Estimating flow data in urban drainage using partial least squares regression

Rita S. Brito, M. Céu Almeida and José S. Matos

ABSTRACT
Flow monitoring in wastewater systems is used for system operation or for billing purposes, among others. Given the difficult measurement conditions, gaps in measurement series occur frequently and stakeholders need an appropriate method to estimate missing data. In data scarcity situations, mathematical modelling of the underlying physical processes may not be feasible and other methods are required. Partial least squares (PLS) regression is a multivariate statistical method suited to correlated data and has been frequently used for water quality estimates. PLS suitability for hourly and daily flow estimations was tested, based on previous flow and precipitation data, which urban water utilities currently monitor. Results were evaluated using proposed performance criteria and classes. The estimation errors were comparable to the ones obtained in physical process modelling. The application of the proposed method for flow estimation in sewers, in two common scenarios of wet and dry weather flows, is presented and discussed.

1. Introduction
Flow monitoring in wastewater systems provides information for multiple uses, including system operation, asset management, regulatory compliance, billing and research purposes. In some of these cases, such as in research projects, only if adequate data quality control is implemented can the results be considered suitable. Monitoring in sewers can be challenging due to flow regime, aggressiveness of the wastewater, or the presence of solids and turbulence, among others (Hager 1994). Data gaps in measurements occur frequently and a procedure to fill in missing data may be required. Such a procedure generally has to be accepted by external entities, if data are used for contractual or compliance purposes.

In dry weather, low flows may contribute to data loss as velocity sensors require a minimum water depth (Larrarte et al. 2008). Nevertheless, in most locations a relatively stable daily dry weather pattern is observed, thus estimating dry weather flow can be quite straightforward. However, rain induced flows (e.g. from storm water connections) impose rapid changes in water depth and velocity. During high intensity rain events, peak flows may increase 10–15 times or even more (Matos 2003) and transport large scale debris. This may contribute to data gaps due to damaged or obstructed equipment.

Modelling processes concerning rainfall and wastewater flows, with a comprehensive description of the phenomena, may not be feasible as a large amount of data may be required (Achleitner 2008). Simpler estimation models might be preferable, as long as accuracy is considered.

The application of partial least squares (PLS) to estimate hydraulic variables is relatively recent (Tootle et al. 2007). Even so, it is considered promising given that a description of the phenomena is not necessary, it may require a reduced number of explanatory variables and it is suitable to correlated data (Abudu 2009).

The purpose of the current work was to test PLS model adequacy for estimating flow data gaps in sewers, based on previous flow and precipitation. Daily and hourly models were developed for several regression orders, and the best model was identified, in order to proceed with flow data estimates and additional tests. Additional tests evaluated the effect of missing data and of data replacement. The benefit of developing specific models dedicated to higher storm water inflows was also evaluated. These evaluations were based on comparisons between measured and estimated flow values.

Model results were classified following model adequacy criteria and performance classes proposed for a sewer specific context.

2. Flow estimation in wastewater systems
2.1. Methods of flow estimation
Analysis and estimation of flow in sewers can be performed through various approaches, based either (i) on the description of physical processes, or (ii) on data and the relationship between variables (Achleitner 2008). While the descriptive approach (i) seeks to explain the phenomena, as detailed as possible, the data-based approach (ii) is based on limited knowledge and uses statistical analysis (Wang 2006). This data-based approach
enables flow estimation, but does not enable studying the system's behaviour or extreme scenarios. The advantages and disadvantages of these two approaches to urban hydrology have been addressed (Veldhuis and Tait 2011).

In situations of limited data availability, the use of complex modelling relying on numerous variables should be avoided (Fischer et al. 2009). In such case, the use of statistical models has become more attractive (Abudu 2009). Among the multivariate regression techniques, PLS regression has been widely used in chemometrics for decades (Kresta 1992), and was recently applied in hydrology (Abudu 2009).

