On-line hydraulic state prediction for water distribution systems

Ami Preis[1], Andrew Whittle [2] and Avi Ostfeld [3]

[1] Postdoctoral Associate, Center for Environmental Sensing and Modeling, MIT-SMART Center, Singapore;
amipreis@smart.mit.edu
[2] Professor, Department of Civil and Environmental Engineering, MIT, Cambridge, MA, USA;
ajwhittl@MIT.EDU
[3] Senior Lecturer, Faculty of Civil and Environmental Engineering, Technion – I.I.T, Haifa, Israel;
ostfeld@tx.technion.ac.il

Abstract

This paper describes and demonstrates a method for on-line hydraulic state prediction in urban water networks. The proposed method uses a Predictor-Corrector (PC) approach in which a statistical data-driven algorithm is applied to estimate future water demands, while near real-time field measurements are used to correct (i.e., calibrate) these predicted values on-line. The calibration problem is solved using a modified Least Squares (LS) fit method. The objective function is the minimization of the least-squares of the differences between predicted and measured hydraulic parameters (i.e., pressure and flow rates at several system locations), with the decision variables being the consumers' water demands. The a-priori estimation (i.e., prediction) of the values of the decision variables, which improves through experience, facilitates a better convergence of the calibration model and provides adequate information on the system's hydraulic state for real time optimization. The proposed methodology is demonstrated on a prototypical municipal water distribution system.

1. Introduction

The integration of near real-time hydraulic data with computer simulations for on-line operation and control of large-scale urban water distribution systems can be used in a variety of applications ranging from real-time optimization of pump and valve settings for efficient power management; to the detection and quantification of leaks. Such a system can also be used for the implementation of water security systems and for the prediction of system performance during emergency events (e.g., pollution events, main pipe rupture, or significant fire).

Over the last 10 years, computer simulation models of water networks have been widely used by water systems operators (Walski et al. 2003). In conventional practice, model calibration is carried out off-line (USEPA 2005) using a short-term (e.g., one week) sample of hydraulic data (e.g., based on a limited set of flow rate and pressure measurements within the network). Thereafter, uncertain system parameters (e.g., water demands and pipe roughness) are adjusted until an acceptable match is achieved between the model outputs and physical observations. Ormsbee (1989) and Lansey and Basnet (1991) were among the first to develop formal optimization methods for determining the uncertain system elements. Datta and Sridharan (1994), and Reddy et al. (1996) used the regression approach in which parameter uncertainties were estimated as part of the calibration process. Greco and Del Giudice (1999) used a "sensitivity matrix" to minimize the least squares differences between observed and predicted values, and Lingireddy and Ormsbee (1998) developed a calibration method using Artificial Neural Networks (ANN). More recently, Kapelan et al. (2007) used the shuffled complex evolution metropolis (SCEM-UA) global optimization algorithm to solve a least-squares-type calibration problem where both calibration parameter values and associated uncertainties were considered in a single optimization model run.

The current research uses a Genetic Algorithm approach following methods developed by Savic and Walters (1995); Wu and Simpson (2001); Kapelan et al. (2002); Wu et al. (2002); Walski et al. (2006); and Clark and Wu (2006).

Overall, the main limitation of all off-line calibration procedures is that they approximate the unknown parameters using a short-term sample of hydraulic data. The calibration results may represent the system hydraulics during the short period of the sampling procedure but they are not expected to accurately represent the system conditions for the full range of operational conditions that can occur. In the case of water demands this issue is even more critical, since water demands have dynamic/stochastic pattern variations which fluctuate with time-changing economic and demographic characteristics and may even show trends with local climatic conditions *(*Maidment and Miaou 1986; Kenward and Howard 1999; Zhou et al. 2000*)*. In principle, much more realistic predictions can be achieved by updating the hydraulic state-

estimation using continuous on-line hydraulic measurements, provided by a sensor network installed on the distribution system.

There have been several recent studies that have assimilated on-line measurements into hydraulic state estimation models. Davidson and Bouchart (2006) propose proportional and target demand methods. These are two techniques for adjusting estimated demands in hydraulic models of water distribution networks to produce solutions that are consistent with available Supervisory Control and Data Acquisition (SCADA) data. The two techniques assume that pipe resistances and SCADA data are accurate and that the combination of SCADA data and demand estimates produce over-determined problems. Nodal demands are regarded as stochastic variables which fluctuate about an estimated mean value. The method of weighted least squares is used to obtain solutions that satisfy all of the constraints imposed by SCADA data with adjusted nodal demands that most closely resemble the estimates. The methods are intended for use in real-time modeling but are limited to quasi-steady state flow.

Shang et al. (2006) presented a predictor-corrector method, implemented in an extended Kalman filter to estimate water demands within distribution systems in real-time. A time-series ARIMA model is used to predict the water demands based on the estimated demands at previous steps. The predictions are corrected using measured nodal water heads or pipe flow rates. As noted by the authors, the proposed methodology is in a preliminary stage and aimed mainly to study the impact of spatial correlation between demand forecast errors on demand estimations. The methodology is demonstrated on EPANET example 3, having 59 demand nodes, through three simulation studies with 20 pressure, 20 flow rate, and 40 flow rate sensors. The main conclusion was that the model performances depend on sampling design, measurement uncertainty, demand forecast error and the spatial correlations among the demand forecast errors. Although only preliminary results were presented, the study provided a modeling framework and mathematical tools for further implementations on more complex case studies.

In this study, a Predictor-Corrector (PC) approach which integrates a limited number of continuous hydraulic observations with a computer simulation model is implemented to continually predict the hydraulic state of a real urban water supply comprised of 10550 demand

nodes. The *M5 Model-Trees algorithm* (Quinlan 1992) is used to forecast future water demands for a rolling planning horizon of 24 h ahead, and *Genetic Algorithms* (Holland 1975) are used to correct (i.e., calibrate) these predicted values in real-time. Thereafter, at each subsequent time step, the corrected outputs of previous iterations are used as inputs for the prediction model. This a-priori estimation of the calibration parameters values, which improves through experience, facilitates a better and quicker convergence of the calibration procedure towards the optimal solution of the problem and provides adequate information on the system's hydraulic state for real time operation and control.

2. Methodology

Model definitions and assumptions

1. Calibration Parameters: The calibration parameters used in this study are the variations in water demands (i.e., defined as demand multiplication factors; see next paragraph for a full description of this parameter). The demand multiplication factors (DMFs) are calibrated at each time step of the overall process. Other uncertain variables are less dynamic and their values are assumed to be constant for a certain period of time. It is assumed that valve and/or pump settings are known inputs, while pipe roughness coefficients are calibrated off-line using conventional procedures (with complete supplementary information on pipe diameter, material type, and age).

2. Demand Multiplication Factors (DMFs): The baseline demand (D_{base}) of each consumption node is a deterministic value that usually equals to the average daily demand (i.e., average daily demand is calculated from monthly or quarterly meter readings and billing records). The patterns in demand on a finer time scale are described by Demand Multiplication Factors (DMFs). For each short-term time step in the demand pattern, the relevant DMFs are multiplied with the baseline demands of the consumption nodes to obtain the actual water consumption (i.e., $D_t = D_{base} \times DMF_t$; where D_t is the actual nodal demand at time step t and DMF_t is the demand multiplication factor at the same time step). It is assumed that the min-max DMFs boundaries are 0 and 3, respectively; previous publications (Walski et al. 2003; Jonkergouw et al. 2008) have shown that these min-max boundaries provide acceptable estimates for hourly basis demand multiplication factors.

3. Calibration parameters grouping: There are thousands of unknown parameters in a typical urban water system and only a relatively small number of direct measurements available. This creates an ill-posed, underdetermined problem which leads to non-unique solutions. This can be overcome by grouping the unknown parameters. Grouping is based on the assumption that water customers in a given area of the system will have the same characteristics and will not need large adjustments to achieve calibration (Wu et al. 2002; Walski et al. 2006). The main advantage of 'grouping' is that the size of the problem is reduced - making it possible to find unique solutions to the optimization problem. In this study, the consumption nodes were grouped based on spatial analysis of the system and each group of consumption nodes is assigned its own set of demand multiplication factors.

