On-line Hydraulic State Estimation in Urban Water Networks Using Reduced Models

A. Preis
Center for Environmental Sensing and Modeling, MIT-SMART Center, Singapore

A. J. Whittle
Department of Civil and Environmental Engineering, MIT, Cambridge, MA, USA & Center for Environmental Sensing and Modeling, MIT-SMART Center, Singapore

A. Ostfeld & L. Perelman
Faculty of Civil and Environmental Engineering, Technion – I.I.T, Haifa, Israel

ABSTRACT: A Predictor-Corrector (PC) approach for on-line forecasting of water usage in an urban water system is presented and demonstrated. The M5 Model-Trees algorithm is used to predict water demands and Genetic Algorithms (GAs) are used to correct (i.e., calibrate according to on-line pressure and flow rate measurements) these predicted values in real-time. The PC loop repeats itself at each subsequent time-step with the forecasting model inputs being the corrected outputs of previous iterations, thus improving the model performances over time. To meet the computational efficiency requirements of real-time hydraulic state estimation, the urban network model which is comprised of over ten thousand pipelines and nodes is reduced using a water system aggregation technique. The reduced model, which resembles the original system's hydraulic performances with high accuracy, simplifies the computation of the PC loop and facilitates the implementation of the on-line model. The developed methodology is tested against the real input data of an urban water distribution system comprised of approximately 12500 nodes and 15000 pipes.

1 INTRODUCTION

On-line operation and control of large-scale urban water distribution systems relies on integrating computer simulation models with near real-time hydraulic data (i.e., with continuous pressure and flow rate measurements, provided by a sensor network installed on the distribution system). This on-line procedure can be used to estimate the system's hydraulic state for operational planning of the water utility.

There have been several recent studies that have assimilated on-line measurements into hydraulic state estimation models. Davidson and Bouchart (2006) proposed 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.

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 was used to predict the water demands based on the estimated demands at previous steps and the forecasts were corrected using measured nodal water heads or pipe flow rates.

This study uses a Predictor-Corrector (PC) approach which integrates a limited number of continuous hydraulic observations to continually update predictions of the hydraulic state of a real urban water supply network. The *M5 Model-Trees algorithm* (Quinlan 1992) is used to forecast future water demands for a rolling planning horizon of 24 hrs 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 repeats itself at each subsequent time-step while the forecasting model inputs correspond to the corrected outputs of previous iterations, thus improving the model performances over time and providing adequate information on the system's hydraulic state for real time operation and control.

To meet the computational efficiency requirements of this on-line procedure, the urban network model is condensed to an equivalent system, with a reduced number of links and nodes through a system aggregation technique (Ulanicki et al. 1996).

The reduced model, which resembles the original system's hydraulic performances with high accuracy, simplifies the hydraulic model and facilitates efficient

implementation of the real-time, state estimation technique.

2 METHODOLOGY

2.1 *Hydraulic Aggregation*

At the initial stage, which is carried out off-line, the algorithm developed by Ulanicki et al. (1996) was used to create an equivalent reduced system.

The algorithm proceeds in a step-by-step elimination of pipes and nodes, allocating the demand at the node being eliminated to its neighboring nodes. All reservoirs, pumps, valves, tanks, and critical nodes (e.g., nodes in which pressure is monitored and nodes that represent significant water customers) remain in the reduced network. The validity of the system's reduction is measured by the similarity of the connectivity of the simplified system with that of the original system and its hydraulic performance (e.g., similarity of pressure at nodes, water levels at tanks, and\or pumps operation) over a wide range of operating conditions. The method of Ulanicki et al. (1996) is based on reducing the algebraic system of mass and energy conservation equations by eliminating variables using Gauss-elimination (Hammerlin and Hoffman 1991). The method involves the following stages:

2.1.1 *Full nonlinear modeling of the system's hydraulics*

The complete nonlinear mathematical description of the system hydraulics can be described by formulating mass conservation equation for each node of the network (i.e., Kirchoff's law 1):

$$AQ = q \tag{1}$$

And energy balance equation for all basic loops of the network (i.e., Kirchoff's law 2):

$$A^T h = \Delta h \tag{2}$$

where $A(nodes, links)$] = directed incidence matrix of the network graph $G(nodes, links)$; Q = vector of unknown flows in the links; h = vector of unknown nodal heads; q = vector of known demands at the nodes; Δh=vector of the head-losses along the links (i.e., pipes).