### 2.2. Partial least squares regression

PLS regression finds the underlying relations between the covariates ($X$, explanatory variables considered relevant) and the response variable ($Y$) (Rosipal and Kramer 2006). $X$ and $Y$ are decomposed in accordance with Equations (1) and (2) (Geladi and Kowalski 1986).

$$X_{(n \times q)} = T_{(n \times d)} \cdot L_{(d \times q)}^T + E_{(n \times q)}$$

$$Y_{(n \times m)} = U_{(n \times d)} \cdot R_{(d \times m)}^T + F_{(n \times m)}$$

where,

- $n$: number of samples
- $q / m$: number of variables in the $X / Y$ matrix
- $d$: number of latent variables (LV)
- $T, U$: scores matrices
- $L, R$, $L^T, R^T$: loadings matrices / transpose matrices
- $E, F$: residuals matrices

PLS searches for a set of latent variables (LV) that best explains the relation between $X$ and $Y$ (Figure A-1). It is a guided decomposition method where $Y$ intervenes directly in the decomposition of $X$ (Geladi and Kowalski 1986).

The PLS regression aims to minimize the matrix of residuals from the internal relationship between $X$ and $Y$ (H, in Figure A-1). This method is particularly suited for correlated data, since LVs are a small number of new uncorrelated variables. A reduced number of LV (compared to the number of variables in $X$) indicates model robustness (Van der Voet 1994), meaning that overall trends in measured data can be perceived.

Outliers in a PLS model may be identified (Wise et al. 2006) through the scores plot (scores in the two first LV, Rosen and Lennox 2001), using Hotelling $T^2$ (measure of the variation, in each sample, within the model) or $Q_{residuals}$ (measure of the variation, in each sample, away from the model).

Further details on PLS can be found in the literature (Wold 1966, Geladi and Kowalski 1986, Tobias 1997).

### 2.3. Using PLS in urban drainage

Flow in sewers relate to the hydrological time series in the drainage basin and to urban water uses, including foul flows. Often, these variables are correlated (e.g., precipitation and flow) (Ragavan 2008).

As PLS regression includes correlated variables in the algorithm itself, it can be assumed as a suitable tool for studying and forecasting hydraulic phenomena (Shalamu 2009) and suits correlated hydraulic variables (Abudu 2009). It has been used to predict monthly or seasonal flow in rivers (Tootle et al. 2007, Abudu 2009) or daily evaporation, presenting performances similar to other regression models (multiple least-squares, principal components and artificial neural networks), but was simpler, more explicit to use (Kovoor and Nandagir 2007, Abudu et al. 2011), and provided reduced variance in parameter estimators (Song and Kroll 2011).

Applying PLS to hydraulic estimates is relatively recent and its use for smaller time steps is still not fully explored (Abudu 2009). Nevertheless, it is considered promising (Wang 2006, Kroll and Song 2013) as it doesn’t require much data, it’s quite simple to use and previous work has provided good results.

### 2.4. Model adequacy

Model adequacy evaluates if the best model identified is acceptable for the intended use. For instance, a certain level of accuracy may be adequate for yearly averages but may be inadequate for billing purposes.

Flow estimates rely on measurement accuracy; low accuracy measurements provide less reliable estimates. Understanding measurement accuracy helps to adequately choose the performance classes for flow estimation models.

Sewer’s harsh environment, along with data transmission problems, contribute to data corruption and uncertainty (Ribeiro et al. 2009). A moveable sediment bed (Harmel et al. 2006), water depth or the chosen flow equation (Bertrand-Krajewski et al. 2000a) significantly contribute to uncertainty. Uncertainties from ±50% up to ±100% have been reported in a few case studies in France (Bertrand-Krajewski et al. 2000b). An uncertainty of 10% has been considered acceptable, given appropriate maintenance and calibration (USBR 2001), which is in line with the accepted errors in mathematical modelling (WaPug 2002). Uncertainty propagation in small watersheds has been presented for several scenarios (Harmel et al. 2006). Best case scenario consisted of a high quality control procedure (e.g. with optimal hydrological and measurement conditions). The worst case scenario referred to reduced quality control (e.g. with significant financial restrictions and a moveable sediment bed). Uncertainty for an averaged scenario was in line with Ribeiro et al. (2009) and Bertrand-Krajewsky et al. (2000a).

Reported results are summarized in Table A-1.

Adequacy evaluation may be performed by a variety of statistics. The coefficient of determination, $R^2$, is frequently used and low values clearly show dispersion in the results, but poor models may produce high values of $R^2$ (Legates and McCabe 1999). In this application, the aim is to compare measured and predicted values and a 45° regression line is aimed at. Therefore, $R^2$ will most likely not provide a good evaluation whenever systematic bias (additive or multiplicative deviations) occurs.