4. Time step: the method developed in this study is general and can be adjusted to a wide range of time-step intervals so that even frequent continuous sensor measurements can be used in the model. The current analysis uses a fixed one hour time step in order to match the frequency of available hydraulic data.

5. Time cycle: The common approach (Shang et al. 2006, Alvisi et al. 2007, and Ghiassi et al. 2008) for the time-series forecasting of water demands relies on direct identification of patterns existing in the archived system data. It has been observed that water demand patterns usually follow a 24-h cycle. This cycle is called the Diurnal Demand Pattern (DDP) and is used by many urban water utilities to plan the system operation one day ahead (i.e., to schedule pump operation and plan tank storage). Weekend demand patterns often differ from weekday patterns and usually follow a 168-h (1 week) cycle. In this study, daily and weekly demand cycles were used, so that the DMFs at each time-step are predicted based on previously calibrated DMFs from past hours 24, 25, 168, and 169.

Predictor-Corrector Model

The predictor-corrector loop process starts at t=169 hr, after performing an off-line calibration procedure for the first 168 h (1 week) of the collected data; the aim of this off-line calculation is to generate initial values for the input data-set of the prediction model; no a-priori information on the first 168 h DMFs values is available except of the min-max boundaries which are 0 and 3,

respectively. At t=169 h the prediction-correction process (Fig. 1) is initiated. The following paragraphs describe the model steps:

1. DMF prediction: DMFs values of each group of demand nodes are predicted using the *M5 Model Trees algorithm*, with the inputs being the calibrated DMFs from past hours t-24, t-25, t-168, and t-169

2. EPANET simulation: the system hydraulics are simulated using the steady-state mode of EPANET, with the predicted DMFs as inputs; the hydraulic simulation outputs are nodal pressures and pipe flow rates

3. On-line hydraulic data integration: pressure and flow measurements (from a set of in-line sensors) are inserted to the model at the current time step

4. DMF correction/calibration: a calibration problem is formulated and solved using *Genetic Algorithms*. The objective function is the minimization of the differences between predicted and measured hydraulic parameters (i.e., pressure and flow rates at the measured locations), with the decision variables being the consumers' water demands (i.e., the Demand Multiplication Factors - DMFs). A modified Least Squares fit method (the Huber function) which takes into account noisy measurements is implemented to solve the optimization problem

5. DMF delay: the calibrated DMFs are being delayed for 24, 25, 168, and 169 h before being used as inputs in the prediction model

- Steps 1 to 5 are repeated at each subsequent time step

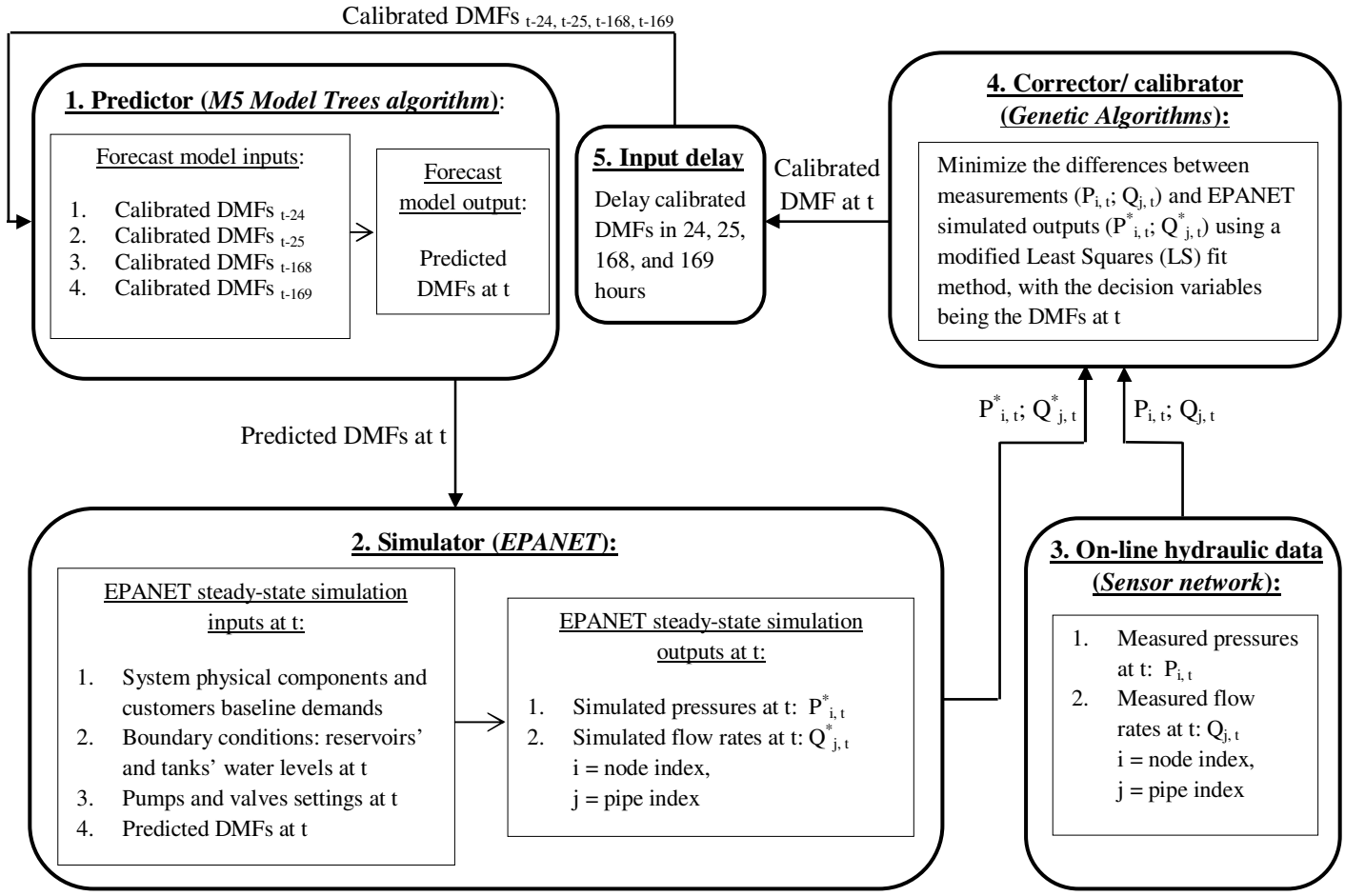

Fig. 1: predictor- corrector loop for Demand Multiplication Factors (DMFs) prediction at the t^{th} time step

The basic building blocks of the predictor-corrector model (e.g., *M5 Model Trees algorithm* and *Genetic Algorithms*) are described in the following paragraphs.

M5 Model Trees algorithm (Quinlan 1992)

The M5 model trees algorithm (Appendix A give a more comprehensive description of the methodology) builds rule-based predictive models using a top-down induction approach. The tree is fitted to a training data set by recursively partitioning the data into homogeneous subsets based on its attributes. Thereafter, the tree is constructed with all training cases being predicted by the tree leaves (i.e., each leaf is a linear regression model that can explain the remaining variability of each homogeneous subset). In order to simplify the tree structure, and thus to improve its ability to classify new instances, the tree is then pruned from the bottom-up by quantifying the contribution of each attribute to the overall predicted value and removing those

attributes that add little to the model. At the last stage, a smoothing process is performed to compensate for the sharp discontinuities that will inevitably occur between adjacent linear models at the leaves of the pruned tree.

As observed in previous studies (Solomatine and Xue 2004, Bhattacharya et al. 2007, and Ould-Ahmed-Vall et al. 2007), Model Trees have several advantages compared to other data-driven techniques such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Regression Trees: Their accuracy is similar to that of both ANN and SVM and much better than that of Regression Trees; they are interpretable to users (in contrast to 'black-box' ANN and SVM models) and hence can provide insights to prevent sources of inefficiencies; and they are much faster in training compared to all other methods.