The relationship between the head loss and the flow in pipe, i, can be expressed with the pipes component law using the Hazen-Williams coefficient:

$$Q_i(\Delta h_i) = g_i(D_i, CHW_i, L_i)\Delta h_i^{1/e_1} \tag{3}$$

where g_i is the pipe's conductance (i.e., a function of the pipe diameter D_i, the Hazen-Williams head-loss coefficient CHW_i, and the pipe length L_i, with the constant e_1=1.852).

2.1.2 *Linearization of the system's hydraulic model*

For a given operating point defined by the nodal head h^0 and demand q^0, the linearized approximation describes the relationships between small changes in nodal head δh and demand δq around the chosen operating point. After linearization, Eqs. (1) and (2) take the following form:

$$G\delta\Delta h = \delta q \tag{4}$$

where $G=A[dQ(\Delta h)/d(\Delta h)]A^T$ is the Jacobian symmetric matrix whose elements are linear pipe conductances , $\tilde{g}$, which can be evaluated using Eq. (5), and $\delta\Delta h$ and δq = fluctuations in the nodal head and demand, respectively.

The elements of the Jacobian matrix are computed using

$$\begin{aligned} \tilde{g}_{i,j} &= \frac{1}{e_1} g_{i,j} \left|\Delta h_{i,j}\right|^{\left(\frac{1}{e_1}-1\right)}, \\ \tilde{g}_{i,i} &= \sum_j \frac{1}{e_1} g_{i,j} \left|\Delta h_{i,j}\right|^{\left(\frac{1}{e_1}-1\right)} \quad \forall i, j \end{aligned} \tag{5}$$

where $\tilde{g}_{i,j}$ and $g_{i,j}$ are the linear and nonlinear conductances of a pipe connecting nodes i and j, respectively; $\tilde{g}_{i,i}$ is the linearized node, i, conductance (i.e., the sum of the linearized pipe conductances of the pipes which connected to the node); and $\Delta h_{i,j}$ is the pipe's head loss.

The linear system of equations [Eq. (4)] describes head-flow relationship around an operating point.

Eq. (4) corresponds to Ohm's law and the flow in the pipe is equal to the linear conductance of the pipe multiplied by its head loss.

2.1.3 *Linear model reduction using Gauss-Elimination procedure*

Following the linearization process, the network is then condensed by applying Hammerlin and Hoffman's Gauss-elimination process (1991) at which node, i, is removed from the network by eliminating the corresponding equation (equation, i). The demand of that node, δq_i, is redistributed among other nodes connected to node$_i$, proportionally to the conductance of the connecting pipes. The connecting pipes of the removed node are removed as well, and

new linear conductances and nodal demands are calculated for the remaining elements of the network.

2.1.4 *Reduced nonlinear model recovery from the reduced linear model*

At the last stage of the aggregation procedure, the reduced nonlinear model is retrieved using the relationships formulated in Eq. (5). The aggregated model contains fewer nodes and links, forms a new network topology, and resembles the hydraulic performance of the original system with high accuracy as will be shown in the results section.

2.2 *Predictor – Corrector (PC) model*

The following paragraphs 2.2.1 to 2.2.5 describe the model steps:

2.2.1 *Demand Multiplication Factors (DMF) prediction*

The patterns in demand over an hourly-basis time-steps are described by the Demand Multiplication Factors (DMFs). For each 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. The consumption nodes are grouped into demand zones based on spatial analysis of the system and each group of consumption nodes is assigned its own set of DMFs.

