CALIBRATION AND VALIDATION OF MULTIPLE REGRESSION MODELS FOR STORMWATER QUALITY PREDICTION: DATA PARTITIONING, EFFECT OF DATA SETS SIZE AND CHARACTERISTICS

M. MOURAD*, J-L. BERTRAND-KRAJEWSKI* and G. CHEBBO**
*Laboratoire URGC Hydrologie Urbaine, INSA de Lyon, 34 avenue des Arts, 69621 Villeurbanne cedex, France.
**CEREVE, ENPC, 6-8 av. Blaise Pascal, Cité Descartes, Champs sur Marne, 77455 Marne la Vallée - cedex 2, France.

Abstract
Two main issues regarding stormwater quality models have been investigated. i) The effect of calibration dataset size and characteristics on calibration and validation results. ii) the optimal split of available data into calibration and validation subsets. Data from 13 catchments have been used for 3 pollutants: BOD, COD and SS. Three multiple regression models were calibrated and validated. The use of different data sets and different models allows viewing general trends. It was found mainly that multiple regression models are case sensitive to calibration data. Few data used for calibration infers bad predictions despite good calibration results. It was also found that the random split of available data into halves for calibration and validation is not optimal. More data should be allocated to calibration. The proportion of data to be used for validation increase with the number of available data (*N*) and reach about 35 % for *N* around 55 measured events.

Keywords Stormwater, modelling, calibration, validation, uncertainty.

INTRODUCTION

Stormwater discharges from both combined and stormwater sewer systems are considered as a large source of pollutants into the receiving waters. Lately, growing concerns about the quality of surface waters have led to more demanding water legislation. The design of control and treatment facilities as well as management strategies, in many cases, require the estimation of discharged pollutant loads at different scales of time in some specific locations of the sewer system. A wide variety of stormwater quality models can be found in the literature, ranging from very simple ones such as the simple method (Schueler, 1987) to complex fully detailed models implemented in commercial softwares such as the well known Infoworks CS (Wallingford Software) and MouseTrap (Danish Hydraulic Institute). The complexity of a model can be quantified by the number of parameters or number of processes included in the model. Most of the existing models are based or coupled with statistical or empirical approaches, generally, incorporating conceptual parameters. Such hardly measurable parameters must be calibrated so the output of the model can match the observations used for calibration. The more the model is complex, the more observations are required for calibration to cover all possible conditions of use.

In practice, the models are generally used for predictive purposes. Good match of calibration data doesn't guarantee that future predictions will be good. To increase confidence in prediction ability of a model, validation is performed. There is not yet a complete agreement on what validation can or should encompass. Nevertheless some methods have been widely admitted like the split sample validation. Split sample validation consists of splitting available data into two samples. One sample is used to calibrate the model and the other one to test the prediction ability of the model. Independently of the modelling approach and research field, available data are

generally splitted randomly into halves (e.g. Kerlinger and Pedhazur, 1973; Schlütter, 1999; Vaze and Chiew, 2003). Jewell *et al.* (1978) suggest the same split ratio but stated also that if limited data are available, it is the usual practice to use the larger portion of the data for calibration and the smaller portion of the data for validation.

Among modelling approaches, multiple regression models (MRMs) have intermediate complexity. This approach has been used widely (e.g. Driver and Tasker, 1990; Saget, 1994). It can be considered as simple at first sight. Nevertheless great attention must be paid to hypotheses and limits of these models, which are frequently neglected by many users. A MRM aims to estimate a single variable by means of a set of other explanatory variables. In this study, the event mean concentration (EMC) is evaluated using the rainfall or/and the flow characteristics.

To carry out calibration and validation properly, large datasets are required. Unfortunately, in practice, the number of measured events is limited because of the high costs of monitoring campaigns and the relatively short time devoted for the studies. Thus engineers and researchers are often found dealing with limited datasets to be used for model identification, calibration and validation. In this case, what would be the effect of calibration dataset size and characteristics on the estimated relationship and how to optimally split the available observations into calibration and validation sets?