Genetic Algorithms (GA) (Holland, 1975)

Genetic Algorithms are heuristic combinatorial search techniques that imitate the mechanics of natural selection and natural genetics of Darwin's principles of evolution. The basic idea is to simulate the natural evolution mechanisms of chromosomes (represented by string structures), involving: selection, crossover, and mutation. This is accomplished by creating a random search technique that combines survival of the fittest among string structures with a randomized information exchange. A typical form of a genetic algorithm involves three main stages: 1) Initial population generation: the genetic algorithm generates a bundle of strings (termed population, or generation), with each string (chromosome) being a set of values of the decision variables/optimization parameters. 2) Computation of string fitness: the genetic algorithm evaluates each string's fitness (i.e., the value of the objective function that corresponds to each string). 3) Construction of a new generation: the genetic algorithm establishes the next generation by performing selection, crossover, and mutation. The process of selection involves choosing strings (chromosomes) from the current population for reproduction according to their fitness values. Crossover involves the partial exchange of information between pairs of strings; and mutation is the random change in one of the string locations. The genetic algorithm parameters are the population size, the mating and mutation rates, and the number of generations.

Calibration Problem Objective Function

The objective of the calibration process is to match the computed and measured sensor node data (pressure and/or flow rates), taking into consideration possible noise in the measurements. In this application, calibration is achieved by minimizing a modified least-squares of differences function known as the Huber function (Huber 1973). The Huber function implementation to the hydraulic state estimation problem is described as follows:

1. The differences (i.e., residuals) between modeled and observed pressures and flow rates at each time step, at sensor node *i* - are defined as $R_{i,t}^{P}$ and $R_{i,t}^{Q}$ respectively
2. The Huber function of each residual R is defined as

$$f\left(R_{i,t}^{P\ or\ Q}\right) = \begin{cases} \frac{1}{2}\left(R_{i,t}^{P\ or\ Q}\right)^2, & \left|R_{i,t}^{P\ or\ Q}\right| \le h \\ h\left|R_{i,t}^{P\ or\ Q}\right| - \frac{1}{2}h^2, & \left|R_{i,t}^{P\ or\ Q}\right| > h \end{cases} \quad (1)$$

   where h is a predefined value that represent the tolerance to noise in measurements; for small residuals ($|R| \le h$) that represent low to zero values of noise in sensor measurements, the Huber function minimizes the usual least squares function (i.e., l_2 norm approximation), for large R ($|R| > h$) that represent high values of noise in sensor measurements, it minimizes a linear penalty function which is relatively insensitive to noise (i.e., l_1 norm approximation)
3. The overall calibration problem objective function to be minimized at each hydraulic time-step *t* is defined as

$$\sum_{i=1}^{N_P} f\left(R_{i,t}^{P}\right) + \sum_{i=1}^{N_Q} f\left(R_{i,t}^{Q}\right) \quad (2)$$

   where *i* is the sensor nodes index, N_p is the total number of pressure sensors, and N_Q is the total number of flow rate sensors

- In this application the value of *h* in each sensor node at each time-step is equal to the average of all previous time-steps sensor node residuals multiplied by a factor of 2.

3. Results

The predictor-corrector approach developed in this study was tested against the real input data of Network 2 (Fig. 2) of the "Battle of the Water Sensor Networks (BWSN): A Design Challenge for Engineers and Algorithms" (Ostfeld et al. 2008). The network corresponds to an anonymous but real water distribution system comprising 12,523 nodes, two constant head

sources, two tanks, 14,822 pipes, four pumps, five valves. The system was subject to highly variable demand patterns over a period of 934 hours (~39 days). Hydraulic simulations for this system are considered valid for this entire duration. The original EPANET input file was downloaded from the University of Exeter Centre for Water Systems (ECWS) web-site (http://www.exeter.ac.uk/cws/bwsn).

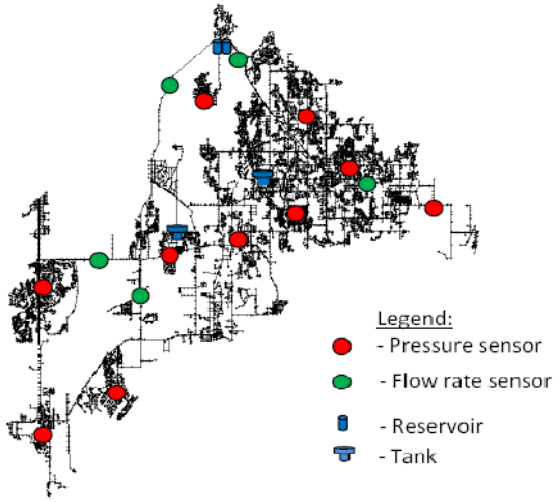

Fig. 2: Network 2 with the sensor nodes locations

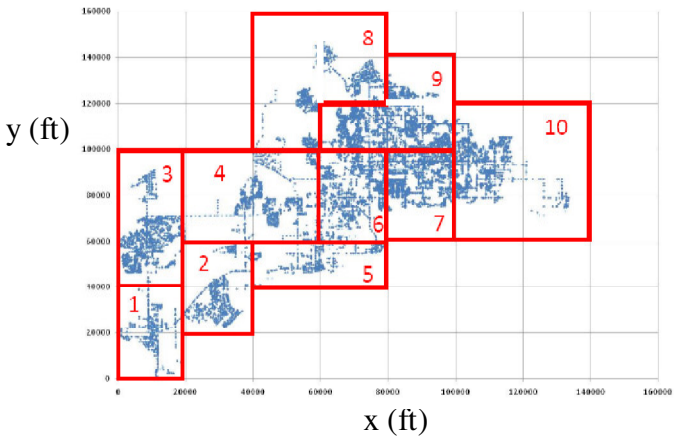

Fig. 3: demand nodes groups on a plane grid of the system

The current application assumes that continuous in-line data are available from 10 pressure and 5 flow rate sensors (Fig. 2). The nodal flow and pressure records from these locations were generated by the EPANET model using real input data for the system. The reservoirs and tank water levels were considered as known inputs. The analysis considers 10 demand zones (i.e., 10 groups of demand nodes; see Fig.3) based on a spatial analysis of the system. It is expected that the consumption nodes in each zone will follow the same demand pattern and each nodal base demand at each zone will be multiplied with the same DMF. Therefore, the number of decision variables to be calibrated/corrected at each time-step is equal to 10.

The total running time on a DELL PC (2.66 GHz, 3.0 GB of RAM) of the GA calibration process (i.e., with a GA population of 48 ‘chromosomes’ and 30 GA generations) is about 5 minutes and the total running time of the data driven prediction process is less than 10 seconds.

Base Run: Demand Multiplication Factors (DMFs) prediction accuracy

Following previous data-driven models implementations (e.g., Solomatine and Xue 2004 and Bhattacharya et al. 2007), about 33% of the data (from t = 680 to 934 hrs, corresponding to a total of 254 hrs) was used for cross validating the model predictions. The predictive ability of the model can be evaluated with several prediction metrics. In this application the following commonly used metrics were applied to evaluate the fit between predicted (p) and actual (a) values:

1. Correlation Coefficient (CC): measures the degree of correlation between predicted and actual values; it ranges from -1 to 1, with 1 corresponding to an ideal correlation:

$$CC = \frac{Cov(p,a)}{\sigma_p \sigma_a} \tag{3}$$

where $Cov(p,a)$ is the covariance between p and a; and σ_p, σ_a are their standard deviations

2. Root Mean Squared Error (RMSE): RMSE is a frequently-used measure of the differences between values predicted by a model or an estimator and the values actually observed; it ranges from 0 to infinity, with 0 corresponding to ideal fit:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(p_i - a_i)^2}{N}} \tag{4}$$

where p_i and a_i are the predicted and actual values of case i; and N is the number of cases

3. Mean Absolute Error (MAE): MAE is similar to RMSE, except it uses absolute error values instead of the squared errors; it ranges from 0 to infinity, with 0 corresponding to ideal fit:

$$MAE = \frac{\sum_{i=1}^{N}|p_i - a_i|}{N} \tag{5}$$

where p_i and a_i are the predicted and actual values of case i; and N is the number of cases

