

Water Utility Case Study of Real-Time Network Hydraulic and Water Quality Modeling Using EPANET-RTX Libraries



Water Utility Case Study of Real-Time Network Hydraulic and Water Quality Modeling Using EPANET-RTX Libraries

FINAL

James Uber^{*}, CitiLogics

Sam Hatchett^{*}, CitiLogics

Stu Hooper^{*}, CitiLogics

Dominic Boccelli[†], University of Cincinnati

Hyoungmin Woo[†], University of Cincinnati

Robert Janke[±], USEPA

** CitiLogics, 615 Madison Avenue, Covington, KY 41011*

†University of Cincinnati, 705 Engineering Research Center, Cincinnati, OH 45221

±U.S. Environmental Protection Agency, 26 W Martin Luther King Dr., Cincinnati, OH 45268

Contents

List of Figures	iii
List of Tables	vi
List of Abbreviations & Acronyms.....	vii
Acknowledgments.....	viii
Disclaimer	viii
Executive Summary	ix
1.0 Introduction.....	1
1.1 Previous Work	2
1.2 Study Goals.....	3
1.3 Document Organization	4
2.0 Field Study Description	5
2.1 Study Area Distribution System Infrastructure.....	5
2.2 Real-Time Data Streams: Measurements, Boundary Conditions, and Key Assumptions	9
2.3 SCADA Data Quality	12
2.4 Tracer Field Study Description	22
2.5 Calcium Chloride Injection Protocol	27
3.0 Real-Time Modeling Using EPANET-RTX.....	28
3.1 Real-Time Simulation Process.....	29
3.2 EPANET-RTX Real-Time Model Configuration and Data Transformations	30
3.2.1 Tank Level Data Streams	30
3.2.2 Pressure Data Streams.....	32
3.2.3 Pump Status Data Streams	33
3.2.4 Pipe Flow Data Streams.....	36
3.2.5 Altitude Valve Data Streams	38
3.2.6 Reconstruction of Missing Plant Production Data Stream.....	39
3.3 District Metered Area Real-Time Demands	42
3.3.1 DMA Demand Time Series Data Transformation Pipeline	42
3.3.2 DMA Demand Disaggregation	45
4.0 Real-Time Model Calibration	45
4.1 Calibration Process	45
4.2 Real-Time Hydraulic Simulation	46
4.3 Real-Time Simulation of Tracer Movement	61
4.3.1 Tracer Data Processing	62

Water Utility Case Study Using EPANET-RTX Libraries

4.3.2 Real-Time Water Quality Model	63
4.3.3 Accuracy Metrics	67
4.3.4 Observed and Simulated Tracer Signals	68
4.3.5 Summary Results	86
5.0 Water Security Application: Demonstration of Using Real-Time Water Quality Simulation Results for Contamination Detection.....	90
5.1 Evaluation of Real-Time Model-Based Event Detection Using the NKWD Tracer Study Data.....	92
6.0 Outcomes	97
6.1 Identification of Infrastructure Model Errors.....	98
6.2 Improved NKWD Model	98
6.3 Identification of Potential Valve Failure or SCADA Sensor Problem.....	98
6.4 Demonstration of SCADA Data Evaluation, Analysis, and Use in Real-Time Modeling.....	99
6.5 Demonstration of Using Real-Time Water Quality Simulation Results for Contamination Detection	99
6.6 Steps for Developing a Real-Time Model and Implementing Real-Time Analytics	100
6.7 Potential Barriers to Real-Time Implementation	101
7.0 Conclusions.....	103
References.....	105
Appendix A: Prediction Accuracy of Chloride Levels Based on Measured Specific Conductance	A-1
Appendix B: A Catalog of Operational Notes and Network Model Updates for the Northern Kentucky Water District (NKWD) System	B-1
B.1 Operational Notes.....	B-1
B.2 Model Updates	B-1
Appendix C: Recommendations and Open Questions	C-1