Three MRMs are implemented in the stormwarter quality module of the French software Canoe (Insa/Sogreah, 1999) having all the same structure. The models are presented in Table 1. In an attempt to explore the questions above, these models served as a bench test and were applied to data from 13 catchments and for three pollutants BOD, COD and SS.

M1	$EL = K \cdot ADWP^{a} \cdot I_{max5}^{b} \cdot Vr^{c}$ $EMC = EL/Vr = K \cdot ADWP^{a} \cdot I_{max5}^{b} \cdot Vr^{c-1}$
M2	$EL = K \cdot TP^{a} \cdot D^{b}$ $EMC = EL/Vr$
M3	$EMC = K \cdot ADWP^{a} \cdot TP^{b} \cdot I_{max5}^{c}$

Where EL is the event total load ; EMC is the event mean concentration ; ADWP is the antecedent dry weather period ; I_{max5} is the maximum intensity on 5 minutes period ; Vr is the total runoff volume ; TP is total precipitation ; D is the event duration ; K, a, b and c are calibration parameters.

Table 1 : the three MRMs used in the study

METHODS

CALIBRATION DATASET EFFECT

In order to study the effect of the size and characteristics of calibration datasets on the results of a model, subsets were sampled from the available data. The size n of subsets ranged from 4 to N - 2 where N is the number of available data. For each n a large number of subsets was drawn randomly without replacement (typically 1000). In fact the objective is to see how the results of a model could vary if a subset of events have been measured instead of all available data for the same period of time. Hence in each subset, each event can not occurs more than once and for each n, all subsets were distinct.

CALIBRATION

The type of the model determines the extent of calibration required. The desirable output of the model must be considered and be of focus. Generally, a model is calibrated by minimizing the distances between calculated and measured values. The sum of the squared errors is one of the

criteria most used for calibration. The use of a least squared errors approach infers constancy of error variance for all observations. However larger uncertainties in hydrologic variables are known to be associated to larger variables values. To cope with this problem, a logarithmic transformation is applied for all variables. This transformation achieves stability in the error variance, normality of residuals and linearity of the regression model, making it more easy to calibrate with ordinary least square method. To allow comparison between data sets independently of n, the Root Mean Squared Error (Eq. 1) is used as a measure of goodness of calibration.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(O_i - T_i\right)^2}{n}} \qquad \text{Eq. 1}$$

where O_i denotes observations and T_i denotes model outputs.
Calibration has been performed for the three models shown in Table 1 for BOD, COD and SS using data from 13 catchments. Data of 12 catchments have been extracted from the French database on stormwater discharges QASTOR (Saget and Chebbo, 1996). The thirteenth catchment is "Le Marais" catchment in Paris, France (Chebbo *et al.*, 1999). A summary of the available data for each catchment is given in Table 2.

Combined sewer systems				Separated Sewer systems			
Catchment	BOD	COD	SS	Catchment	BOD	COD	SS
La Briche d11	12	13	13	aixnord	33	41	41
La Briche dd11	9	9	9	aixzup	41	47	47
La Briche enghien	10	10	10	ulissud	29	29	29
La Briche PHI	16	16	16	velizy	22	22	22
La Briche PLB	16	16	16	Maurepas	59	59	59
Mantes	23	23	23	ulisnord	57	57	57
Le Marais	-	64	67				

Table 2 : Summary of available data (EMCs)

OPTIMAL DATA SPLIT

Let Y denote the $N \times 1$ vector of the dependent variable to be estimated and X denote the $N \times K$ matrix of independent variables, where N is the number of observations and K is the number of independent explanatory variables in the model. Suppose that $N_1 \left(N_1 \leq N\right)$ observations are selected randomly from the available dataset for calibration. The relationship between $Y1$ and $X1$ can be written as follows:

$$Y1_i = f\left(X1_{i\cdot}, \beta\right) + \varepsilon 1_i \qquad \text{Eq. 2}$$

where $i\cdot$ means that all the ith row of X is considered
β is the vector of the estimated parameters of the model
$\varepsilon 1$ is the ($N_1 \times 1$) vector of the model calibration errors