4. Root Relative Squared Error (RRSE): RRSE value is relative to what it would have been if a naive predictor had been used. More specifically, this simple predictor is just the average of the actual values; it ranges from 0 to infinity, with 0 corresponding to ideal fit:

$$RRSE = \sqrt{\frac{\sum_{i=1}^{N}(p_i - a_i)^2}{\sum_{i=1}^{N}(a_i - a_{av})^2}} \tag{6}$$

where p_i and a_i are the predicted and actual values of case i; a_{av} is the mean of the observed data; and N is the number of cases

The accuracy of the 10 zones DMF predictions is summarized in Table 1:

Zone index	CC	RMSE	MAE	RRSE
1	0.87	0.05	0.06	0.16
2	0.89	0.06	0.06	0.18
3	0.92	0.04	0.05	0.18
4	0.85	0.06	0.06	0.18
5	0.85	0.06	0.06	0.18
6	0.86	0.06	0.07	0.17
7	0.91	0.07	0.06	0.16
8	0.92	0.05	0.06	0.16
9	0.87	0.04	0.05	0.15
10	0.92	0.04	0.05	0.15

Table 1: Predictive metrics for DMFs in 10 demand zones from base run

The results indicate that the predictor – corrector model has relatively good predictive accuracy for all 10 demand zones according to all four metrics. All of the DMFs were predicted with a correlation coefficient exceeding 0.85 with mean absolute error below 10%. Figures 4 to 6 show the improvement achieved in the predictor-corrector model predictions through experience, as demonstrated on one of the system demand nodes (e.g., Junction 7631). In this example, the correlation coefficient (CC) is used as the characteristic measure of predictive accuracy. The data set of results from t = 169 – 934 hrs was divided into 3 segments (Δt_1 =169 - 424 hrs; Δt_2 =425 - 679 hrs; and Δt_3 = 680 to t=934 hrs). Figure 4 shows a correlation coefficient, CC = 0.66 for time period Δt_1. This low value is explained by insufficient input data for the Model Trees predictor in forecasting future DMFs. For the second (Fig. 5) and third (Fig. 6) time periods, with the increase in training data, there is an improvement in the predictor-corrector performances that is refelected in higher correlation coefficients, CC = 0.85 and 0.92, for periods Δt_2 and Δt_3, respectively.

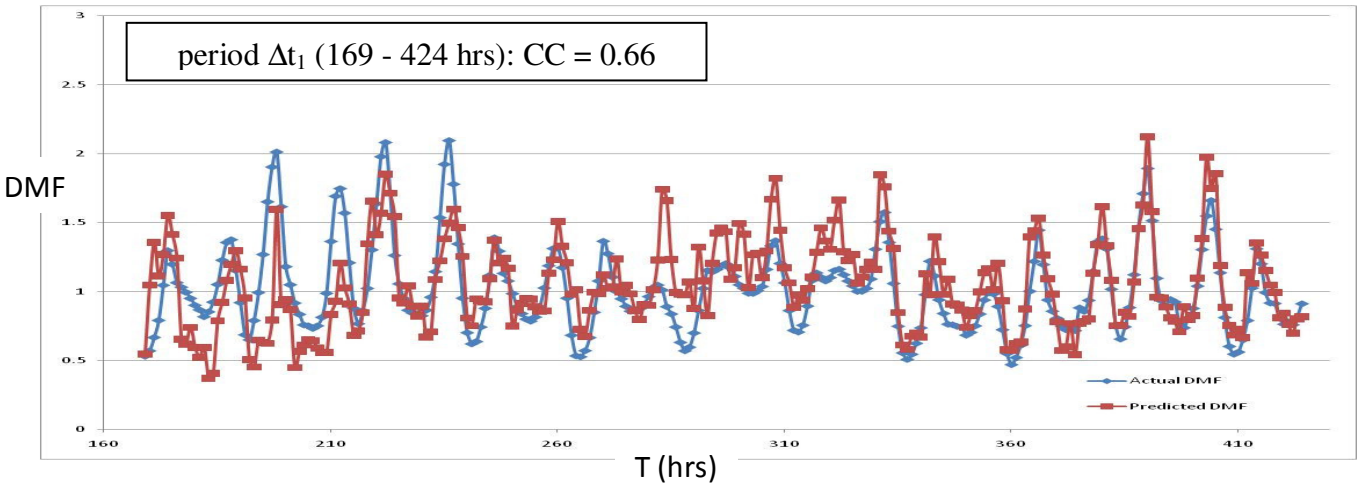

Fig. 4: Comparison between predicted and actual DMFs for time period 1;

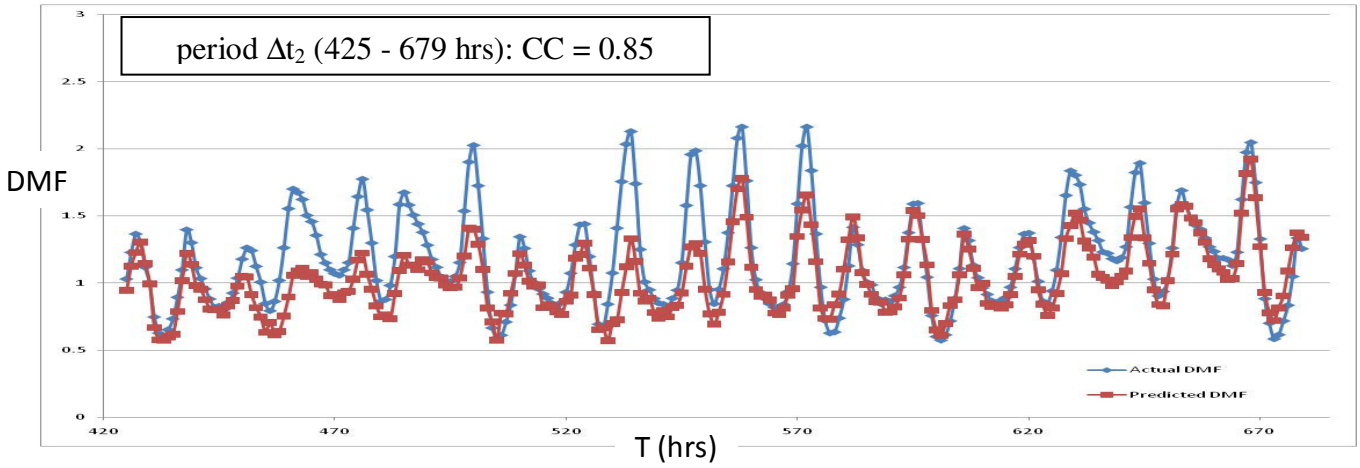

Fig. 5: Comparison between predicted and actual DMFs for time period 2;

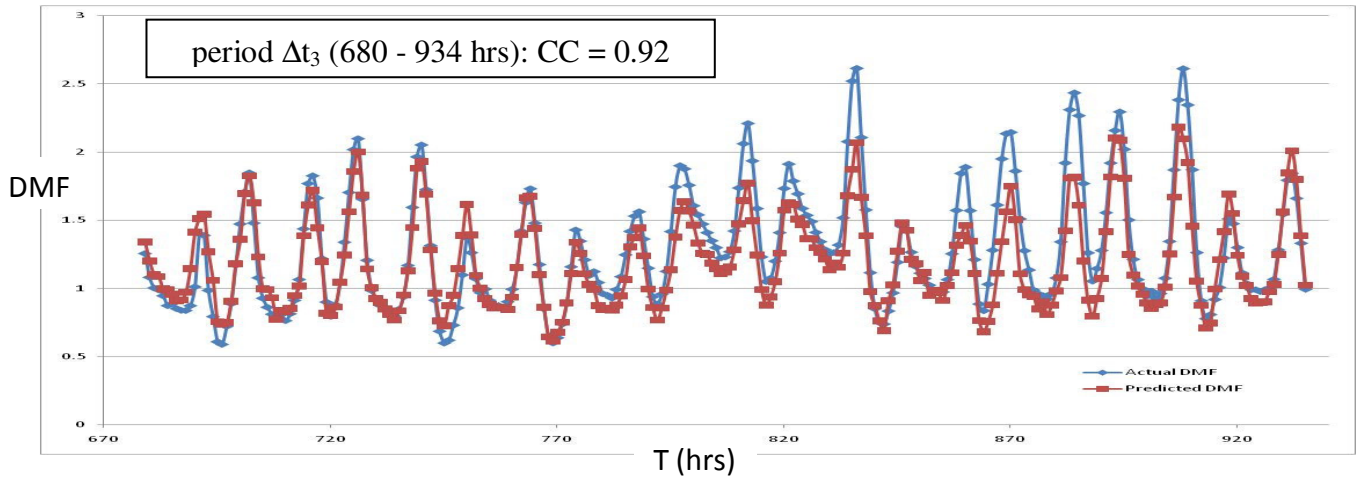

Fig. 6: Comparison between predicted and actual DMFs for time period 3;

