeurbanne cedex, France
(Email: jean-luc.bertrand-krajewski@insa-lyon.fr)

Abstract In urban drainage, stormwater quality models have been used by researchers and practitioners for more than 15 years. Most of them were initially developed for research purposes, and have been later on implemented in commercial software packages devoted to operational needs. This paper presents some epistemological problems and difficulties with practical consequences in the application of stormwater quality models, such as simplified representation of reality, scaling-up, over-parameterisation, transition from calibration to verification and prediction, etc. Two case studies (one to estimate pollutant loads at the outlet of a catchment, one to design a detention tank to reach a given pollutant interception efficiency), with simple and detailed stormwater quality models, illustrate some of the above problems. It is hard to find, if not impossible, an “optimum” or “best” unique set of parameters values. Model calibration and verification appear to dramatically depend on the data sets used for their calibration and verification. Compared to current practice, collecting more and reliable data is absolutely necessary.

Keywords Calibration; epistemology; field data; modelling; sensitivity analysis; separate and combined sewers; stormwater; verification

Introduction
In urban hydrology, models have been widely used by researchers and practitioners (consulting companies, sewer operators, water authorities) since the middle of the 1980s for hydrological and hydraulic aspects (rainfall-runoff processes, flow propagation in both combined and separate sewer systems, including hydraulic ancillaries like, e.g., overflows, detention tanks, etc.), and since the second half of the 1990s for pollutant loads and concentrations. If most of the models were initially developed for research purposes, they have been later on implemented in commercial software packages devoted to operational needs. This implementation has often been carried out by promoting the models’ features and capabilities, with examples of application aiming to convince users that these models had great potential to solve practical problems. Unfortunately, limits and conditions of application have not been sufficiently emphasised, especially regarding the critical questions related to model calibration and model verification based on observed or experimental data sets (difference between observation and experimentation will be introduced in section 2).

The main objective of this paper, based on real examples and case studies, is to present and discuss some points regarding the joint use of stormwater quality models and of experimental data. Its (modest) ambition is to contribute to a more pertinent perception and a more rational application of models and data.

The following sections will present successively:

• some epistemological aspects of modelling and their consequences for the application of models and for the interpretation of modelling results (section 2)
some reminders about practical difficulties in model calibration and verification (section 3)
a first case study with a rather simple model showing how field data sets strongly affect model calibration and model verification (section 4)
a second case study with a more detailed model, including its operational use to design a stormwater detention tank (section 5).

Some epistemological aspects of modelling

There are several possible approaches and classifications of models, depending on the selected criteria (see e.g. Badiou, 1969; Walliser, 1977; Legay, 1997; Morton and Suarez, 2001). In this section, we will consider models mainly from an engineering point of view, i.e. the ability of models to answer operational questions like, e.g., conception, design, operation and/or control of systems. In our case, the systems are separate and/or combined sewer systems, which involve a complex combination of both natural (rainfall events, topography, soils, surface water and groundwater, etc.) and anthropologic (buildings, urbanisation and impervious surfaces, sewer systems, etc.) elements and phenomena. This paper is not dealing explicitly with the elaboration and the creation of models, but with the application of existing models.

An engineering model is not the reality, but a limited, simplified and partial representation of reality, i.e. a tool built for practical applications. It is then intrinsically false, but this may be compensated by other evaluation criteria like simplicity, ability to simulate and/or reproduce experiments or observations, or to deliver operational results. The phenomena occurring in sewer systems during storm events are very complex, heterogeneous, with very large space and time scales and variability. As written by the philosopher Pascal Nouvel, “if one wants to have a clear vision of complex aspects of reality, one shall neglect numerous aspects […] In order to understand something, one shall neglect many other things, and the model is the expression of such a process, the controlled process of negligence. ‘Don’t look at this, look only at that, and then things become clear’. […] The question is not to make some aspects to appear, but on the contrary to make some aspects to disappear in order to keep only a limited number of them” (Nouvel, 2002, p. 193). In a few words, the model is resulting from “a strategy of negligence” (Nouvel, 2002, pp. 193). As this process of negligence is necessarily, at least partly, subjective, it is obvious that any model is a representation, a “point of view” (Legay, 1997, pp. 48–49). The consequence is that many different models may be proposed, with the same level of legitimacy, to simulate the same phenomena. The question about the choice of the best model is then very hard, if not impossible to be answered: “Even the validation and the confrontation to reality by means of measurements do not allow to dissipate this plurality of approaches” (Bouleau, 1999, pp. 14–15).