List of Figures

Figure 2.1-1. Distribution system study area map showing supply infrastructure, pressure zones, district metered areas, and categorized real-time data streams.	7
Figure 2.3-1. Maximum time gap (minutes) between valid measurements for all flow measures data streams, for each day from October 1, 2012, through December 31, 2012.	15
Figure 2.3-2. Maximum time gap (minutes) between valid measurements for all tank level measure data streams, for each day from October 1, 2012, through December 31, 2012.	18
Figure 2.3-3. Maximum time gap (minutes) between valid measurements for all SCADA pump runtime measure data streams, for each day from October 1, 2012, through December 31, 2012.	21
Figure 2.4-1. Tracer study area description illustrating Conductivity Monitoring Areas A, B, C, D, E, and F.	23
Figure 2.4-2. Typical monitor setup.....	26
Figure 2.4-3. Typical components of the monitor (e.g., conductivity sensor, display, logger, piping and tubing).	26
Figure 3.1-1. Prototype EPANET-RTX application code (C++) for executing real-time simulation using an EPANET-RTX configuration file.	29
Figure 3.2-1. Representative raw storage tank water level data with 1st and 2nd moving average filters.	32
Figure 3.2-2. Time series data transformation pipeline for resampling and smoothing raw SCADA tank level data.	32
Figure 3.2-3. Representative raw pressure data with moving average filter.....	33
Figure 3.2-4. Time series data transformation pipeline for resampling and smoothing raw SCADA pressure data.....	33
Figure 3.2-5. Time series data transformation pipeline using derivative and threshold to derive binary pump status from raw SCADA pump runtime data.	34
Figure 3.2-6. Representative cumulative pump runtime calculated from raw SCADA pump runtime data, as well as three pump status data streams derived from runtime data.	35
Figure 3.2-7. Time series data transformation pipeline using special-purpose RuntimeStatus class to derive binary pump status directly from raw SCADA pump runtime data.....	35
Figure 3.2-8. Raw pump station flow SCADA data (gallons per minute) and trimmed/smoothed station flow (left axis), along with station status (right axis) used to produce the trimmed data stream.....	37
Figure 3.2-9. Time series pipeline for trimmed and smoothed pump station flow measures.	38
Figure 3.2-10. Raw and smoothed tank level data (left axis), along with altitude valve status (right axis) used to model control action of altitude valves on specific tank inlet/outlet lines.....	39
Figure 3.2-11. Time series data transformation pipeline to determine status of tank inlet/outlet pipe with altitude control valve.....	39
Figure 3.2-12. Treatment plant clearwell schematic showing flow and level measures used for construction of boundary flow data stream.	40
Figure 3.2-13. Time series data transformation pipeline to determine flow measure at northern treatment plant, from conservation of fluid volume within clearwell.....	41
Figure 3.2-14. Representative results from the associated time series data transformation pipeline, including the total Actiflo flow ($F_a + F_b$), the clearwell net inflow (dV/dt), and the resulting supply flow (F).....	41
Figure 3.3-1. Time series data transformation pipeline constructed automatically by EPANET-RTX to aggregate boundary flows (DMA 3 demand).	43
Figure 3.3-2. Real-time aggregate demand (gallons per minute) for DMAs 1 through 3 for November 19 through 26, 2012 (top), and expanded for a 1-day period within the same time frame (bottom).	44

Figure 4.2-1. Pearson's correlation coefficients between measured and real-time simulated heads, flows, and tank levels.....	48
Figure 4.2-2. Measured and real-time model simulated heads.	50
Figure 4.2-3. Measured and real-time model simulated heads.	51
Figure 4.2-4. Measured and real-time model simulated heads.	52
Figure 4.2-5. Measured and real-time model simulated heads.	53
Figure 4.2-6. Measured and real-time model simulated heads.	54
Figure 4.2-7. Measured and real-time model simulated flows.....	55
Figure 4.2-8. Measured and real-time model simulated flows.....	56
Figure 4.2-9. Measured and real-time model simulated flows.....	57
Figure 4.2-10. Measured and real-time model simulated tank levels.	58
Figure 4.2-11. Measured and real-time model simulated tank levels.	59
Figure 4.2-12. Measured and real-time model simulated tank levels.	60
Figure 4.2-13. Measured and real-time model simulated tank levels.	61
Figure 4.3-1. Illustrative results from data processing applied to raw tracer data.	62
Figure 4.3-2. Conductivity boundary condition at the injection site used for tracer simulations.	64
Figure 4.3-3. Evidence used to determine elevated background conductivity at northern treatment plant.....	66
Figure 4.3-4. Tracer time series characteristics used for error analysis.	68
Figure 4-5. Observed.3 and simulated tracer movement, Locations A2 (top), A3 (middle), and A4 (bottom).	69
Figure 4.3-6. Observed and simulated tracer movement, Locations A5 (top), A6 (middle), and A7 (bottom).	70
Figure 4.3-7. Observed and simulated tracer movement for Location A8.....	71
Figure 4.3-8. Observed and simulated tracer movement, Locations B1 (top), B2 (middle), and B3 (bottom).	72
Figure 4.3-9. Observed and simulated tracer movement, Locations B5 (top), B7 (middle), and B9 (bottom).	73
Figure 4.3-10. Observed and simulated tracer movement, Locations C1 (top), C2 (middle), and C3 (bottom).	75
Figure 4.3-11. Observed and simulated tracer movement, Locations C4 (top), C6 (middle), and C7 (bottom).	76
Figure 4.3-12. Observed and simulated tracer movement, Location C8.....	77
Figure 4.3-13. Observed and simulated tracer movement, Locations D1 (top), D2 (middle), and D3 (bottom).	78
Figure 4.3-14. Observed and simulated tracer movement, Locations D4 (top), D6 (middle), and D7 (bottom).	79
Figure 4.3-15. Observed and simulated tracer movement, Location D8.	80
Figure 4.3-16. Observed and simulated tracer movement, Locations E1 (top), E2 (middle), and E3 (bottom).	81
Figure 4.3-17. Observed and simulated tracer movement, Locations E4 (top) and E6 (bottom).	82
Figure 4.3-18. Observed and simulated tracer movement, Locations F1 (top), F2 (middle), and F3 (bottom).	84
Figure 4.3-19. Observed and simulated tracer movement, Locations F4 (top), F5 (middle), and F7 (bottom).	85
Figure 4.3-20. Comparison of tracer data and real-time simulations at 37 monitoring locations, as a function of pipe diameter or location characteristic.....	90