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 (Walski et al. 2003).

Thereafter, the demand zones DMFs are predicted using the *M5 Model Trees algorithm* (Quinlan 1992), with the inputs being the calibrated DMFs from past hours t-24, t-25, t-168, and t-169 [i.e., daily (24-h) and weekly (168-h) demand cycles are used for the DMFs forecasts].

The M5 model trees algorithm (Quinlan 1992) 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.

2.2.2 *EPANET simulation*

The system hydraulics are simulated using the steady-state mode of EPANET (USEPA 2002), with the predicted DMFs as inputs. The simulation outputs are nodal pressures and pipe flow rates.

2.2.3 *On-line hydraulic data integration*

Pressure and flow measurements (from a set of in-line sensors) are inserted to the model at each time step.

2.2.4 *DMF correction/calibration*

A calibration problem is formulated and solved using *Genetic Algorithm* (Holland 1975) which is a heuristic combinatorial search technique that imitate the mechanics of natural selection and natural genetics of Darwin's principles of evolution.

Genetic Algorithms (GAs) 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.

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 (Huber 1973)] which takes into account noisy measurements is implemented to solve the optimization problem.

The Huber function implementation to the hydraulic state estimation problem is described as follows:

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

$$f\left(R_{i,t}^{P\ or\ Q}\right)=\begin{cases}\frac{1}{2}\left(R_{i,t}^{P\ or\ Q}\right)^2, & \left|R_{i,t}^{P\ or\ Q}\right|\le h\\ h\left|R_{i,t}^{P\ or\ Q}\right|-\frac{1}{2}h^2, & \left|R_{i,t}^{P\ or\ Q}\right|>h\end{cases} \quad (6)$$

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 approx-

imation), for large residuals ($|R| > h$) 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). In this application, h = 2 × [average of previous time-steps sensor node residuals].

The overall calibration problem objective function to be minimized at each hydraulic time-step t is defined as

$$\sum_{i=1}^{N_P} f\left(R_{i,t}^{P}\right) + \sum_{i=1}^{N_Q} f\left(R_{i,t}^{Q}\right) \quad (7)$$

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.

2.2.5 *DMFs delay*

The calibrated DMFs are being delayed for 24, 25, 168, and 169 hrs before being used as inputs in the prediction model.

2.2.6 *DMFs prediction-correction loop*

Steps 2.2.1 to 2.2.5 (Figure 1) start at t=169 hr, after performing an off-line calibration procedure for the first 168 hrs (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 hrs DMFs values is available except of the min-max boundaries which 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.

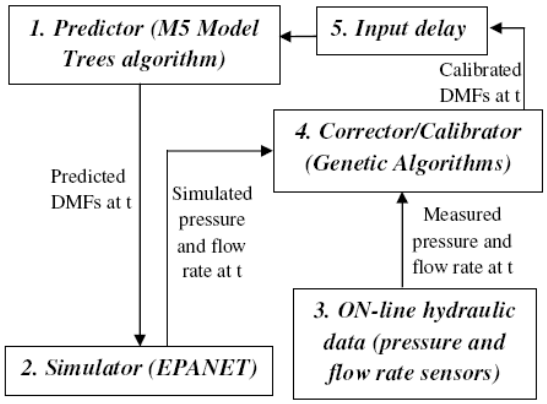

Figure 1. Predictor- Corrector loop for Demand Multiplication Factors (DMFs) prediction at the t^{th} time step

2.3 *Full network and reduced model interaction*

The predictor-corrector method is implemented on the reduced model of the water system to meet the computational efficiency requirements of this on-line procedure. Once the future water demands are being predicted for each demand-zone in the reduced model those predictions are assigned to the same demand zones in the full network model and can be used for further analysis of the full system performances.

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, and five valves. The system was subject to highly variable demand patterns over a period of 934 hrs (~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: www.exeter.ac.uk/cws/bwsn. The current application assumes that continuous in-line data are available from 30 pressure sensors (Fig. 2).