From an epistemological point of view, the Popperian approach based on the refutability of scientific theories by experiments (e.g. Popper, 1959, 1972) should be replaced by the Quinian approach (Quine, 1951, 1960) based on the underdetermination of theories (and models) by the facts. Indeed, engineering models should not be considered as scientific theories which are usually more general and, in principle, without any internal and formal contradiction (even if scientific theories may be included or embedded in models). Contrarily to the Popperian method according to which experiments and observations should be used to reject false theories, the Quinian approach claims that measurements and field data do not allow us to clearly refute any model: there are always several possible models to simulate the same phenomena. Engineering models cannot be falsified and can be constantly improved (Bouleau, 1999, p. 348). Then, the criterion to distinguish and select models is not their ability to reproduce experiments and observations, but their
ability to be used for some given purposes. “The quality of a model cannot be thought in
the absolute, it depends essentially on what the model will be used for, and on the prac-
tical possibilities of validation” (Bouleau, 1999, p. 273). “A model is a useful simulacrum”
(Bouleau, 1999, p. 301). From this point of view, the most detailed models are not
necessarily the most useful ones. As written differently by the French poet Paul Valéry:
“What is simple is always false, what is not simple is unusable” (Valéry, 1942, p. 143).
One may also find some epistemological reasons to consider a model as a simulacrum:
the person “who rebuilds a simulacrum of a phenomenon or a being, by means of tech-
niques named modelling, necessarily knows something about the phenomenon or the
object he/she simulates” (Moles, 1995, p. 190).

When considering complex environmental phenomena like those occurring in sewer
systems during storm events, data sets allowing us to fully test models (in the Popperian
way) are extremely rare. Using the terminology proposed by Abraham Moles, one may dis-
tinguish observation sciences with weak interaction with phenomena such as astronomy,
and experimental sciences with strong interaction with the phenomena such as physics
(Moles, 1995, pp. 99 and 164–165). These two categories represent the extremities of a
continuum with various degrees from strict observation to full experimentation. Among
environmental sciences, urban hydrology may be considered as being essentially an obser-
vation science (even if some degree of experimental work under controlled conditions may
exist sometimes for some aspects). Consequently, not all variables, parameters, influencing
factors and phenomena can be known or controlled; measurements cannot be exhaustive in
time and space. Moreover, due to both the high natural variability of phenomena and the
unavoidable uncertainty in field data (measurements done and sensors used in sewer sys-
tems), observation data sets to which models are compared are uncertain and, as a conse-
quence, do hinder the test, comparison and selection of models. In a few words, if many
models reproduce observations with errors equivalent to the uncertainty in observations, it
will be impossible to rank the models according to this test.

Most frequently, stormwater quality models have been developed as sets of equations,
each equation (or sub-set of equations) being considered as the description of a given
part or of a fraction (sub-process) of the whole process analysed as a chain of successive
steps or sub-processes (see e.g. Ashley et al., 1999; Rauch et al., 2002). A strong trend,
since the 1970s, lies in “improving” the models by adding more and detailed equations,
each “improvement” being justified either by the inability of the previous models to
reproduce some given observations (even if this inability has not always been clearly
established) or by the willingness to include in the model a better level of scientific
knowledge obtained from small scale observations (laboratory and in situ experiments
and observations). In many cases, these model refinements do not really improve their
engineering usability. On the contrary, they lead to increased difficulties. Some are
numerical difficulties: they will be described briefly in section 3. Other difficulties are
more epistemological difficulties, presented in the following paragraphs.

Following centuries of scientific analytical approach initiated by famous scientists like
Descartes, Newton, Leibniz, etc., some models are based on the belief that knowing pro-
cesses at small scale and under known conditions is sufficient to build adequate models
at large scale by integration. If this approach has been very successful, e.g. in mechanics,
is limits are reached very rapidly when considering large and complex systems like
urban catchments and sewer systems. As an example, let us consider solid transport in
sewer systems. In the laboratory or at small in situ scale, under controlled conditions, it
is possible to derive models representing solid transport and related phenomena like
deposition and erosion. Once such a model is established at small scale, it is tempting to
include it in a software package for large scale simulations (scaling-up process) e.g. to
simulate solid transport along hundreds of kilometres of sewer pipes in a large city. Unfortunately, many difficulties appear (see e.g. Ashley et al., 2005). Some are theoretical, some are practical.

Among the theoretical difficulties, the first one lies in the scaling-up process. Contrary to the traditional analytical approach, the systemic approach (e.g. von Bertalanffy, 1968; de Rosnay, 1975) revealed 50 years ago that properties and behaviour of complex systems are not simply the combination of properties and behaviour of their elementary components. Some emerging properties of complex systems, at large or global scale, cannot be explained only by the integration of elementary processes. Moreover, “we know that, generally, ‘the scale creates the phenomenon’ (C.E. Guye) and that the very essence of phenomena or forms changes when the scale [of observation] changes” (Moles, 1995, p. 136).