It is advisable not to rely solely on correlation indicators. Several efficiency criteria were applied to watershed models, namely RSR (root mean squared error standard deviation ratio of observations), NSE (Nash-Sutcliffe coefficient of efficiency) and PBias (percentage bias). NSE determines the relative magnitude of the residual variance compared to the observed data variance. NSE equal to 1.0 is the best value, whereas NSE minor than or equal to 0.0 means that using the mean observed value is a better
estimate than the predicted values. Both RSR and NSE highlight the dispersion of predicted values related to the dispersion of measures values. None of these criteria performed ideally, so a combination is preferred (Krause et al. 2005). Both NSE and PBias are recommended by ASCE (1993). NSE is largely used in watershed models and is widely reported (e.g. Moriasi et al. 2007). PBias measures the average trend of the predicted values being higher or lower than the corresponding measured values. The ideal PBias is 0. PBias clearly indicates a poor model, when predicted values are mostly over or underestimated (Gupta et al. 1999). On the other hand, a poor model with dispersion in the results may perform well in PBias whenever the average trend is not either over or underestimated. The inclusion of graphical information is also recommended (Legates and McCabe 1999, Moriasi et al. 2007).

The equations for the referred criteria are the following:

\[
\text{RMSE} = \sqrt{\frac{\sum (O_i - E_i)^2}{n - 1}}
\]

\[
\text{RMSE}_{\text{rel}} = \frac{\text{RMSE}}{\bar{O}} \times 100
\]

\[
\text{RSR} = \frac{\text{RMSE}}{\bar{O}} = \sqrt{\frac{\sum (O_i - E_i)^2}{\sum (O_i - \bar{O})^2}}
\]

\[
\text{NSE} = 1 - \frac{\sum (O_i - E_i)^2}{\sum (O_i - \bar{O})^2}
\]

where:
- \( n \): number of samples in the set
- \( O_i \): observed value for sample \( i \)
- \( E_i \): estimated value for sample \( i \)
- \( \bar{O} \): average observed values for samples in the set
- \( S_O \): standard deviation of the observed values

Such criteria must be rated and, as already mentioned, performance ratings ought to be related to confidence in measured data. Values presented in Table A-1 could guide the choice of performance ratings for RMSE_{rel}, which determines the dimensionless magnitude of the error. Moriasi et al. (2007) established guidelines for small watershed model evaluation (namely for RSR, NSE and PBias) based on a thorough review of literature, classifying criteria from Unsatisfactory to Very Good.

Models were based on detailed basin descriptions and were developed for monthly and daily time steps, with less accurate estimates being expected for shorter time steps. Such a classification has not yet been applied to drainage systems, but given the similarities with small watersheds, it might be applicable.

### 3. Methodology

#### 3.1. Overview

The proposed methodology (Figure 1) consists of five phases: (i) data selection and pre-processing; (ii) model identification; (iii) additional tests; (iv) model adequacy evaluation and (v) final analysis. Detailed explanation is given in Sections 3.2–3.5.
3.2. Data selection and pre-processing

PLS models were intended to estimate dry or wet weather flow at time $t_i$ ($Q_i$ in matrix $Y$), using a small number of covariates. Drainage basin response to rain events has also been modelled and values included in matrix $X$: previous values for flow ($Q_{t-1}, \ldots, Q_t$) together with precipitation. In the later, previous values as well as values at $t_i$ ($P_{t-1-6}, \ldots, P_t$) were included. Index $p$ stands for the model’s regression order. Only periods with simultaneous data both from the flow meter and the rain gauge were studied.

Precipitation varies significantly during rain events; the smaller the time step, the more accurately the phenomena is represented. The data time step in the model should be small enough to enable adequate estimates, but not so small that model robustness is compromised. Preliminary tests were carried out using 5 minute, hourly and daily data. A large number of LV was required in the 5 minute PLS models, indicating lack of model robustness (Van der Voet 1994). Therefore, model identification proceeded for hourly and daily data.

Data in X and Y were mean centred (Wise et al. 2006), i.e. the arithmetic mean was subtracted from all the values, to account for the scale effect between precipitation and flow. Preliminary tests using auto-scaling (i.e. subtracting the mean and dividing by standard deviation, Wise et al. 2006) were made on one flow meter and results were similar.

3.3. Model identification

PLS models were developed in MATLAB ® 7.4.0 (The Mathworks Inc., Natick, Massachusetts, U.S.A.) using the PLS toolbox 3.0 package (Eigenvector Research Inc., Wenatchee, Washington, U.S.A.), expecting to determine how many antecedent days/hours were required to obtain lower errors in flow predictions (optimal $p$).