Let $Y2$ and $X2$ denote the N_2 observations held out for validation ($N_2 = N - N_1$). The prediction error is vector $\varepsilon 2$:

$$\varepsilon 2 = Y2 - \hat{Y}2 \qquad \text{Eq. 3}$$

where $\hat{Y}2$ is the predicted concentrations given by

$$\hat{Y}2_i = f\left(X2_{i\cdot}, \beta\right) \qquad \text{Eq. 4}$$

For each possible split of the available data (from $N_1 = 4$, $N_2 = N - 4$ to $N_1 = N - 2$, $N_2 = 2$) a large number of random possible splits is available. The number of returned splits was limited to 1000. The root mean squared error (RMSE) is used as an overall measure of goodness for validation as like as for calibration.

For each value N_1 one gets two distributions of the RMSE, one for calibration and one for validation. Since the validation aims to prove that the model is able to give results for prediction as good as for calibration, the two distributions can be compared. The results showed that the distributions are not always normally distributed. Hence the use of non-parametric tests is recommended. The idea is to test if the two distributions are identical. The best split will be the one that maximizes the probability to have two identical distributions. The most appropriate test in our case is the Wilcoxon rank sum test (Hollander and Wolfe, 1973). This is a test of the null hypothesis H_0 that two samples are drawn from the same distribution, against the alternative hypothesis H_1 that the distributions have different origins (Figure 1). The test returns the significance probability "*p*" that the two distributions are identical.

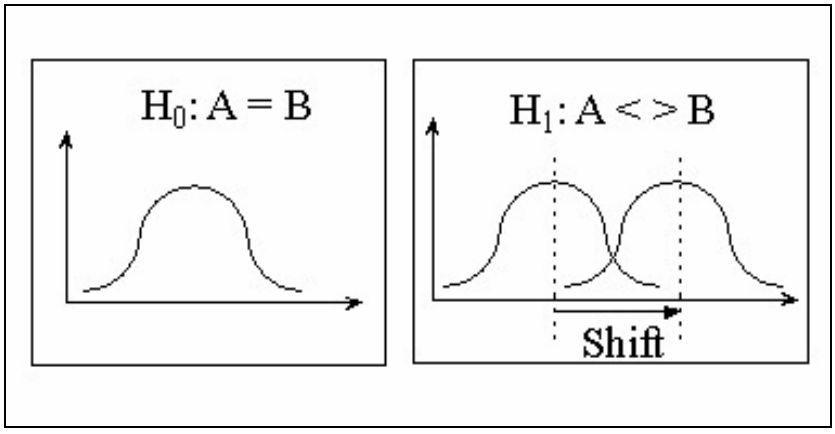

Figure 1: Illustration of the Wilcoxon test

RESULTS

CALIBRATION AND VALIDATION VARIABILITY

The procedures above were applied to each catchment, for each pollutant and the three models. Regardless the model, the pollutant and the catchment that have been used, the distribution evolution of the calibration RMSE as a function of the number of data used in calibration is quite the same (Figure 2). With only four events used for calibration, the models were able to fit perfectly the observations, of course with a different set of parameters for each calibration subset. This is not surprising, as it is easy to perfectly fit 4 rainfall events with a model having 4 calibration parameters. It was found also that generally more than 20 events are necessary to have a calibration RMSE distribution centred on the RMSE value obtained using all available data for calibration.

As well as for calibration distribution, the evolution of the validation distribution was also identical for all the catchments and pollutants using the three models. A typical plot is given in Figure 3. For n = 4 the RMSE values are very high and situated out of the figure for scale purposes. In fact, when few data are used for calibration, the model not only fits the data but it fits also measurement errors. In addition to this, data don't cover enough possible conditions. This explains the weakness in prediction ability of the calibrated model.