Another theoretical difficulty is the frequent semantic sliding of vocabulary which reflects some over-confidence in models and over-interpretation of their results. Once a model has been calibrated and verified (i.e. it is able, after having adjusted its parameters, to reproduce some specific sets of observations with a level of approximation which, according to uncertainties in data and parameters, is considered as acceptable by the user), it is applied to simulate phenomena and events for which no observations are available. This application is based on some hypotheses which are usually implicitly accepted without serious discussion. Reproducing observations is not equivalent to explaining phenomena: this error is analogous to the assumption that correlation is equivalent to a cause–effect relationship. Even if the model gives good results in verification, it does not mean that it really explains the involved phenomena: “Predicting is not explaining” (Thom, 1991). “Reproduction of historical data may be no more than mimicry” (Oreskes and Belitz, 2001, p. 32). This limits strongly the application of the model to new contexts or conditions. Indeed, it is rarely (or only very lately) demonstrated that a verified model can be applied to other contexts than those used for its calibration and verification: for example, models are used to simulate the future, assuming that the future will be like the past. But some trends or aspects in the model and in the past observations may have been ignored or compensated, and the prediction may be wrong. One of the main risks in modelling is the frequent “undue extension of the domain of validity of the model” (Legay, 1997, p. 54). This is particularly true in environmental engineering, with significant consequences for decision making processes (see e.g. Lung, 2004). “Why do people make and use models that they know are conceptually flawed and have perhaps already failed in a predictive mode?” (Oreskes and Belitz, 2001).

A strong limitation in the scaling-up process is the amount of information. At small scale, it is rather easy to collect all the necessary and detailed information in order to establish the model, to calibrate and to verify it. But, at large scale, collecting the necessary information is frequently impossible. In the example of solid transport modelling in sewers, a key variable is the solid inputs in a given sewer reach. This information can be measured or at least be rather well estimated at small scale, but it cannot be obtained at the large scale of a whole city. Consequently, the detailed model obtained at small scale, even if its equations and concepts could be valid at large scale (which is not guaranteed but usually assumed by the user), cannot be applied at this large scale because it is not feasible (with available time, money, etc.) to collect the required information. Sometimes,
default values are suggested to replace the missing local information, but default values are truly a lack of values which cannot be compensated.

**Numerical difficulties in model calibration and verification**

Including more equations frequently leads to stormwater quality models which are over-parameterised, with high levels of auto-correlation and possible compensation effects leading to difficult, if not random, calibration. Model simplification and reformulation may then be necessary for adequate operational use. This has been frequently observed, even with simple models like the surface pollutant build-up model based on the results of Sartor *et al.* (1974), cited by Alley and Smith (1981), and implemented in many software packages. The model is given by Equation 1 which, after integration, leads to Equation 2:

\[
\frac{dM}{dt} = AS_a - DM \tag{1}
\]

\[
M(t) = \frac{AS_a}{D} (1 - e^{-Dt}) + M_r e^{-Dt} \tag{2}
\]

with

- \(M\) mass of pollutants (kg) accumulated on the catchment surface at time \(t\) (days) since the previous storm event
- \(S_a\) active area of the catchment (ha)
- \(A\) accumulation rate (kg ha\(^{-1}\) day\(^{-1}\))
- \(D\) disappearing coefficient (day\(^{-1}\))
- \(M_r\) mass of pollutant (kg) remaining on the catchment surface after the previous storm event.

During dry weather, \(M\) is growing asymptotically until a maximum mass equal to 
\(\frac{AS_a}{D}\) is reached. As a consequence, the parameters \(A\) and \(D\) are correlated and cannot be calibrated independently: many couples of values are usable. Recently, Kanso (2004) worked again, with probabilistic methods, on the calibration of this model and suggested a re-writing of the equation in order to obtain independent parameters. If model calibration is highly recommended by researchers, its effective application by practitioners remains difficult. Many commercial packages are very poor in assisting the user, who has to proceed with trials and errors based on roughly qualitative (sometimes only visual) estimations to reach a hypothetical optimum fitting, based on very limited data sets (Gromaigre *et al.*, 2002).

Due to the high number of their parameters, many models are affected by equifinality and indeterminability, especially physically based distributed models (see e.g. Beven, 1996, 2001; Beven *et al.*, 2001): several sets of parameters are able to give the same model outputs, or at least outputs which cannot be meaningfully distinguished from one another. Consequently, the best or optimum set of parameters cannot be obtained. Search for it may even be vain, as the unique optimum set of parameters does not exist independently from the calibration algorithms used, even the most sophisticated ones. The choice of one set of parameters among all possible sets is a very difficult question. This is usually an opportunity to reduce the complexity of the model, by applying the principle of parsimony (Occam’s razor): do not introduce more hypotheses, equations and parameters than necessary for a given objective and a given level of uncertainty. An alternative consists of increasing the number of observations under various conditions, in order to decrease the number of acceptable sets of parameters.

Once a model has been calibrated, it is usually submitted to verification. We recommend to abandon the term validation, which, according to many authors, leads to an undue confidence in the model. Indeed “the term ‘validation’ is an unfortunate one,
because its root – valid – implies a legitimacy that we may not be justified in asserting” (cited by Anderson and Bates, 2001, p. 1). As revealed by an enquiry among French practitioners (Gromaire et al., 2002), calibration itself is not systematic and verification is seldom applied. Among the main reasons explaining this situation, the two most important ones are i) the insufficient assistance for calibration and verification in commercial packages and ii) the lack of field data sets. The lack of field data is the most critical aspect, with serious consequences for both calibration and verification.