Several regression orders $p$ were tested. For daily models, $p$ varied between 1 and 5 (i.e., previous 1 up to 5 days) and for hourly models, between 1 and 6 (i.e., 1 up to 6 hours). Preliminary tests were carried out on the regression order. These preliminary tests were based on Phase 1 and Phase 2 of the methodology (Figure 1), up to the determination of RMSECV (the root mean squared error applied to the cross validation set), and were applied to every flow meter under study. Tests showed that higher $p$ did not provide lower errors.

Available data for each flow meter were divided into two sets (training and test sets). The training set was used for model identification, consisting of calibration and validation. Several calibration and validation procedures are available in the PLS toolbox; a full cross-validation iterative procedure by contiguous blocks was adopted (Wise et al. 2006). In each iteration, a different subset of the training set is used for validation. At the end of model identification, every element in the training set was used for calibration and once for validation. The error of each run was calculated by full cross-validation iterative procedure by contiguous blocks was adopted (Wise et al. 2006). In each iteration, a different subset of the training set was used for validation. At the end of model identification, every element in the training set was used for calibration and once for validation. The error of each run was calculated by
data, they were kept in the model; otherwise, they were removed, as they could distort the model.

The predictive ability of the models was tested using samples that were not used in model development, i.e., the test sets. The model with optimal $p$ was identified based on lower RMSEP rel (RMSE rel applied to the predictions on the test set, according to Equation (4)) and higher $R^2$.

3.4. Additional tests

After model identification, other aspects were evaluated:

• Data omission: data in the test set may not be available in all previous $p$ corresponding moments; to evaluate the effect of data gaps, flow data was deleted from the test sets, for 1 previous day/hour, 2 previous days/hours, etc., up until $p-1$; predictions were made for these edited test sets, with data gaps of 1 up to $p-1$ values.

• Data replacement: data omissions can be replaced with estimated data. Predictions can be made for the edited test sets. To evaluate this procedure, data gaps in the data omission tests were replaced with estimates, using the following rules:

  (i) if only 1 antecedent day/hour was missing, it was replaced with its corresponding prediction from the model identified in Section 3.3;

  (ii) if 2 antecedent days/hours were missing: for the 1st antecedent value, (i) was applied; the 2nd antecedent value was replaced with its prediction from Data Omission models (with a data gap of 1 value);

  (iii) if the data gap had more than 2 antecedent days/hours missing ($n$ values, up to $p-1$), the procedure in (ii) was followed using, for the $n$th replacement, the Data Omission model with a data gap of $n-1$ values.

• Storm water inflows: when PLS models include a full year of data, models can be constrained by dry weather records (more frequent and with lower variability); to evaluate whether specific models for higher flows provided better estimates during rain events, PLS models were developed for an edited training set (high flow data in the training set) and predictions were made for edited test sets (high flow data in the test sets). High flow data records due to rain events were previously identified (Brito 2012) based on the following sequential criteria: (i) after a rain event, 5 days of flow data were kept (ii) dry weather flow patterns were identified and only flow records higher than such pattern were kept; (iii) operation and maintenance registers were analysed; high flows due to such procedures were removed. Various regression orders $p$ were tested. Data omission and data replacement tests were also performed.

3.5. Model adequacy evaluation

Model adequacy evaluation was carried out using RMSEP rel, RSR, NSE and PBias (Equations (4)–(7)). Given the recommendations in Section 2.4, the performance classes in Table 1 were used.

Classes proposed in Table 1 were mostly proposed in literature for contexts other than urban drainage. Even though some of
Table 1. Selected performance criteria for model adequacy classification.

| Performance classes | RMSEP<sub>rel</sub><sup>a</sup> | RSR<sup>b</sup> | NSE<sup>b</sup> | |Bias|<sup>b</sup> |
|---------------------|----------------|-------------|-------------|----------------|-------------|
| Very good           | <5%            | 0.0–0.5     | 75–100%     | <10%           |
| Good                | 5–10%          | 0.5–0.6     | 65–75%      | 10–15%         |
| Satisfactory        | 10–20%         | 0.6–0.7     | 50–65%      | 15–25%         |
| Unsatisfactory      | ≥20%           | ≥0.7        | ≤50%        | ≥25%           |

<sup>a</sup> Proposed by the authors.