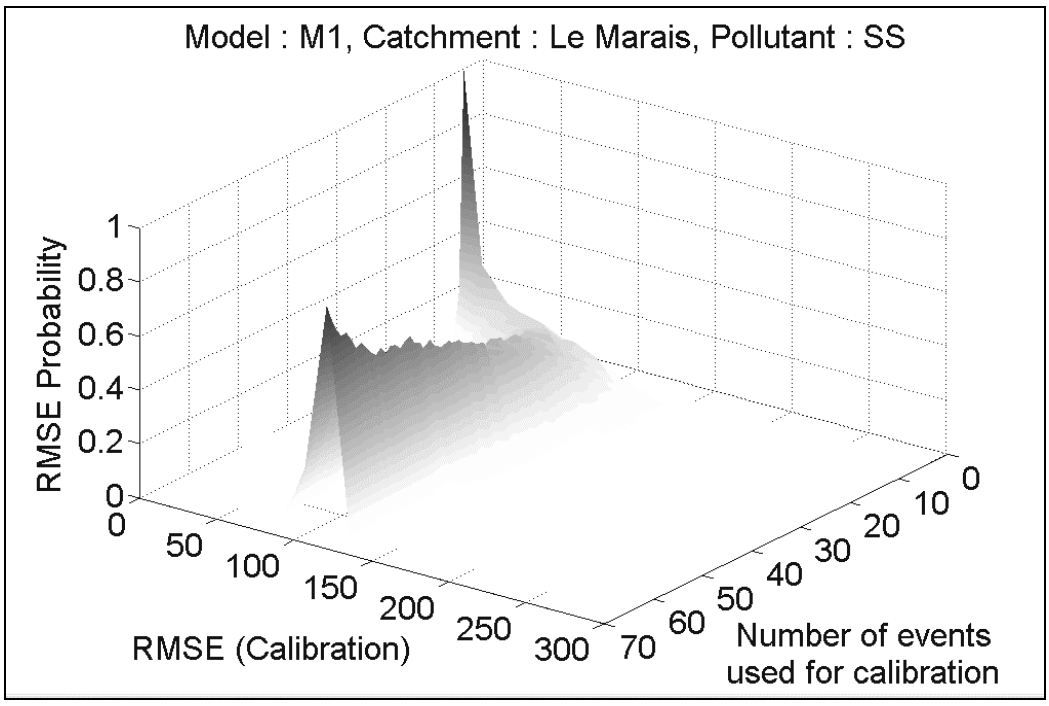

Figure 2 : Typical calibration distributions plot

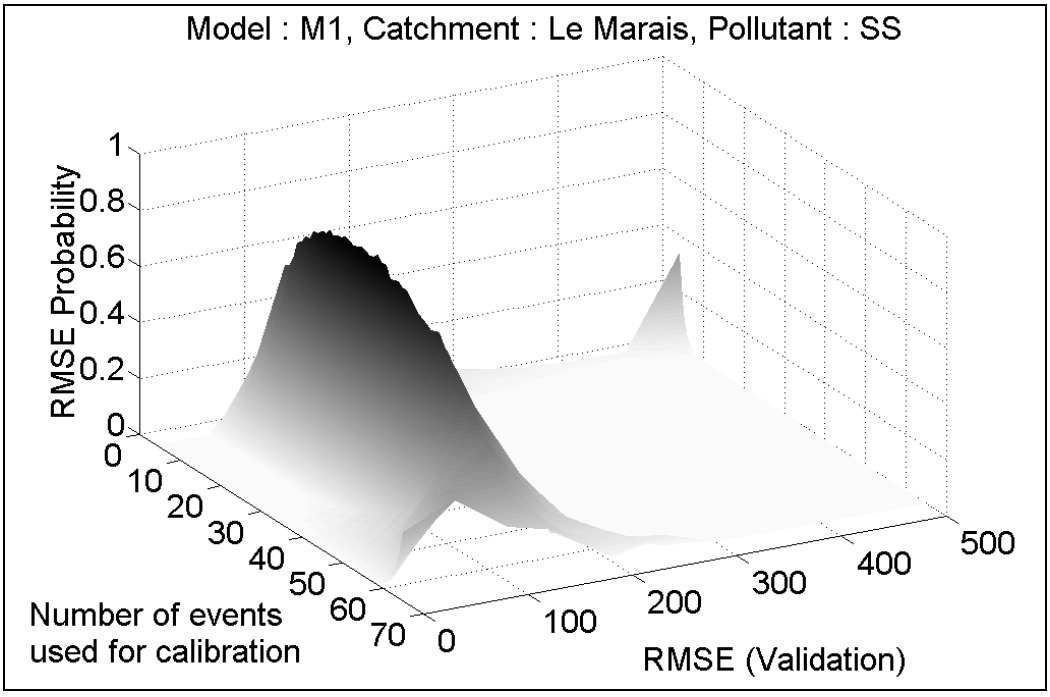

Figure 3 : Typical validation distributions plot

An analysis of the calibration subsets giving best results (most likely the RMSE using all data) will be carried out to identify potential characterized groups of events. This might help in optimising data collection and reducing costs.

DATA SPLIT

As it was mentioned previously, the Wilcoxon rank sum test was applied to test if the RMSE distributions for calibration and validation are identical for each split. For each data split, the significance probability is returned. The example of "Le Marais" catchment for COD using model M3 is shown in Figure 4. In this case the calibration and validation RMSE distributions are most likely identical when using N_1 = 49 events for calibration and N_2 = N - N_1 = 15 events for validation. This means that using approximately 25 % of the data for validation is statistically optimal. This ratio has been calculated for the three pollutants and the three models. A summary of the results is shown in Figure 5.