In the following sections, case studies will illustrate how some of the above difficulties appear in stormwater quality models and their consequences for operational objectives.

**First case study with simple models**

**Regression model**

Among the various models proposed to simulate stormwater quality, regression models are frequently used to calculate EMC values (Event Mean Concentration) (see e.g. Servat, 1984; Driver and Tasker, 1990; Hoos and Sisolak, 1993; Saget, 1994; Hoos, 1996; Irish et al., 1998). In this section, we use one of the regression models implemented in the French commercial software Canoè (INSA/SOGREAH, 1999), given by Equation 3:

\[ C = aH^bI_{max 5}^cADWP^d \]  

with

- \( ADWP \) antecedent dry weather period (hours)
- \( C \) event mean concentration (mg.L\(^{-1}\))
- \( H \) rainfall depth (mm)
- \( I_{max 5} \) rainfall maximum intensity in 5 minutes (mm.h\(^{-1}\))
- \( a, b, c, d \) parameters to be calibrated.

**Observations**

Seventeen storm events have been monitored in a 12 ha urban catchment equipped with a combined sewer system. All data (event mean concentrations of TSS – total suspended solids, rainfall depth, intensity and antecedent dry weather period) are presented in chronological order in Table 1. These \( N = 17 \) events show the typical variability of all variables.

**Table 1** Values of \( C, H, I_{max 5}, \) and \( ADWP \) measured for 17 storm events

<table>
<thead>
<tr>
<th>n° event</th>
<th>( C ) (mg TSS.L(^{-1}))</th>
<th>( H ) (mm)</th>
<th>( I_{max 5} ) (mm.h(^{-1}))</th>
<th>( ADWP ) (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>834</td>
<td>3.2</td>
<td>2.4</td>
<td>141.2</td>
</tr>
<tr>
<td>2</td>
<td>602</td>
<td>3.4</td>
<td>3.2</td>
<td>374.2</td>
</tr>
<tr>
<td>3</td>
<td>707</td>
<td>4.6</td>
<td>8.8</td>
<td>7.3</td>
</tr>
<tr>
<td>4</td>
<td>625</td>
<td>4.0</td>
<td>5.2</td>
<td>11.0</td>
</tr>
<tr>
<td>5</td>
<td>288</td>
<td>6.4</td>
<td>3.2</td>
<td>16.4</td>
</tr>
<tr>
<td>6</td>
<td>410</td>
<td>6.0</td>
<td>3.2</td>
<td>48.3</td>
</tr>
<tr>
<td>7</td>
<td>402</td>
<td>4.8</td>
<td>3.2</td>
<td>44.0</td>
</tr>
<tr>
<td>8</td>
<td>914</td>
<td>15.0</td>
<td>6.4</td>
<td>48.0</td>
</tr>
<tr>
<td>9</td>
<td>632</td>
<td>3.4</td>
<td>4.8</td>
<td>26.6</td>
</tr>
<tr>
<td>10</td>
<td>399</td>
<td>3.0</td>
<td>3.2</td>
<td>0.6</td>
</tr>
<tr>
<td>11</td>
<td>743</td>
<td>3.6</td>
<td>4.8</td>
<td>43.5</td>
</tr>
<tr>
<td>12</td>
<td>474</td>
<td>2.8</td>
<td>2.4</td>
<td>56.0</td>
</tr>
<tr>
<td>13</td>
<td>331</td>
<td>4.6</td>
<td>4.0</td>
<td>0.7</td>
</tr>
<tr>
<td>14</td>
<td>625</td>
<td>5.2</td>
<td>10.4</td>
<td>14.8</td>
</tr>
<tr>
<td>15</td>
<td>548</td>
<td>10.6</td>
<td>2.4</td>
<td>21.6</td>
</tr>
<tr>
<td>16</td>
<td>508</td>
<td>5.6</td>
<td>4.8</td>
<td>86.1</td>
</tr>
<tr>
<td>17</td>
<td>1,760</td>
<td>14.0</td>
<td>31.2</td>
<td>30.1</td>
</tr>
</tbody>
</table>
Calibration

Three different calibrations are presented in this section, as given in Table 2. Calibration 1 corresponds to a short measurement campaign, during which only the $n = 6$ first events have been measured. Calibration 2 is similar, but with the $n = 6$ last events. Calibration 3 is based on all $n = 17$ events, in order to use all available information. This 17 events data set is rather large compared to the usual operational practice revealed by the French enquiry: in 50% of the cases, up to 5 events are available for calibration, and up to 10 events in more than 80% of the cases (Gromaire et al., 2002).