<sup>b</sup>Moriasi et al. (2007).

the proposed criteria have a similar insight on data, they were all calculated for model adequacy evaluation, in order to allow a better understanding of their application to urban drainage. Only classes for RMSEP<sub>rel</sub> were not available in literature for this specific context. The authors proposed the limits in Table 1 based on literature review (see Section 2.4). Since it is aimed to evaluate whether the proposed classes are realistic, the results on RMSEP<sub>rel</sub> were compared with the results from other criteria.

This classification was applied to model identification and to the additional tests on data omission and data replacement.

4. Experimental data

The case study is located in Portugal. The Costa do Estoril wastewater system serves circa 800,000 inhabitant equivalent and has a flow monitoring network operating since 1998, with 5 rain gauges and more than 50 flow meters, installed in various hydraulic conditions. It was designed as a separate wastewater system, but storm water inflows are significant. SANEST, S.A, the utility manager, uses flow data for asset management since 2001 and for billing purposes since 2005. Asset management recurs to 5 min, hourly, daily, monthly and annual data, depending on the decision level (operational, tactical or strategic), and billing relies on daily and monthly data.

Three flow meters were selected for this study (Q41, Q03 and Q08), considering their different sewer diameters (680–1790 mm), upstream contributing areas (20–107 km²), concentration times (2–8 h) and magnitude of flows (70–3500 L/s).

Q03 was selected because of its good overall performance (concerning an evaluation of hydraulic variables and maintenance interventions) even though infiltration and storm water inflow occur (Cardoso 2008). Q41 is a more problematic location, due to structural problems, low dry weather flows and hydraulic problems upstream, contributing to increased maintenance (Brito 2003). The upstream catchment area in Q41 is mostly urban. Q08 presents a larger cross section and drainage basin and higher dry weather flows, and is very important for billing. The influence of rain events on flows lasts longer than in other measurement locations. Maintenance interventions are rare in Q08. Q03 and Q41, being nearer to laterals, show higher flow variability.

Q08, located further downstream in the system, presents smoother hydrographs.

Data series have a 5 minute time step. Flow and precipitation data from the years 2006, 2007 and 2008 were used in this study. Data from 2008 integrated the training set, the year with the largest amount of data; two test sets were used, data from 2006 and 2007.

Figure A-2 illustrates daily precipitation and flow in Q08, where rain induced flows are evident and long lasting.

Rain events presented diverse characteristics, such as event duration (15min - 6h), total rainfall (1–99 mm), average intensity (3–30 mm/h), maximum intensity (6–60 mm/h) and previous dry period (4–9 days).

5. Results And Discussion

5.1. Model identification

In PLS models, each matrix had up to 360/8800 lines (daily/ hourly models). The various models presented 0–5% of outliers (rejected, as clusters were not observed). Most models required 2–4 LV, revealing model robustness. LVs express the overall trends in data and are represented by linear equations using variables in matrix X. Flow data variables contributed more to the LV equations, when compared to rainfall data variables, given the associated coefficients’ weight.

Figure 2 represents daily measured and estimated flow for 2006, with p = 4 days, in Q03. Figure A-3 presents a similar example for an hourly model.

Although some dispersion is visually noticeable in Figure 2a, estimated data almost overlapped measured data in Figure 2b. The goodness of fit was confirmed by the overall RMSEP<sub>rel</sub> for the three flow meters, for 2006 and 2007, and for each of the tested p. Lower RMSEP<sub>rel</sub>, indicating the optimal p, varied mostly between 5% and 10%. In daily models, in most cases the error decreased slightly with increasing p.

Lower RMSEP<sub>rel</sub> was obtained when previous 3–4 days of data were used and R² was higher than 0.90. Using more than 4 days did not improve the results (see Table A-2).

In hourly models, RMSEP<sub>rel</sub> values varied in a similar way. Lower RMSEP<sub>rel</sub> was obtained for 2 or 3 hours of data (9–13%) and R² was higher than 0.95. Using more than 4 hours of data did not improve the results (see Table A-3).

In 2006 daily models, site Q41 was clearly an exception. Possibly due to its location in upstream areas, this site is subject to the direct effect of user behaviour. Like water consumption patterns (Alegre et al. 1993), social behaviour can be explanatory variables, but such data were not available for these PLS models. In contrast, downstream sites show smoother hydrographs, given the attenuation due to hydraulic transport in the sewers. These factors can influence the forecasting ability for lower flows.