In Figure 5 the data ratio to be used for validation is drawn as a function of the size of the available dataset. Each point on the figures refers to a catchment. Hence, on each figure, 13 points can be seen except for BOD where only 12 points are present. It is shown that none of the optimal splits reached a 50 % ratio for validation. All cases showed data validation ratios less than 40 %.

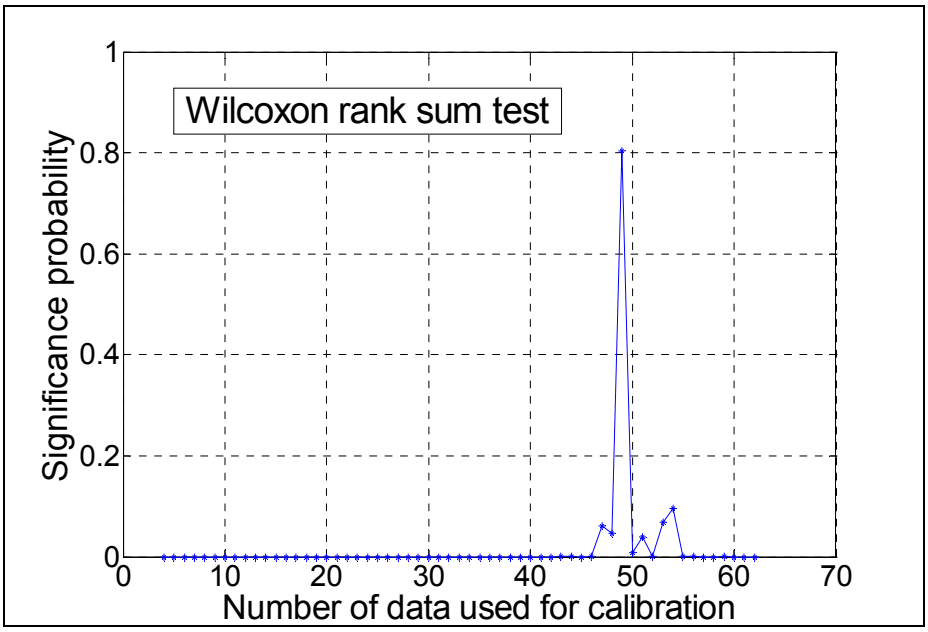

Figure 4 : Significance probability for Le Marais catchment for COD using M3

A general increasing trend can be observed for BOD, COD and less obvious for SS. For validation data ratio equal to zero, the validation and calibration RMSE distributions are not identical for all tested splits. If the number N of available data is less than 30 the optimal validation proportion varied between 0 and about 25 %. This proportion ranges from 25 to 40 % for N > 30. For SS it is more difficult to draw conclusions because of more dispersed points. For COD and SS "Le Marais" catchment (N > 60) reverses the trend and gives less validation proportion with higher N. the proportion of the data to be used for validation declines to near 20 %. One point beyond N = 60 doesn't allow any generalisation of this trend.