Comparison between on-line and off-line calibration model performance

Since the main hypothesis of this study is that on-line hydraulic-state prediction models are more accurate than off-line calibration models in representing the hydraulic state of the water system, the predictive accuracy of the proposed predictor-corrector analysis must be compared with reference results from off-line calibration. The off-line reference result are based on a calibration using the same pressure and flow records for a one-week period (Δt_0 = 0 – 168 hrs). The off-line calibration process was much more intensive then the on-line model (200 GA iterations instead of 30, with a GA population of 120 decision variables' strings instead of 48) and resulted in a relatively high fit for the DMF's during the calibration time period.

The predicted Demand Multiplication Factors (DMFs) for time period 3 (680 - 934 hrs) were used together with EPANET analyses to calculate unknown hydraulic outputs (e.g., flow rates; flow directions; and nodal pressures). Thereafter, these parameters, which represent the hydraulic state of the water system - were used to evaluate and compare the predictions accuracy of the on-line and off-line models.

1. Flow rate prediction

The statistical evaluation of the predicted flow rates includes 5 ranges of predictive accuracy (i.e., within 5, 10, 15, 20, and 25 % of the actual flow rates for two groups of pipes: 1) pipes with diameters less than 24'' which correspond to more than 80% of the network length and 2) medium and large pipes with diameter greater than 24" The statistics are computed for all the pipes in the system (e.g., 15000 pipes) throughout time period 3 (i.e., over a period of 254 hrs); such that the total number of flow rates considered is 3.76×10^6 (14822x254). The results for pipe groups 1 and 2 are shown in Figures 7 and 8, respectively. The proposed predictor-corrector model outperforms the off-line calibration process and provides significantly better estimates of flow rates for both sets of pipes. The predictions are generally much more accurate for the medium and large pipes (group 2, Figure 8) than for the small diameter pipes (group 1, Figure 7). For the medium and large diameter pipes, the predictor-corrector model is within 5% of the actual flow rates for 68% of the samples and within 25% for 95% of the sample population (Fig. 8). For the smaller pipes the 5% and 25% accuracy rates drop to 48% and 86% of the

population. In contrast, the off-line calibration is only within 5% of the actual flow for 22% of the small pipes and 43% of the medium to large pipe sample population.

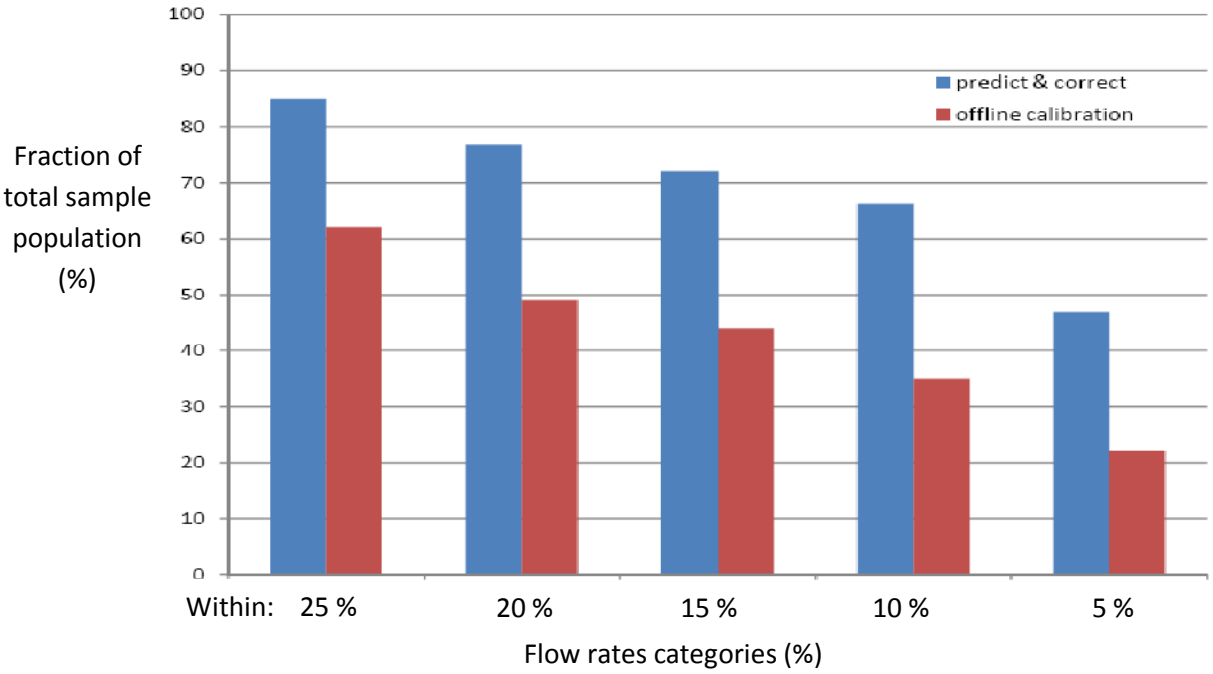

Fig. 7: Comparison of estimated and measured flow rates using the predictor-corrector and off-line calibration methods small pipes (with diameter less than 24")

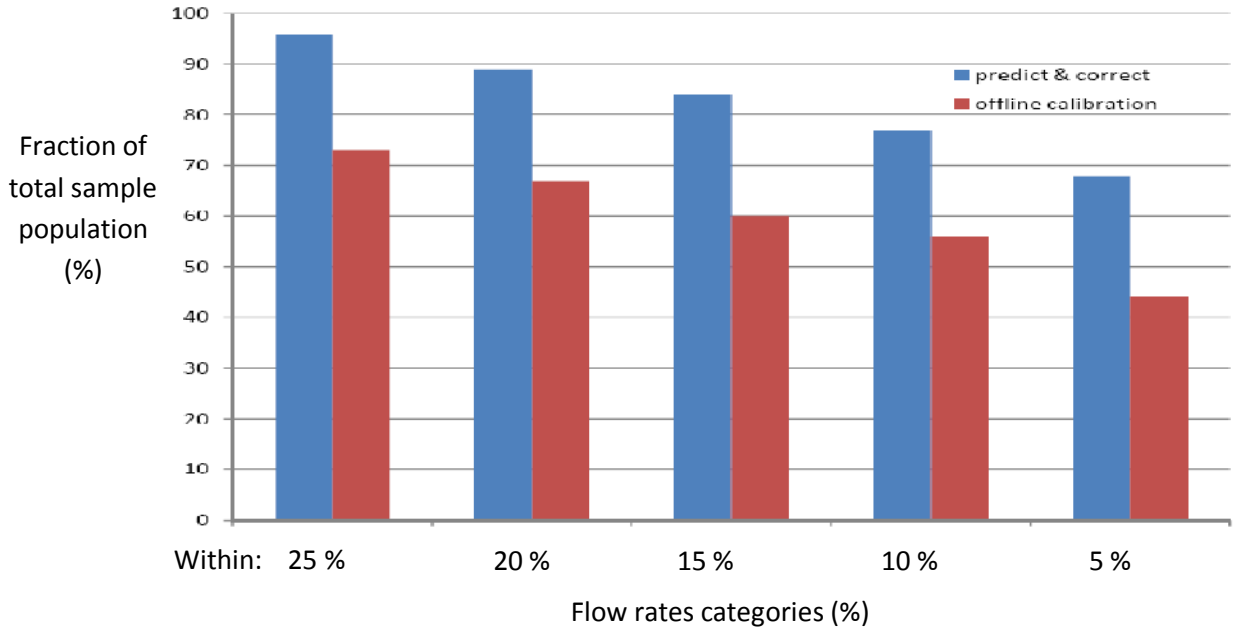

Fig. 8: Comparison of estimated and measured flow rates using the predictor-corrector and off-line calibration methods for medium and large pipes (diameter greater than 24")

The proposed predictor-corrector model correctly estimated the flow direction for 100 % of the sample population of 114822 pipes from t = 680 to 934 hrs while the off-line calibration was accurate for 97% of the actual flow directions in the system. The main conclusion from the above is that flow directions are less sensitive to variations in the DMFs (at least for this reference pipe network) as both the on-line and off-line models were able to provide very accurate estimations of this parameter.