The calibration has been made simply by means of the ordinary least squares method after logarithmic transformation of Equation 3:

$$
\ln(C) = \ln(a) + b \ln(H) + c \ln(I_{max}) + d \ln(ADWP)
$$

and by minimising the sum $S$:

$$
S = \sum_i (C_i - C_{mi})^2
$$

with $C_i$ and $C_{mi}$ respectively the observed and calculated EMC values for the event $n^o i$, and $i$ the index corresponding to the event numbers given in Table 2. This method gives directly the optimum values of the four parameters $a$, $b$, $c$ and $d$, and also their 95% confidence intervals.

Results

For the three calibrations, the results obtained are:

- the values of the four parameters and their 95% confidence intervals (Table 3)
- the graphs showing $C_m$ versus $C$, with circles for the events used for calibration and diamonds for the events used for verification (Figure 1)
- the total mass calculated for the 17 events (Figure 2).

### Table 2 Events used for calibration and verification

<table>
<thead>
<tr>
<th>Calibration n°</th>
<th>n° of the events used for calibration</th>
<th>n° of the events used for verification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 to 6</td>
<td>7 to 12</td>
</tr>
<tr>
<td>2</td>
<td>12 to 17</td>
<td>1 to 11</td>
</tr>
<tr>
<td>3</td>
<td>1 to 17</td>
<td>none</td>
</tr>
</tbody>
</table>

### Table 3 Calibration results

<table>
<thead>
<tr>
<th>Calibration</th>
<th>Parameter</th>
<th>Optimum</th>
<th>95% min. value</th>
<th>95% max. value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$a$</td>
<td>3.220</td>
<td>5.62</td>
<td>1.843,510</td>
</tr>
<tr>
<td></td>
<td>$b$</td>
<td>-1.308</td>
<td>-3.359</td>
<td>0.743</td>
</tr>
<tr>
<td></td>
<td>$c$</td>
<td>0.209</td>
<td>-1.360</td>
<td>1.777</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>-0.030</td>
<td>-0.591</td>
<td>0.530</td>
</tr>
<tr>
<td>2</td>
<td>$a$</td>
<td>128</td>
<td>22.12</td>
<td>743</td>
</tr>
<tr>
<td></td>
<td>$b$</td>
<td>0.326</td>
<td>-0.703</td>
<td>1.355</td>
</tr>
<tr>
<td></td>
<td>$c$</td>
<td>0.365</td>
<td>-0.241</td>
<td>0.972</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>0.110</td>
<td>-0.183</td>
<td>0.402</td>
</tr>
<tr>
<td>3</td>
<td>$a$</td>
<td>193</td>
<td>107.64</td>
<td>347</td>
</tr>
<tr>
<td></td>
<td>$b$</td>
<td>0.006</td>
<td>-0.332</td>
<td>0.345</td>
</tr>
<tr>
<td></td>
<td>$c$</td>
<td>0.482</td>
<td>0.216</td>
<td>0.748</td>
</tr>
<tr>
<td></td>
<td>$d$</td>
<td>0.110</td>
<td>0.018</td>
<td>0.202</td>
</tr>
</tbody>
</table>
Figure 1  \( C_m \) versus \( C \) for all 17 events (arrows in the top graph correspond to inaccurately simulated events: see main text)

J.-L. Bertrand-Krajewski
Discussion

According to Figure 1, both calibrations (i.e. circle dots) 1 and 2 appear satisfactory and not very different from one another. The regression explains respectively 87% and 91% of the total variance in calibrations 1 and 2. This difference can be seen on the graphs, where the circle dots are a little bit closer to the diagonal in calibration 2 than in calibration 1. However, in the verification shown by diamond dots, the results are significantly different: the dispersion is higher in calibration 1 than in calibration 2, especially for the four events marked by an arrow. This indicates that, despite a rather equivalent level in calibration, verification levels are very different. As a consequence, the total masses predicted by calibrations 1 and 2 are very different, as shown in Figure 2: respectively 5,032 and 8,656 kg of TSS, i.e. a relative increase of 72% in calibration 2. Compared to the observed total mass equal to 9,099 kg of TSS, it appears clearly that calibration 2 is better. But, frequently, practitioners have only a limited data set and such comparisons are rarely made. This example shows that model calibration is very sensitive to the observations used: for the same amount of observations (six events), results may differ dramatically. The influence of event n° 17, with a high concentration equal to 1,760 mg/L, appears very important.

The results given in Table 3 show that the parameters $a$ and $b$ are the most sensitive ones to the data set used for calibration. They vary significantly between calibrations 1 and 2. For example, $a$ changes from 3,220 to 128, with very large confidence intervals. Similarly, $b$ changes from −1.308 to 0.326, also with very large confidence intervals. The fact that $b$ can be either negative or positive dramatically affects its role in the model. A further analysis shows that confidence intervals for $b$ decrease from calibration 1 to calibration 3, as shown in Table 3 and in Figure 3. In calibration 3, with 17 observations instead of 6 in calibrations 1 and 2, the optimum value of $b$ becomes 0.006, i.e. very close to zero, and is located in a rather narrow confidence interval [−0.332, 0.345]. One may infer that $b$ could be likely taken as equal to zero, which means that the rainfall depth $H$ has no influence for this catchment and this data set, and that, consequently, it could be removed from Equation 3. One observes here very clearly how increasing the number of observations has a clear effect on the width of confidence intervals.