Figure 4.3-21. Comparison of tracer data and real-time simulations at 37 monitoring locations, as a function of location region (A-F).....	90
Figure 5.1-1. Histogram of detection threshold ratio, T_m^*/T^* , for all 38 tracer monitoring locations.	95
Figure 5.1-2. Illustrated event detection results for three different locations.	96
Figure 6.3-1. Identification of potentially excessive water pumping (over a long period of time) at partnering utility.....	98
Figure A-1. Relationship between specific conductance and chloride measurements.....	A-1

List of Tables

Table 2.2-1. Summary of measurement and boundary data streams used within the NKWD study area. .	11
Table 2.3-1. SCADA flow tags and indices.	13
Table 2.3-2. SCADA tank level tags and indices.	17
Table 2.3-3. SCADA pump runtime tags and indices.	19
Table 2.4-1. Characteristics of selected monitoring locations.	25
Table 3.3-1. Summary of boundary elements for DMA 3 demand time series aggregation.	43
Table 4.3-1. Characteristics of simulated and observed tracer time series data.	86
Table 4.3-2. Differences between simulated and observed tracer time series data.	88
Table 5.1-1. Minimum event detection thresholds for filtered measurement signals (T^*), filtered prediction error signals (T_m^*), and detection threshold ratio (T_m^*/T^*).	94
Table B-1. Summary of structural model changes.	B-2
Table B-2. Summary of parametric model changes or confirmation.	B-3

List of Abbreviations & Acronyms

AWWA	American Water Works Association
CAAB	conductivity area above background
CWS	contamination warning system
DMA	district metered area
DTW	dynamic time warping
EC	electrical conductivity
EPANET-RTX	EPANET's "Real-Time eXtension" open source software libraries
FPM	feet per minute
FTTP	Fort Thomas Treatment Plant (southern TP)
GIS	geographic information system
GPM	gallons per minute
GUI	graphical user interface
IQR	inter-quartile range, measure of time spread of pulse
LIDAR	light detection and ranging — a remote sensing technology that measures distances by illuminating a target with a laser and analyzing the reflected light
MGD	million gallons per day
MPTP	Memorial Parkway Treatment Plant (northern TP)
MySQL	open source, community edition of structured query language (SQL) database
NHSRC	National Homeland Security Research Center
NKWD	Northern Kentucky Water District
NSF	National Sanitation Foundation
OPC	online process control
PRV	pressure reducing valve
Q ₁	Time when 25% of the tracer pulse signature has passed the sensor
Q ₂	Time when 50% of the tracer pulse signature has passed the sensor
Q ₃	Time when 75% of the tracer pulse signature has passed the sensor
RMSE	root-mean-squared error
SCADA	supervisory control and data acquisition
SQL	structured query language
TDS	total dissolved solids
TMTP	Taylor Mill Treatment Plant
TOC	total organic carbon
TP	treatment plant
UK	United Kingdom
USEPA	United States Environmental Protection Agency

Acknowledgments

The U.S. Department of Health and Human Services' Centers for Disease Control and Prevention was instrumental in helping to secure the necessary field equipment for the tracer study. Personnel from the Northern Kentucky Water District provided many hours of their time, along with full access to information that was critical for making this project possible.