The time of concentration (t<sub>c</sub>) of the catchments was only a few hours, while the optimal daily p was 3–4 days. Apparently, surface runoff was not the only explanatory variable for daily p. The indirect rain-induced flow lasted from several hours up to a few days (Figure A-2). The average velocity of this inflow was very low (0.02–0.06 m/s), indicating another factor contributing to optimal daily p: infiltration into sewers following rainfall, due to rising groundwater level, through pipes in poor structural condition. Although not explicitly incorporated, these processes were included in the models, as previous daily flows were covariates.

Some relation was expected between the hourly optimal p and the catchment's t<sub>c</sub>. In daily models, the optimal p was of several days. However, for hourly models, 2–3 hours were enough, even for larger catchments. In fact, catchment response was integrated in the models as a base flow (as previous hourly flows), resulting in hourly optimal p lower than t<sub>c</sub>.
Overall, PLS hourly models provided acceptable results. PLS daily models can be considered acceptable for most sites; however, sites with lower flows may require more robust modelling.

5.2. Additional tests

5.2.1. Data omission test

PLS models were tested to evaluate the effect of data omission. In daily models, as expected, lower RMSEP$_{\text{rel}}$ occurred for smaller data gaps. Models responded quite well to 1–2 days of missing data (RMSEP$_{\text{rel}}$ 9—17%), excepting Q41 (RMSEP$_{\text{rel}}$ over 25%). For more than 2 missing days, $R^2$ were under 0.80. Graphically, results scattered when data was missing for longer periods (see Figure A-4).

Hourly models were not acceptable. RMSEP$_{\text{rel}}$ were mostly above 80%, which is unsatisfactory. Hourly flows usually present greater variability than daily flows therefore the values immediately before are extremely relevant. The omission of antecedent flow was responsible for the poor results in PLS hourly models. Overall, daily models could be applied for data gaps smaller than 2 days, but hourly models were not recommended, and will not be evaluated in section 5.3.

Whenever antecedent hourly data gaps occur, it is preferable to opt for an overall daily estimate.

5.2.2. Data replacement test

Tests on the effect of replacing data gaps with flow estimates were also implemented. Data replacement clearly improved results. The scattered results referred to in Section 5.2.1, when more previous data was missing, was overcome; by replacing such data, flow estimates were more similar to the measured values, and to each other (see Figure A-5).

In daily models, RMSEP$_{\text{rel}}$ was quite stable with an increased number of days, which did not occur in data omission tests. Generally, $R^2$ was above 0.80 and data omission tests had smaller errors than data replacement tests for gaps of 1 day. RMSEP$_{\text{rel}}$ varied mainly between 9-14%, but was above 20% for Q41.

In hourly models, data replacement improved results significantly. RMSEP$_{\text{rel}}$ varied between 10—20% for up to 2h of missing data, and were quite close to 20% for 3h or more, which did not occur in data omission tests. In summary, PLS models improved with data replacement, for data gaps above 1 day in daily models, and for all the replacement tests in hourly models.

5.2.3. Storm water inflow model

Storm water inflow model results were very similar to the aforementioned. In daily models, using $p$ of 3–4 days provided the lowest errors. For tests on data omission and data replacement, RMSEP$_{\text{rel}}$ were similar, with small variations from ±0.1–3%. The benefit of studying storm water inflows separately was not evident.

In hourly models, using $p$ of 2–3 hours provided the lowest errors. Tests on data omission presented unsatisfactory results. Most tests on data replacement presented RMSEP$_{\text{rel}}$ slightly lower than the ones presented in Section 5.2.2 (less 0.2-1.6%). The benefit of studying storm water separately was small, but given the limits in Table 1, such a difference enable satisfactory from unsatisfactory results to be distinguished. Hourly models might benefit from a separate wet weather model, because during this selected periods, the models make full use of rainfall data and are not constrained by dry weather records.

5.3. Model adequacy classification

Model adequacy was assessed using RMSEP$_{\text{rel}}$, RSR, NSE and PBias. These criteria were calculated for each model and classified from Unsatisfactory to Very Good (Table 1). All cases were classified (three sites, for 2006 or 2007, for various values of $p$) for each test (model identification, data omission and data replacement). Box-and-whiskers plots for all cases are presented in Figure 3, for an overall evaluation.