Calibration 2 and calibration 3 presented in Table 3 and in Figure 1 appear as rather similar. The calculation of the total mass is also equivalent: respectively 8,656 and 8,485 kg of TSS, i.e. a relative difference of −2%, as shown in Figure 2. Both are close to the observed value of 9,099 kg, with differences respectively equal to −5 and −7%,
which is smaller than the uncertainty in the observed (i.e. measured) total mass estimated at approximately 25–30%. In this particular case, it seems that increasing the number of observations from 6 to 17 does not significantly change the results. But this conclusion is erroneous. There is a significant difference: in calibration 3, confidence intervals are narrower for all parameters (see Table 3), which allows the user to have a higher confidence in the model.

Additional remarks
This case study with a simple model shows clearly how calibration is sensitive to the data sets. For a given number of observations, very different parameter values are obtained. If the calibration may be similar (which is not always the case), the verification may be different. Increasing the number of observations may improve calibration and verification, and may also decrease the width of the parameters’ confidence intervals, but this is not systematically observed.

An extended analysis has been carried out for three regression models M1, M2 and M3 (M1 and M2 are given respectively by Equations 6 and 7, while M3 is given by Equation 3):

\[ C = a ADWP^b I_{\text{max}}^c V_r^d \]

\[ C = a H^b D^c V_r^d \]

with
\( ADWP \) antecedent dry weather period (hours)
\( C \) event mean concentration (mg/L)
\( D \) storm event duration (hours)
\( I_{\text{max}} \) rainfall maximum intensity in 5 minutes (mm/h)
\( H \) rainfall depth (mm)
\( V_r \) total runoff volume (m³)
\( a, b, c, d \) parameters to be calibrated.

with EMCs and event loads of TSS measured for \( N = 40 \) events in Le Marais catchment in Paris (Mourad, 2005; Mourad et al., 2005). To evaluate the influence of data sets, the
A re-sampling method has been applied in order to simulate virtual measurement campaigns with different amounts of measured events and data characteristics. A measurement campaign is considered as a random sampling of events occurring within a given period of time. This approach can be supported by the fact that sampling may not be done systematically for successive events. Among the $N = 40$ available observations, $n \leq N$ observations are sampled randomly to create a subset of events. Each event can only appear once in a given subset. The three models are calibrated with this subset of $n$ events, and the $N - n$ remaining observations are used for verification. This procedure is repeated 1,000 times for a given size $n$, in order to analyse how different subsets of the same size may lead to different calibration and parameters values. The complete procedure is applied for $n = 4$ to 37 ($n$ has been limited to 37 to keep $N - n = 3$ events for verification), in order to analyse the effect of the number $n$ of observations. The root mean squared error $RMSE$ (Equation 8) is used as an indicator of the calibration and verification quality:

$$RMSE = \sqrt{\frac{\sum (C_i - C_{mi})^2}{n}}$$

(8)

with $C_i$ and $C_{mi}$ respectively the observed and the calculated event mean concentrations of the $i$th event and $n$ the number of events in the subset used for calibration.

Figure 4 shows the results of the $RMSE$ calculated for EMCS. The top series of three graphs shows the confidence intervals of the $RMSE$ for the three models as a function of the number $n$ of events used for calibration. The bottom series of three graphs shows the same information, but for verification. For example, with the model M3 and $n = 6$ events as in the above paragraph 4.4, the mean error in EMC, for the 1,000 calibrations, ranges from approx. 5 mg/L to approx. 90 mg/L depending on the subset of observations used for calibration (dashed lines), and ranges from 85 mg/L to more than 250 mg/L in verification. The width of the calibration confidence interval decreases when $n$ increases. On the contrary, the width of verification confidence intervals is at its minimum for $n$ approximately equal to 15–20 events. If very few events are used for calibration, the probability is high, in verification, to find events which are badly reproduced by the model. This is similar to calibration 1 above. When $n$ increases up to 15–20 events, both calibration and verification confidence intervals diminish. When $n$ increases beyond 20 events, the calibration has a high probability of being improved (more information is included), but, as the calibration is global, the probability to find some events in the verification which are badly reproduced increases again. Both models M1 and M3 give zero as the confidence interval width for $n = 4$ because they have four parameters; thus the model can fit perfectly any set of four observations. Both models showed also identical confidence intervals for calibration and verification. The model M2 gives less satisfactory results. More detailed results are given in Mourad (2005).