The nodal 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.

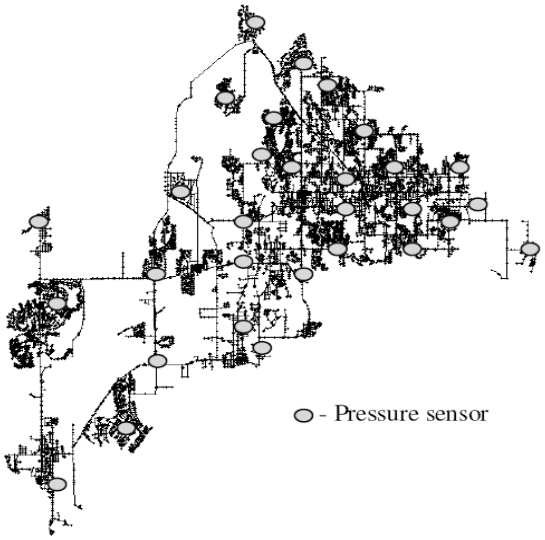

Figure 2. Network 2 full model with the sensor nodes locations

3.1 *Network aggregation results*

The method of Ulanicki et al. (1996) was used to create the reduced model (Fig. 3). All reservoirs, tanks, pumps, valves, nodes connected to 24" diameter pipes (main system's skeleton), pressure monitor-

ing nodes and significant consumption nodes (total of 347 nodes) remained in the aggregated network.

The validity of the reduction is measured by the similarity of sensor nodal pressure data over time in the reduced model with that calculated by the full model (as will be shown in Table 1). The pressure data for the validation test was generated using a representative sample of hourly demand pattern for 168 hrs (one week) of the utility operation. The min-max demand multiplication factors boundaries are 0 and 3, respectively.

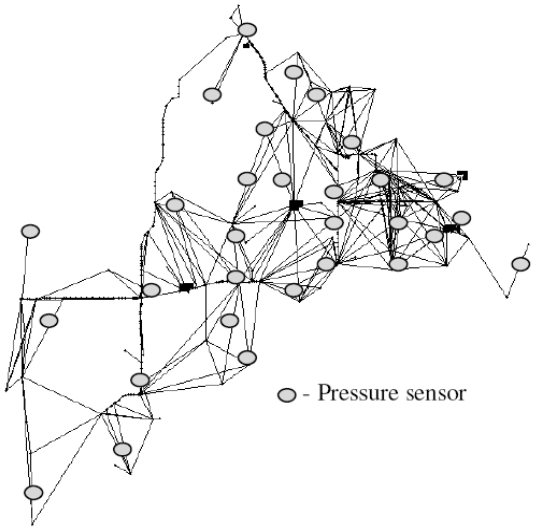

Figure 3. Network 2 reduced model with the sensor nodes locations

The reduced model resembles the original system's hydraulic performances with high accuracy (Table 1). The aggregated network contains 347 nodes and 1100 pipes, a reduction by a factor of about 35, and computation time for the hydraulic simulation is thereby reduced by 89 %.

Table 1. Comparison of sensor nodes pressure data over 168 hrs in the reduced model with that calculated by the full model

Ranges of pressure data accuracy (psi)	Fraction of the total sample population (%)
within 0.02 psi	99.88
within 0.04 psi	99.93
within 0.06 psi	99.96
within 0.08 psi	100

The analysis considers 20 demand zones [i.e., 20 groups of demand nodes (see 20 indexed squares in Fig.4)] which were chosen 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 DMFs.