2. Pressure prediction

The statistical evaluation of nodal pressures considers 7 ranges of predictive accuracy (i.e., ± 2, 4, 6, 8, 10, 12 and 14 psi of the actual pressures). The statistics are computed for all the nodes in the system (e.g., 12523 nodes) through time period 3 (t = 680 - 934 hrs); corresponding to a total sample population of 3.81×10^6. It can again be observed that the proposed predictor-corrector model provides significantly better estimations for the pressures than the off-line calibration process. In this case, 87% of the samples are estimated within 4psi using the predictor-corrector scheme compared to 42% using the off-line calibration.

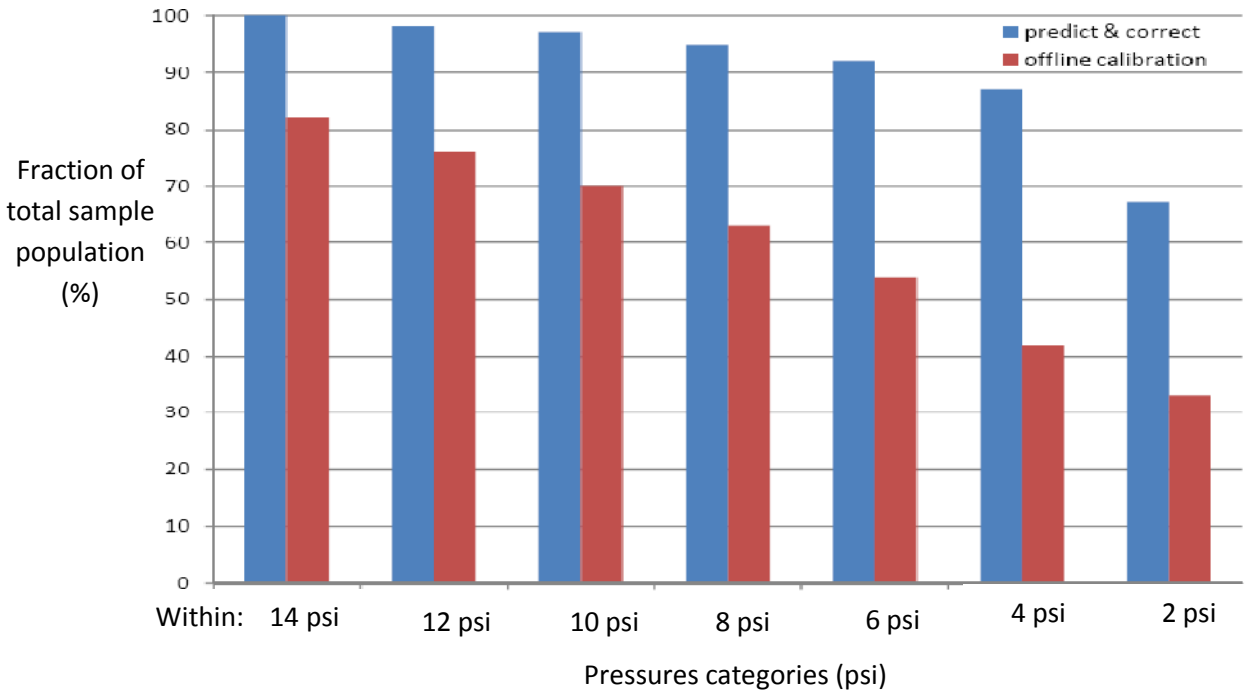

Figure 9: Comparison of predictive accuracy for nodal pressures using the proposed predictor-corrector scheme and an off-line calibration process

Effect of Demand Zone selection

The Authors have carried out a sensitivity analysis to evaluate how the partition of the system into demand zones can affect performance of the predictor-corrector model. Figure 10 shows a revised partition of the network into 20 demand zones. In order to keep the problem slightly over-determined, the number of the sensor locations was increased to 25. In order to comply with the increase in the number of decision variables in the calibration problem, the genetic algorithm parameters were altered to 60 iterations with a population of 96 decision variable strings (vs. 30 and 48, respectively, used in the base run).

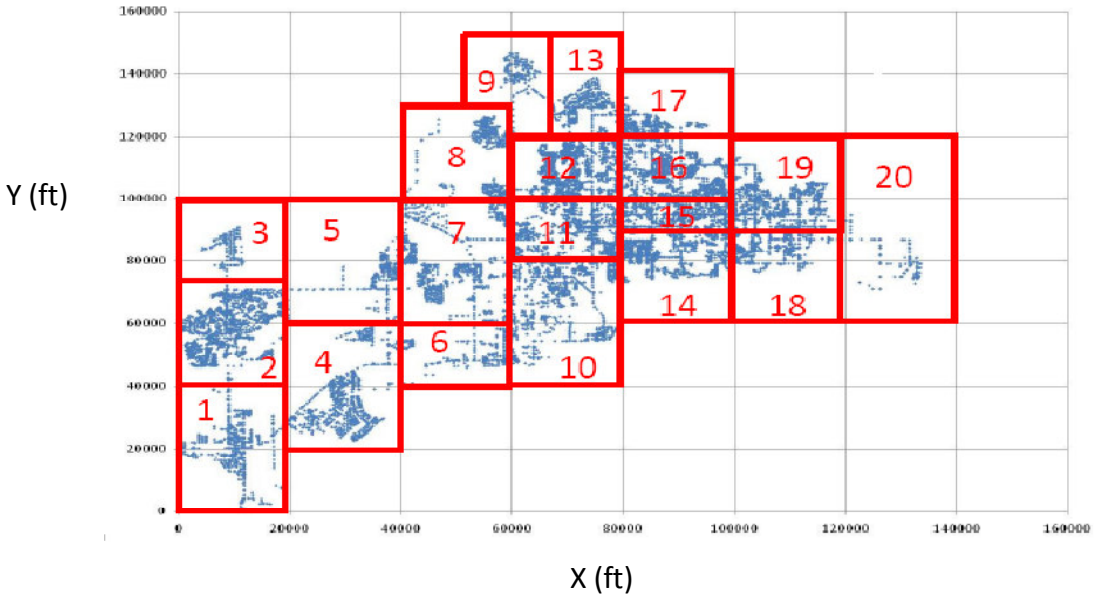

Fig. 10: Sensitivity analysis using 20 demand zones for the reference pipe network

The correlation coefficients for the DMF's in the 20 zones DMF are summarized in Table 2:

Zone index	CC	Zone index	CC
1	0.88	11	0.85
2	0.86	12	0.88
3	0.83	13	0.87
4	0.89	14	0.78
5	0.84	15	0.89
6	0.88	16	0.92
7	0.88	17	0.91
8	0.86	18	0.86
9	0.93	19	0.86
10	0.92	20	0.87

Table 2: Correlation coefficients of DMFs for the 20 zone model of the network

The average correlation coefficient for the entire system (CC = 0.87) is slightly lower than that obtained in the base run (average CC = 0.89) but requires much higher computational effort (e.g., about double the total running time of the base run) making it not worthwhile to change the demand zones selection used in the base run. Although, the revised partition of the network into 20 demand zones need not necessarily improve the predictive accuracy of the model (at least using CC values) this process of testing different system partitions is important as it may improve the model performances. As a conclusion of the sensitivity analysis, grouping demand nodes into demand zones requires high level of judgment based on good knowledge of the system structure. Therefore, if Geographical Information System (GIS) data of the municipality is available, then it is advisable to incorporate it in the demand zone selection procedure and to test the model performances with several system partitions.