**Second case study with an EMC model and a detailed model**

**Brief presentation of the detailed model**

Numerous detailed models have been proposed during the last two decades. Most of them distinguish between the surface build-up process (i.e. pollutant accumulation on catchment surfaces), wash-off process (i.e. erosion and transport by surface runoff of accumulated pollutants), and accumulation, transport, deposition and erosion of solids and deposits in sewer pipes. Equations 9 to 11 are typical examples of equations implemented in commercial...
software like MouseTrap or InfoWorks (see e.g. Bouteligier et al., 2003).

\[ M_a(t) = \frac{ACCU}{\text{DISP}} \cdot C_{\text{imp}} \cdot A \cdot (1 - e^{-\text{DISP} \cdot t}) + MR \cdot e^{-\text{DISP} \cdot t} \]  

(9)

wash-off \[ M_e(t, t + \Delta t) = Md(t) \cdot \left(1 - e^{-a_1 \left(\frac{Q}{A} + \frac{s}{b} \right)^{a_2}} \cdot \Delta t \right) \]  

(10)

in sewer transport, erosion and deposition processes

\[ C' = j \cdot \left(\frac{W_e \cdot R}{A}\right)^{\alpha} \cdot \left(\frac{d_{50}}{R}\right)^{\beta} \cdot \lambda^\gamma \cdot \left(\frac{|u|}{\sqrt{g \cdot (s - 1) \cdot R}} - K \cdot \lambda \cdot \left(\frac{d_{50}}{R}\right)^{\delta}ight)^m \]  

(11)

with \( M_a \) accumulated pollutant load at time \( t \) (kg), \( ACCU \) build-up rate (kg.day\(^{-1}\).ha\(^{-1}\)), \( DISP \) decay factor (day\(^{-1}\)), \( C_{\text{imp}} \) imperviousness coefficient, \( A \) catchment area (ha), \( MR \) residual load after the previous event (kg), \( M_e \) load washed off during the time step \( \Delta t \) (kg), \( Q \) surface runoff (m\(^3\).s\(^{-1}\)), \( \alpha_1, \alpha_2 \) and \( \alpha_3 \) parameters, \( C' \) dimensionless transport capacity, \( R \) hydraulic radius (m), \( d_{50} \) median particle size (m), \( s \) specific gravity, \( W_e \) effective bed width (m), \( \lambda \) composite friction factor, \( u \) flow velocity (m.s\(^{-1}\)), \( g \) gravity (m.s\(^{-2}\)), and \( A, \alpha, \beta, \gamma, \delta, \epsilon, m, n, j, K \) coefficients related to the dimensionless grain size \( D_{gr} \) (CIRIA, 1996).

The above set of equations has been coupled to the hydrodynamic software Canoé (INSA/SOGREAH, 1999) based on the Barré de Saint-Venant equations in order to simulate TSS pollutographs in Le Marais catchment in Paris. The sensitivity of the detailed model to the data sets used for its calibration and verification has been analysed as above, with 200 sub-sets of events for each size \( n \) of data sets ranging from 4 to 37. Results identical to those given in Figure 4 for EMC models have been obtained (see Mourad, 2005 or Mourad et al., 2005 for detail).

Operational application: design of a stormwater detention tank

Among the several possible operational applications of stormwater quality models, the design of a stormwater detention tank has been selected as a representative example (Mourad, 2005). The questions to be analysed were: how the design of the tank may depend on the type of model used? How the design may be sensitive to the data sets used
for calibration and verification? Only some aspects are presented in the following paragraphs.

The stormwater detention tank has been designed according to the scheme presented in Figure 5. The total discharge at the outlet of the catchment is $Q$. The maximum inflow in the downstream WWTP is $Q_1$. If $Q > Q_1$, the discharge $Q_2 = Q - Q_1$ is entering into the stormwater detention tank, if the tank is not full. If the tank is full and if $Q$ remains higher than $Q_1$, the water is sent directly into the receiving water and $Q_4 = Q_2$. When $Q$ is again equal to its dry weather values, the tank is emptied with an outflow $Q_5$ such that $Q + Q_5$ cannot be greater than $Q_1$. Three annual rainfall series with a total of 542 storm events have been simulated continuously, accounting for the fact that the tank may not be empty at the beginning of a new storm event if two storm events are close to each other, and with different hypotheses for the maximum value of $Q_1$.

The results presented hereafter have been obtained for a design criterion equal to the annual interception efficiency $IE$ expressed as the ratio (annual pollutant load entered into the tank and the WWTP/annual pollutant load produced by the catchment during all storm events during the year), i.e.:

$$IE = \frac{M_1 + M_2}{M_{tot}}$$  \hspace{1cm} (12)

The tank has been designed with three models:
- the hydraulic model (i.e. the ratio of the loads is simply the ratio of the volumes)
- the EMC regression model M3
- the above detailed model.

The EMC pollutant model has been calibrated and applied 1,000 times respectively for $n = 5, 10$ and $35$ events used for calibration. The detailed model has been calibrated and applied 200 times only for each $n$ value due to high calculation times. The hydraulic model has been calibrated independently with measured discharges. Some of the final results are presented in Table 4.