Most daily models presented larger whiskers due to Q41, which delivered less satisfactory results. In hourly models, data omission tests were not graded, as they were considered unacceptable. Nevertheless, most tests presented either Very Good or Good results, with the exception of RMSEP$_{\text{rel}}$ where most models were classified as Good or Satisfactory. These results may derive from the RMSEP$_{\text{rel}}$ classes proposed herein by the authors and not tested previously. It is possible that RMSEP$_{\text{rel}}$ ratings could be less demanding, taken into account the measurement errors expected in urban drainage (Table A-1). RMSEP$_{\text{rel}}$ Classifications may also be compared to NSE’s. RMSEP$_{\text{rel}}$ provides information for data gaps above 1 day in daily models, and for all the replacement tests in hourly models.
6. Conclusion

In this study, the use of PLS models to estimate daily and hourly flow in sewers was evaluated as an alternative to physical process modelling. A reduced number of LV was a sign of model robustness.

Daily PLS models presented lower errors using previous 3–4 days of data.

Separate daily models for wet weather were shown not to be justifiable. Errors increased slightly when previous data was missing, but for two or more days of missing data, models benefited from their replacement with estimated data.

Hourly PLS models presented lower errors using previous 2–3 hours of data. There was a small advantage in identifying separate hourly models for wet weather. When previous hourly flows were missing, models were considered inadequate.
However, results clearly benefited from their replacement with estimated data.

The proposed adequacy evaluation identified the models that were good enough for this specific context (flow estimation in sewers). Models were mostly classified as Satisfactory or Good, and some were Very Good. However, sites with lower flows may require more robust modelling.

Overall, the proposed methodology can be a useful tool for utilities, for both wet and dry weather. Requiring only previous data regarding flow and precipitation, most hydrological explanatory phenomena were represented by these variables. PLS models provided errors similar to those recommended for flow measurement, leading to estimates with a level of accuracy compatible with the intended uses.

Acknowledgement

The authors acknowledge SANEST, for providing access to data.

Funding

The first author acknowledges the financial support from FCT (Portugal) through a doctoral (SFRH/BD/46665/2008) research grant.

References

Abudu, S., 2009. Monthly and seasonal streamflow forecasting in the Rio Grande basin. Thesis (PhD). New Mexico State University.

Abudu, S., et al., 2011. Modeling of daily pan evaporation using partial least squares regression. Science China Technological Sciences, 54 (1), 163–174.

Achleitner, S. 2008. Modular conceptual modeling in urban drainage development and application of city drain. Thesis (PhD). Universität Innsbruck.

Alegre, H. and Coelho, S.T., 1993. Integrated computer applications in water supply. In: A methodology for the characterisation of water consumption, Vol. 1, New York, NY: John Wiley & Sons, 369–384.

ASCE, 1993. Criteria for evaluation of watershed models. Journal of Irrigation and Drainage Engineering, 119 (3), 429–442.

Bertrand-Krajewski, J.L., et al., 2000a. Mesures en hydrologie urbaine et assainissement. Paris: Editions Technique & Documentation. Bertrand-Krajewski, J.L., Barrau, S., and Chocat, B., 2000b. Need for improved methodologies and measurements for sustainable management of urban water systems. Environmental Impact Assessment Review, 20, 323–331.

Brito, R.S., 2003. Avaliação Integrada do Comportamento de Sistemas Pseudo-Separativos de Águas Residuais Domésticas. Thesis (Master). Universidade Técnica de Lisboa.

Brito, R.S., 2012. Monitorização de variáveis hidráulicas e de qualidade da água em drenagem urbana. Thesis (PhD). Universidade Técnica de Lisboa.

Cardoso, M.A., 2008. Avaliação de Desempenho de Sistemas de Drenagem Urbana. Thesis (PhD). Universidade Técnica de Lisboa.

Fischer, A., et al., 2009. Possibilities of sewer model simplifications. Urban Water Journal, 6 (6), 457–470.

Geladi, P. and Kowalski, B., 1986. Partial least square regression. A tutorial. Analytica Chemica Acta, 35, 1–17.

Gupta, H.V., Sorosohian, S., and Yapo, P.O., 1999. Status of automatic calibration for hydrologic models. Comparison with multilevel expert calibration. Journal of Hydrologic Engineering, 4 (2), 135–143.

Hager, W., 1994. Wastewater hydraulics - Theory and practice. Berlin: Springer–Verlag.

Harmel, R.D., et al., 2006. Cumulative uncertainty in measured streamflow and water quality data for small watersheds. Transactions of the ASABE, 49 (3), 689–701.