The scatters differ from one pollutant to another and from one model to another. Different model and different pollutant means different modelling errors. This gives an indication that *N* may not be the single determinant factor for an optimal split. Other factors like model error to signal ratio might tend to influence the optimal split. In fact, one can expect that more data must be used to calibrate a less good model or when calibration data have larger measurement errors.

However, it can be thought that beyond a given *N* (very high) any split might perform well. This is due to the fact that even a small proportion of the data can cover most of the possible conditions. Unfortunately this is not the case in this field of research where data collection is very expensive and only limited number *N* of data is used in modelling.

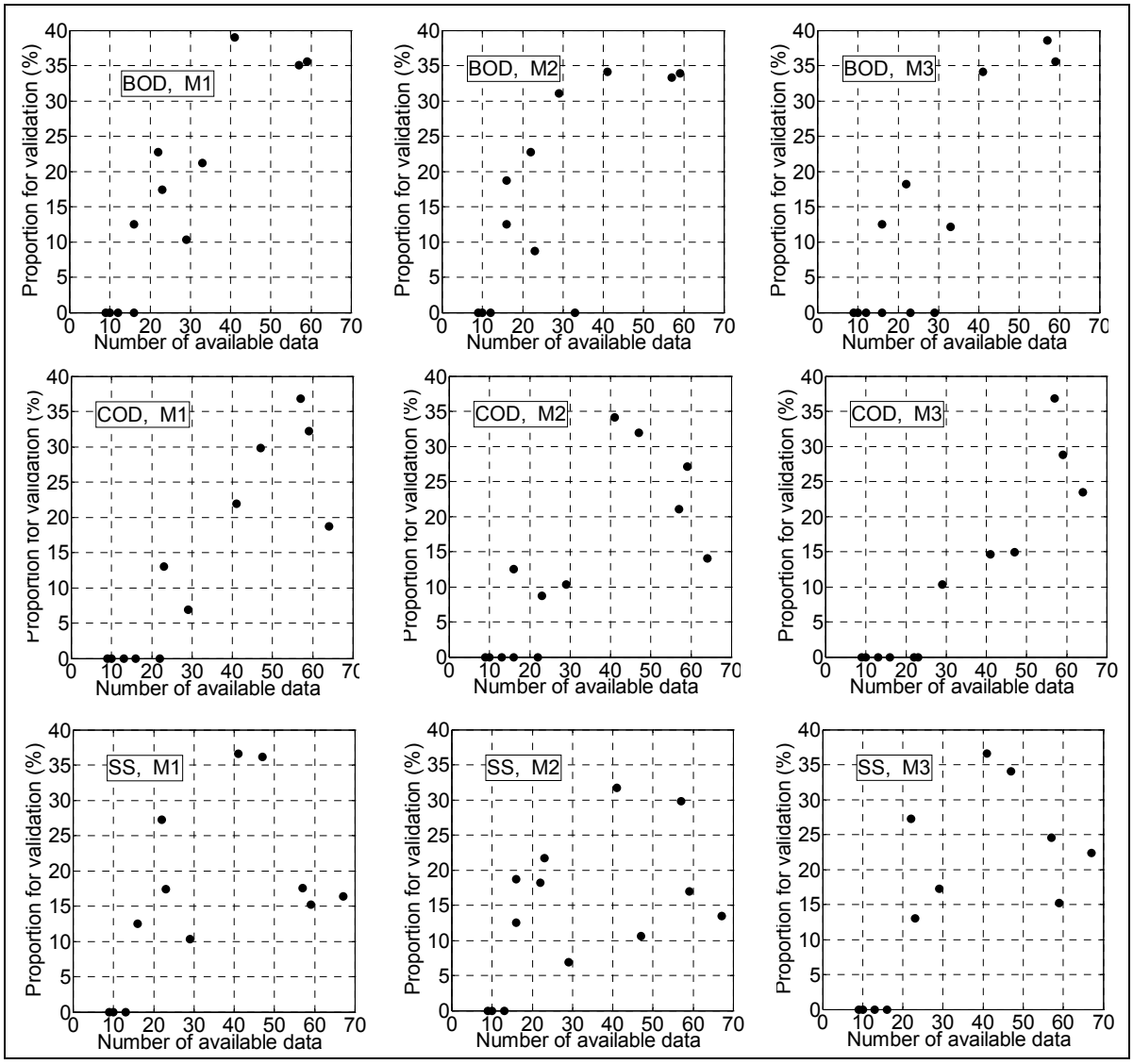

Figure 5 : summary of optimal split results

CONCLUSIONS

In this paper two main issues have been investigated. The first one is the effect of calibration dataset size and characteristics on multiple regression models performance. It was found that using few data (less than 20) for calibration infers important variability on the results. In addition, contrary to lower values of the RMSE for calibration, validation RMSE values can be extremely

high and thus the predictive ability of the model is very poor. Multiple regression models are found to be very sensitive to calibration data. Outliers can easily affect the calibration results.

The second issue is about the optimal split of available dataset into calibration and validation subsets. Results for BOD and COD showed an obvious correlation between the number of available data and the optimal proportion to be used for validation and this for *N* between 10 and 60. For *N* less than 20 it was found that that less than 15 % of the available data are to be used for validation. The main finding in this part is that the usual split of data into halves is not optimal. More data must be allocated for calibration than for validation. Validation proportion did not exceed 40 % in all cases.

A further analysis of the relation between the validation ratio and the noise to signal ratio in model performance, might explain the differences between the scatters of different pollutants and models. Another extension to this work could be the investigation of balanced split sample (McCarthy, 1976) instead of random split. The balanced split sample produces subsets covering approximately the same conditions, which could partly solve some of the above difficulties.