For the hydraulic model, the results of course do not depend on the number of events used to calibrate the pollutant model. Differences from year to year are only due to rainfall series: for the year 1 (resp. years 2 and 3), the tank volume should be 15.7 m$^3$/ha (resp. 21.2 and 17.9 m$^3$/ha) to intercept 90% of the annual volume. The effect of the
annual variability is evident here, with a difference of +35% in the tank volume when comparing 21.2 to 15.7 m³/ha. When considering the complete period of 3 years, the required tank volume should be 19.7 m³/ha.

For the EMC model, the tank volume strongly depends on the data sets used for calibration. For example, when the model is calibrated with $n = 10$ events, the mean volume required to intercept 90% of the annual pollutant load is equal to 14.5 m³/ha for year 1 (resp. 22.4 and 17.2 m³/ha for years 2 and 3). In 95% of the cases (95% confidence intervals), the volume ranges from 0 to 23.9 m³/ha for year 1 (resp. 4.4 to 41.6 and 0 to 43.4 m³/ha for years 2 and 3): this high variability is only due to the data sets used to calibrate the model. In other words, the tank design is highly sensitive to the model calibration. One observes also that confidence intervals decrease when $n$ increases: for year 1, the 95% confidence intervals are respectively 0–44.9, 0–23.9 and only 12.8–17.3 m³/ha for $n = 5$, 10 and 35 events. This indicates that the design will be more reliable when the model is calibrated with more events. Lastly, it is important to observe that, for the whole period of 3 years, mean volumes are not very different from those obtained with the hydraulic model. This means that, for this catchment, an EMC model does not lead to a significantly different design for a given annual interception efficiency, compared to a hydraulic model. This conclusion is of course not valid for all possible operational applications of EMC models.

For the detailed model, conclusions regarding the annual variability and the calibration variability are analogous to those obtained for the EMC model. For example, when the model is calibrated with $n = 35$ events, the mean volume required to intercept 90% of the annual pollutant load is equal to 24.6 m³/ha for year 1 (resp. 27.0 and 22.3 m³/ha for years 2 and 3). In 95% of the cases (95% confidence intervals), the volume ranges from 22.2 to 33.8 m³/ha for year 1 (resp. 23.6 to 35.7 and 19.9 to 31.0 m³/ha for years 2 and 3). Similarly, the confidence intervals decrease when $n$ increases. But the mean volumes calculated with the detailed model are significantly different from the mean volumes calculated with the hydraulic and EMC models: for $n = 10$ events, the mean volume is 27.1 m³/ha, compared respectively to 19.7 and 20.2 m³/ha. In other words, in this case,
using a detailed model describing pollutographs leads to higher tank volumes. Once again, this conclusion cannot be generalised to any other case study. The difference between simple and detailed models may also differ from case to case.

Conclusions
From the above sections, the most important conclusions regarding stormwater quality models are the following.

Models should be considered as engineering tools rather than scientific theories and description of knowledge. Their main interest is their ability to answer operational questions like design, conception, management, etc. Accordingly, the most detailed and complex models are not necessarily the most useful and applicable ones. There is no absolute quality of models, but various degrees of usefulness. The model usefulness depends on a number of criteria.

Models are not the truth, but a partial, simplified, incomplete and subjective representation of reality. They can hardly be falsified, but can always be improved. They are underdetermined by facts and observations.

In complex systems like urban catchments and sewer systems, scale effects and scaling-up of models are very difficult questions. Various models should be proposed at the corresponding various scales of interest.

All models shall be properly calibrated and verified before any application. However, even well calibrated and verified models usually only reproduce observations: their real ability for prediction remains highly questionable. Reproducing observations is not explaining phenomena. Reproducing observations is not a guarantee for good and valid predictions.

In complex and large systems like urban catchments and sewer systems, only limited and non-exhaustive information (especially regarding space and time variability and heterogeneity) is available. This hinders calibration and verification. This also frequently hinders the application of detailed models requiring a lot of information.

Many models are over-parameterised. It is hard, if not impossible, to find the “optimum” or “best” unique set of parameters values. *Model calibration and verification, and consequently model outputs and results, dramatically depend on the data sets used for their calibration and verification.* One possibility to decrease this dependence consists in using large data sets incorporating many contexts and conditions, in order to reflect, as much as possible, the natural variability of the phenomena observed in urban drainage (problem of representativity). Compared to the present practice, collecting more and reliable data is absolutely necessary.

Further efforts should be devoted to account for uncertainties (in data, in models, in parameters and in model outputs).

Acknowledgements
The author gratefully acknowledges the financial support of the French RGCU “Réseau Génie Civil et Urbain” and of the SIAAP “Syndicat Interdépartemental pour l’Assainissement de l’Agglomération Parisienne” for the “MCFP” research project carried out in collaboration with both the Cereve in Paris, France and the University of Montpellier, France. This programme supported both PhD theses of Kanso (2004) and Mourad (2005). This paper also benefited from various discussions with colleagues in France, The Netherlands and Australia.
References