4. Summary

This paper has presented and demonstrated a Predictor-Corrector (PC) model for on-line, hydraulic state prediction of urban water networks. The method uses a statistical data-driven algorithm (M5 Model Trees algorithm) to estimate future water demands, while near real-time field measurements are used to correct (i.e., calibrate) these predicted values on-line. The calibration problem is solved using Genetic Algorithm with a modified Least Squares (LS) fit method (Huber function) to account for noisy measurements. The a-priori estimation (i.e., prediction) of the decision variables values, which improves through experience facilitates a better convergence of the calibration model towards the optimal solution of the problem; and provides adequate information on the system's hydraulic state for real time optimization. Future research efforts will focus on the implementation of the developed methodology on large scale urban water system using physical data from an in-situ sensor network. Additional efforts will focus on the ability to detect anomalies such as leakage and burst events in real time. The integration of water networks aggregation methods to reduce the computational time required for the calibration process will also be explored.

References

Alvisi, S., Franchini, M., Marinelli, A., (2007),"A short-term, pattern-based model for water-demand forecasting" Journal of Hydroinformatics, Vol. 9, No. 1

Bhattacharya, B., Price, R.K., and Solomatine, D.P., (2007), "A machine learning approach to modeling sediment transport", ASCE J. of Hydraulic Engineering,133(4), 440-450.

Clark , C., Wu, Z, Y. (2006), "Integrated Hydraulic Model and Genetic Algorithm Optimization for Informed Analysis of A Real Water System", 8th Annual Water Distribution Systems Analysis Symposium, Cincinnati, Ohio, USA, CD-Rom.

Davidson, J. W., Bouchart, F. J.-C., (2006), "Adjusting Nodal Demands in SCADA Constrained Real-Time Water Distribution Network Models", Journal of Hydraulic Engineering, Vol. 132, No. 1

Datta, R.S.N., and Sridharan, K. (1994), "Parameter Estimation in Water-Distribution Systems by Least Squares", Journal of Water Resources Planning and Management, ASCE, 120(4), 405-422.

EPANET. (USEPA 2002) Available on line at: www.epa.gov/ORD/NRMRL/wswrd/epanet.html

Ghiassi, M., Zimbra, D. K., and Saidane, H., (2008), "Urban Water Demand Forecasting with a Dynamic Artificial Neural Network Model", Journal of Water Resources Planning and Management, Vol. 134, No. 2

Greco, M., and Del Guidice, G. (1999), "New Approach to Water Distribution Network Calibration", Journal of Hydraulic Engineering, ASCE, 125(8), 849-854.

Holland J. H. (1975). "Adaptation in natural and artificial systems." The University of Michigan Press, Ann Arbor.

Huber, P. J., (1973), "Robust regression: Asymptotics, conjectures, and Monte Carlo" Ann. Statist., 1, 799-821.

Jonkergouw, P. M. R., Khu, S.-T., Kapelan, Z. S., and Savić, D. A., (2008), "Water Quality Model Calibration under Unknown Demands" Journal of Water Resources Planning and Management, Vol. 134, No. 4

Kapelan, Z. S., Savic, D. A., and Walters, G. A. (2002) ''Hybrid GA for calibration of water distribution system hydraulic models.'' Proc., 1st Annual Environmental & Water Resources Systems Analysis (EWRSA) Symp., CD-Rom.

Kapelan, Z. S., Savic, D. A., and Walters, G., A., (2007), "Calibration of Water Distribution Hydraulic Models Using a Bayesian-Type Procedure" Journal of Hydraulic Engineering, Vol. 133, No. 8,pp. 927-936

Kenward, T. C., and Howard, C. D. (1999). "Forecasting for urban water demand management." *Proc., 26th Annual Water Resources Planning and Management Conf.*, ASCE, Reston, Va.

Lansey, K. E., and Basnet, C. (1991). "Parameter estimation for water distribution networks." Journal of Water Resources Planning and Management, ASCE, 117(1), 126– 144.

Lingireddy, S. and Ormsbee, L. (1998) "Neural Networks in Optimal Calibration of Water DistributionSystems," Artificial Neural Networks for Civil Engineers: Advanced Features and Applications, ASCE.

Maidment, D. R., and Miaou, S. P. (1986). "Daily water use in ninecities." *J. Water Resour. Plann. Manage.*, 110 (1), 90–106.

Ormsbee, L. E. (1989). "Implicit network calibration." Journal of Water Resources Planning and Management, ASCE., 115(2), 243–257.

Ostfeld A. et al. (+ 34 co-authors) (2008). "The battle of the water sensor networks: a design challenge for engineers and algorithms." Journal of Water Resources Planning and Management Division, ASCE, Vol. 134, No. 6, pp. 556- 568.

Ould-Ahmed-Vall, E., Woodlee, J., Yount, C., Doshi, K., (2007) "On the Comparison of Regression Algorithms for Computer Architecture Performance Analysis of Software Applications", First Workshop on Statistical and Machine learning approaches applied to Architectures and compilation**,** Ghent, Belgium

Quinlan J. R. (1992). "Learning with continuous classes." Proceedings 5th Australian Joint Conference on Artificial Intelligence. World Scientific, Singapore, 343-348

Reddy, P.V.N., Sridharan, K., and Rao, P.V. (1996), "WLS Method for Parameter Estimation in Water Distribution Networks", Journal of Water Resources Planning and management, ASCE, 122(3), 157-164.

Savic, D.A., and Walters, G.A. (1995), "Genetic Algorithm Techniques for Calibrating Network Models", Report No. 95/12, Centre for Systems and Control Engineering, University of Exeter, p. 41.

Shang, F., Uber, J., van Bloemen Waanders, B., Boccelli, D. , Janke, R.(2006) "Real Time Water Demand Estimation in Water Distribution System", 8th Annual Water Distribution Systems Analysis Symposium, Cincinnati, Ohio, USA, CD-Rom.

Solomatine, D.P. and Xue, Y., (2004) "M5 model trees compared to neural networks: application to flood forecasting in the upper reach of the Huai River in China" ASCE Journal of Hydrologic Engineering, 9(6)

USEPA (2005), Water Distribution System Analysis: Field Studies, Modeling and Management. A Reference Guide for Utilities. USEPA Cincinnati, Ohio, USA

Walski, T. M. (1983). "Technique for calibrating network models." Journal of Water Resources Planning and Management, ASCE, 109(4), 360–372.

Walski, T.M., Chase, D.V., Savic, D.,A., Grayman, W., Beckwith, S. and Koelle, E., (2003) "Advanced Water Distribution Modeling and Management," Haestad Methods, Inc. Waterbury, CT.

Walski, T.M., Defrank, N., Voglino, T., Wood, R., Whitman, E. (2006), "Determining the Accuracy of Automated Calibration of Pipe Network Models", 8th Annual Water Distribution Systems Analysis Symposium, Cincinnati, Ohio, USA, CD-Rom.

Wu, Z. Y. and Simpson A. R. (2001) "Competent Genetic Algorithm Optimization of Water Distribution Systems." J. of Computing in Civil Engineering, ASCE, Vol 15, No. 2, pp 89-101.

Wu, Z, Y., Walski, T.M., Mankowski, R., Herrin, G., Gurrieri,, R. and Tryby, M. (2002), "Calibrating Water Distribution Models via Genetic Algorithms," Proceedings AWWA Information Management Technology Conference, Kansas City, Mo.

Zhou, S. L., McMahon, T. A., and Lewis, W. J. (2000). "Forecasting daily urban water demand: A case study of Melbourne." *J. Hydrol.*, 236, 153–164.