ACKNWLEGMENTS

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

REFERENCES

Chebbo G., Gromaire M.-C., Garnaud S., Gonzalez A. (1999). The Experimental Urban Catchment "Le Marais" in Paris. *Proceedings of the 8th International Conference on Urban Storm Drainage (ICUSD)*, Sydney, Australia, 30 August - 3 September 1999, 1520-1527

Driver N.E., Tasker G.D. (1990). Techniques for estimation of storm-runoff loads, volumes, and selected constituent concentrations in urban watersheds in the United States. U.S. Geological Survey Water-Supply Paper 2363, 44 p.

Hollander M., Wolfe D.A.(1973). *Nonparametric Statistical Methods*. New York (USA) : Wiley, 1973.

INSA/SOGREAH (1999). *CANOE User Manual.* Villeurbanne (France) : ALISON, INSA LYON, 1999.

Jewell T.K., Nunno T.J., Adrian D.D. (1978). Methodology for calibrating stormwater models. *Journal of the Environmental Engineering Division*, 104(3), 485-501.

Kerlinger F.N., Pedhazur E.J. (1973). *Multiple regression in behavioural research*. New York: Holt, Reinhardt and Winston.

McCarthy P.J. (1976). The use of balanced half sample replication cross-validation studies. *Journal of the American Statistical Association*, 71 (September), 596-604.

Saget A. (1994). Base de données sur la qualité des rejets urbains de temps de pluie : distribution de la pollution rejetée, dimensions des ouvrages d’interception. PhD Thesis, Ecole Nationale des Ponts et Chaussées, Paris (France), 333p.

Saget A., Chebbo G. (1996). Qastor: The French Database about the Quality of Urban Wet Weather Discharges. Proceedings of the 7th International Conference on Urban Storm Drainage (ICUSD), Hannover, Germany, 9 – 13 September, vol. 3, 1707-1713.

Schlütter F. (1999). Numerical modelling of sediment transport in combined sewer systems. PhD thesis, Aalborg University, Aalborg, Denmark, 1999, 172p.

Schueler T.R. (1987). Controlling Urban Runoff : A Practical Manual for Planning and Designing Urban BMPs. Publication No. 87703. Metropolitan Washington Council of Governments, Washington, DC., (USA), 275 pp.

Vaze J., Chiew F.H.S. (2003). Comparative evaluation of storm water quality models. *Water Resources Research*, 39(10), 10 p.