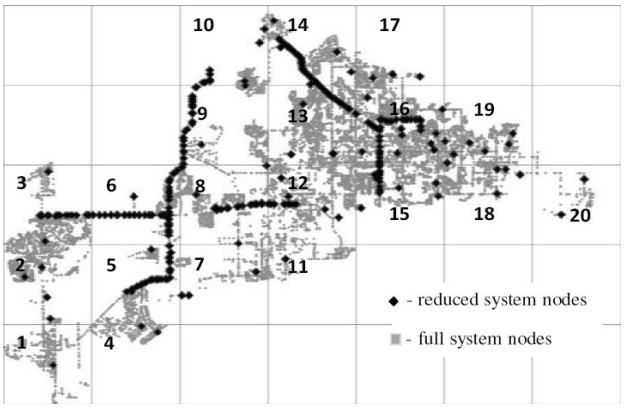

Figure 4. Demand nodes groups (20 demand zones) on a plane grid of the system

3.2 *Demand Multiplication Factors (DMFs) prediction accuracy*

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 120 decision variable strings and 100 GA iterations) is less than 5 seconds and the total running time of the data driven prediction process is less than 3 seconds.

The predictive ability of the model can be evaluated with several prediction metrics. In this application, the commonly used Correlation Coefficient (CC) was applied to evaluate the fit between predicted (p) and actual (a) DMF values.

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:

$$CC=Cov(p,a)/\sigma_p\sigma_a \qquad (8)$$

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

The accuracy of the 20 zones DMF predictions starting at t=169 hrs is summarized in Table 2. The improvement achieved in the predictor-corrector model predictions through experience, is also demonstrated in the table using 3 data segments of results at which the data set from t = 169 to t=934 hrs was divided into 3 time segments (Δt_1 = 169 - 424 hrs; Δt_2 = 425 - 679 hrs; and Δt_3 = 680 to t = 934 hrs).

The relatively low CC values of Δt_1 (average of 0.72) are explained by insufficient input data for the Model Trees predictor in forecasting future DMFs. For the second (Δt_2) and third (Δt_3) time periods, with the increase in training data, there is an improvement in the predictor-corrector performances that is refelected in higher correlation

coefficients (e.g., average CC(Δt_2) = 0.84 and average CC(Δt_3) = 0.9).

Table 2. Predictive metrics for DMFs in 20 demand zones for t=169 hrs to t=934 hrs

Demand zone	CC (Δt_1)	CC (Δt_2)	CC (Δt_3)
1	0.72	0.83	0.88
2	0.69	0.81	0.86
3	0.72	0.84	0.91
4	0.76	0.86	0.92
5	0.73	0.85	0.92
6	0.68	0.82	0.90
7	0.71	0.83	0.88
8	0.75	0.83	0.90
9	0.74	0.85	0.89
10	0.68	0.82	0.87
11	0.73	0.86	0.92
12	0.72	0.84	0.88
13	0.72	0.85	0.91
14	0.75	0.86	0.92
15	0.74	0.84	0.89
16	0.69	0.80	0.86
17	0.70	0.82	0.88
18	0.72	0.84	0.91
19	0.73	0.84	0.90
20	0.71	0.86	0.92
Average:	0.72	0.84	0.90

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 Algorithms 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.

To meet the computational efficiency requirements of real-time hydraulic state estimation, the urban network model which is comprised of over ten thousand pipelines and nodes is reduced using a water system aggregation technique.

The reduced model, which resembles the original system's hydraulic performances with high accuracy, simplifies the computation of the PC loop and facilitates the implementation of the on-line model.

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.

5 ACKNOWLEDGMENT

This work has been supported by the National Research Foundation of Singapore (NRF) and the Singapore – MIT Alliance for Research and Technology (SMART) through the Center for Environmental Modeling and Sensing.

6 REFERENCES

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.

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

Hammerlin, G., and Hoffman, K. H. (1991). "Numerical mathematics."*Graduate texts in mathematics*, Springer, New York.

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

Ostfeld A. et al. (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.

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

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.

Ulanicki, B., Zehnpfund, A., and Martinez, F. (1996). "Simplification of water distribution network models." *Proc., 2nd Int. Conf. on Hydroinformatics*, Zurich, Switzerland, 493–500.

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.