Appendix A: *Model Trees* Construction using the *M5* Algorithm (Quinlan 1992)

Suppose we have a set of *T* training cases. Each case is specified by its value of a fixed set of attributes and has an associated target value. The aim is to construct a model that relates the target values of the training cases to their values described by the input attributes. The performance of the model will generally be measured by the accuracy with which it predicts the target values of unseen cases (cross-validation data set). Fig. A1 illustrates a set *T* which is described by two input attributes (x_1, x_2) and one target attribute (y):

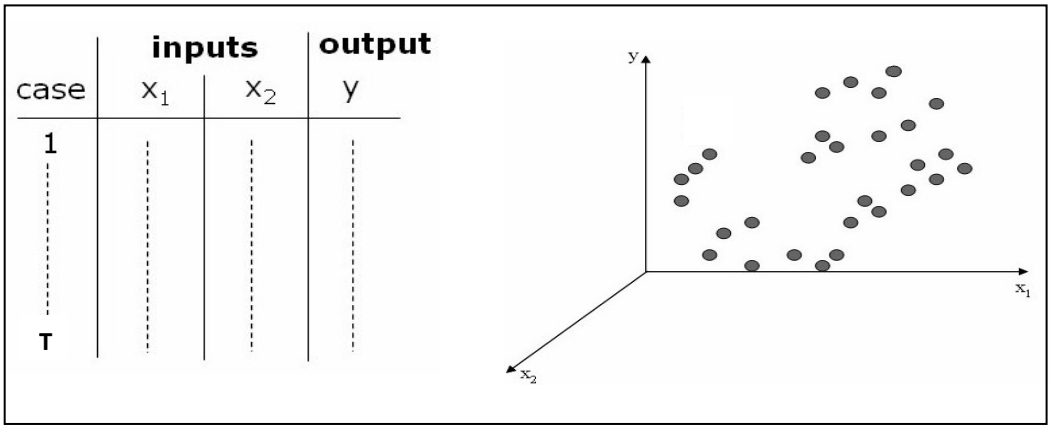

Fig. A1: Set T - described by 2 input attributes (x_1, x_2) and 1 target attribute (y)

In this tree based model, the set *T* is either associated with a leaf, or some test is chosen that splits *T* into subsets corresponding to the test outcomes. The same process is applied recursively to the subsets. This persistent division often produces over-elaborate structures that must be pruned back (i.e., this process is described by the following steps).

Step 1: Building the Initial Tree

The standard deviation of the target values of cases in *T* is computed. Unless *T* contains very few cases (e.g., three examples or less) or their values vary only slightly (e.g., 5% of the standard deviation of the class values of the original set of examples), *T* is split on the outcome of a test. Every potential test is evaluated by determining the subset of cases associated with each outcome; let T_i denote the subset of cases that have the *i-th* outcome of the potential test. If we treat the standard deviation $sd(T_i)$ of the target values of cases in T_i as a measure of error, the expected reduction in error (*SDR*) as a result of this test can be written as:

$$SDR = sd(T) - \frac{1}{|T|}\sum_i |T_i| \, sd(T_i) \qquad \text{(A1)}$$

After examining all possible tests, the model tree chooses one that maximizes this expected error reduction (*SDR*). At this step an initial model tree has been grown and a multivariate linear model is constructed for the cases at each node of the model tree using standard regression techniques. Fig. A2 illustrates this process on the data set *T* introduced in Fig. A1.

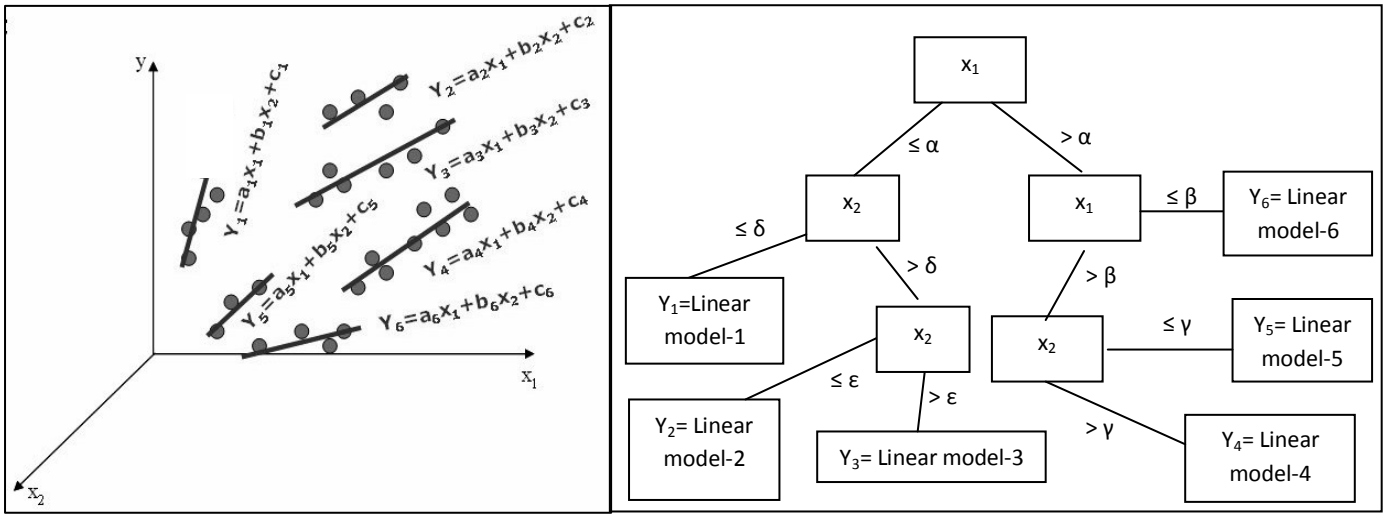

Fig. A2: model tree construction based on the data set - *T*

After that, the tree is transferred into a simple set of *If-Then* rules that simplify the tree structure and make the tree internal structure interpretable to users. Fig. A3 illustrates this process on the data set *T* introduced in Fig. A1.

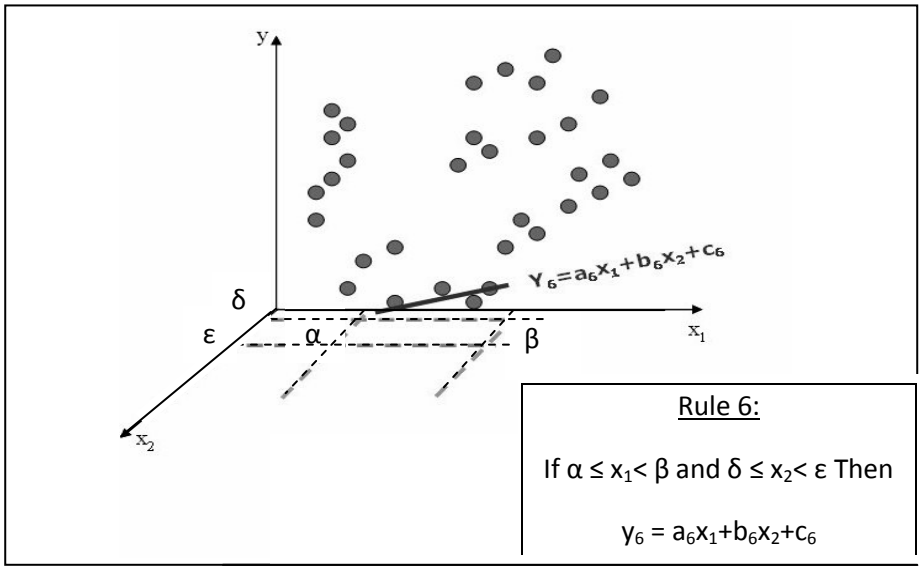

Fig. A3: model tree transformation into an *If-Then* set of rules demonstrated on sub-set Y_6

Step 2: Pruning the Tree

At the second step, the over-elaborated model tree structure is simplified, and pruned bottom-up - and thus its ability to classify new data sets (i.e. cross validation data) is improved. The simplification of the model tree is performed mainly by removing variables that contribute little to the model; in some cases the algorithm removes all variables, leaving only a constant.

Step 3: Smoothing

The smoothing process is performed to compensate for the sharp discontinuities that will inevitably occur between adjacent linear models at the leaves of the pruned tree, particularly for some models constructed from a smaller number of training examples. In smoothing, the adjacent linear equations are updated in such a way that the predicted outputs for the neighboring linear input equations are becoming close in value.