Kovoor, G. and Nandagir, L., 2007. Developing regression models for predicting pan evaporation from climatic data — a comparison of multiple least-squares, principal components, and partial least-squares approaches. Journal of Irrigation and Drainage Engineering, 133 (5), 444–454.

Krause, P., Boyle, D.P., and Base, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Advances in Geosciences, 5, 89–97.

Kresta, J., 1992. Application of Partial Least Squares to problems in chemical engineering. Thesis (PhD). McMaster University, Canada.

Kroll, C.N. and Song, P., 2013. Impact of multicollinearity on small sample hydrologic regression models. Water Resources Research, 49, 3756–3769.

Larrarte, F., et al., 2008. Flow measurement and instrumentation acoustic doppler flow-meters. A proposal to characterize their technical parameters. Flow Measurement and Instrumentation, 19, 261–267.

Legates, D.R. and McCabe, G.J., 1999. Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation. Water Resources, 35 (1), 233–241.

Matos, J.S., 2003. Ambiente e Saneamento - Sistemas de Drenagem Urbana. Lisbon, Portugal: IST Press.

Moriasi, D.N., et al., 2007. Model evaluations guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE, 50 (3), 885–900.

Nash, J.E. and Sutcliffe, J.V., 1970. River flow forecasting through conceptual models. Part 1. A discussion of principles. Journal of Hydrology, 10 (3), 282–290.

Ragavan, A., 2008. Comparing efficiency of cross validated partial least squares regression and ridge regression models in predicting spatial variability of total nitrogen concentration in Truckee River, Nevada. Nevada: WUSS – Western Users of SAS Software.

Ribeiro, A., Almeida, M.C., and Palma, J., 2009. Uncertainty evaluation of multi-sensor flow measurement in a sewer system using Monte Carlo method. XIX IMEKO World Congress. Fundamental and Applied Metrology, 6–11 September 2009 Lisbon. 1287–1292. Available from: www.researchgate.net

Rosen, C. and Lennox, J.A., 2001. Multivariate and multiscale monitoring of wastewater treatment operation. Water Research, 35, 3402–3410.

Rosipal, R. and Kramer, N., 2006. Overview and recent advances in Partial Least Squares. In: Saunders et al., eds. Subspace, latent structure and feature selection, Lecture Notes in Computer Science Statistical and Optimization, 3940, Berlin: Springer Berlin Heidelberg, Vol 3940, 34–51.

Shalamu, A., 2009. Monthly and seasonal streamflow forecasting in the Rio Grande basin. Thesis (PhD). New Mexico State University.

Song, P. and Kroll, C., 2011. The impact of multicollinearity on small sample hydrologic regional regression. World Environmental and Water Resources Congress, 2011, 3713–3722.

Tobias, R.D., 1997. An introduction to Partial Least Squares Regression. Institute for Digital Research and Education (online). Available from: http://ftp.sas.com/techsup/download/technote/ts509.pdf [Accessed 2 October 2010].

Tootle, G.A., et al., 2007. Long lead-time forecasting of U.S. streamflow using Partial Least Squares Regression. Journal of Hydrologic Engineering, Sept/ Oct, 442–451.

USBR, 2001. Water measurement manual. Washington, DC: U.S. Government Printing Office.

Van der Voet, H., 1994. Comparing the predictive accuracy of models using a simple randomization test. Chemometrics and Intelligent Laboratory Systems, 25, 313–323.

Veldhuis, J. and Tait, S., 2011. Data-driven urban drainage analysis: an alternative to hydrodynamic models? 12th International Conference on Urban Drainage, 11–16 September Porto Alegre: International Water Association, 1–9. Available from: http://repository.tudelft.nl

Wang, W., 2006. Stochasticity, nonlinearity and forecasting of streamflow processes. Amsterdam: IOS Press.

WaPug, 2002. Code of practice for the hydraulic modelling of sewer systems. Version 3.001, UK: Wastewater Planning Users Group. Available from: http://www.ciwem.org.uk/

Wise, B.M., et al., 2006. PLS_Toolbox 4.0 for use with MATLAB™. USA: Eigenvector Research. Available from: http://www.eigenvector.com

Wold, H., 1966. Estimation of principal components and related models by iterative least squares. In: P.R. Krishnaiah, ed. Multivariate Analysis, Amsterdam: North-Holland Publishing Company, 391–420.