Stu Hooper (1970-2014)

We echo the words of Stu's colleagues at CitiLogics, *"His absence is deeply felt, but his imprint continues to be at the very core of what we do."*

Disclaimer

The United States Environmental Protection Agency through its Office of Research and Development funded and managed the research described here. Funding for this research was provided by the EPA (contract EP-C-10-060; Work Assignment WSD 2-29 "Field Demonstration of a Real-Time Water Infrastructure Monitoring and Data Fusion Technology to Improve Operations and Enhance Security of Water Systems"). EPA funds came from the National Homeland Security Research Center and the Water Technology Innovation Cluster. The U.S. Department of Homeland Security's Science and Technology Directorate provided funds that supported the tracer study through a Technology Development and Deployment Program managed by The National Institute for Hometown Security, under an Other Transactions Agreement, OTA #HSHQDC07300005, Subcontract #0210UK. This report has been reviewed by the Agency but does not necessarily reflect the Agency's views. No official endorsement should be inferred. EPA does not endorse the purchase or sale of any commercial products or services. Mention of trade names or commercial products does not constitute endorsement or recommendation for use. Neither the United States Government, nor any of their employees, makes any warranty, express or implied, including the warranties of merchantability and fitness for a particular purpose, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product or process disclosed or represents that its use would not infringe privately owned rights.

Executive Summary

The U.S. Environmental Protection Agency (USEPA) National Homeland Security Research Center (NHSRC) has developed an object-oriented software library called EPANET-RTX (the EPANET “Real-Time eXtension”) that comprises the core data access, data transformation, and data synthesis (modeling) components of a real-time hydraulic and water quality modeling system. EPANET-RTX was released as an open source software project on September 24, 2012, to advance real-time modeling capabilities.

In this report we provide a comprehensive description of the development and performance of an EPANET-RTX-based real-time hydraulic and water quality network model, including a description of the data processing steps, and an evaluation of model accuracy using all available operational supervisory control and data acquisition (SCADA) data streams in a complex real distribution system. We describe a field scale evaluation of a real-time hydraulic and water quality model of the Northern Kentucky Water District (NKWD). The NKWD system is a complex system serving approximately 81,000 customer accounts, or nearly 300,000 people. The work described here, however, is not meant to be complete, but rather only illustrative of the insight and value that can be obtained from the fusion of a network model with SCADA data assets.

Our field demonstration includes a one-week evaluation period where real-time model simulations are compared to SCADA operational data and calcium chloride tracer data. We fully describe a field tracer experiment in the NKWD system. The water quality model field experiment described is one of a few distribution system water quality studies to follow a large volume of finished water through an extensive portion of the distribution system. Our study is the first study to specifically use real-time modeling to drive the tracer simulations, and thus evaluate the fidelity of real-time simulation data processing techniques. Our study design represented a challenging test of model accuracy, as 24 of 38 monitors were located on small diameter distribution mains (17) or dead-end mains (7); thus our test not only evaluated the ability of a real-time model to predict movement through transmission mains, but also evaluated the accuracy at a neighborhood scale.

Our real-time simulation results were fully automated by EPANET-RTX data processing algorithms, and they prove the feasibility of calculating accurate real-time simulations for complex distribution systems. We find correlation coefficients averaging approximately 0.80 for flows, pressures, and tank levels for the NKWD study area. We present utility case study demonstration results without the use of complex micro-calibration of system parameters. That is, our real-time hydraulic simulation results are demonstrated and shown to be sufficiently accurate that water utilities should be able to now investigate improving their existing work flows and designing new ones to achieve desired endpoints, such as improved operations and water quality management, emergency preparedness, or water loss determination.

Our report fills a needed gap in understanding the methods that can be used to connect SCADA operational data to a network distribution system model. By describing and implementing a real-time network modeling process on the NKWD network model, we determine and present results on the accuracy of the NKWD hydraulic and water quality model.

We describe a set of outcomes that resulted from the development and application of the EPANET-RTX technologies to NKWD’s network model and SCADA data assets. We define an Outcome as a specific deliverable provided to our partnering utility (e.g., improved network model) or a finding, strategy, or product provided to the wider water community to address a need or to demonstrate a useful result that could be obtained with real-time modeling. The following outcomes are described:

- We demonstrate the application of the EPANET-RTX technologies on a large and complicated water distribution system to support the refinement and calibration of the utility's hydraulic and water quality model. We provide as a product to NKWD the identification of infrastructure model errors and an improved water distribution system network model. For improving the NKWD hydraulic and water quality model, we provide a list of significant recommendations and existing open questions that were generated through the development of the real-time model. While this list is not complete, it is illustrative of the insight and value that can be obtained from the fusion of a network model with SCADA data assets. Our list of issues and recommendations includes issues related to SCADA, model infrastructure and operations data, and real-time model configuration.
- We provide an example of the value and benefit that can be obtained from a real-time model and data fusion application. While the investigation is ongoing, we provide a discussion of how the application of the EPANET-RTX technologies worked to identify a potential critical valve failure or sensor failure in the NKWD distribution system. If the critical valve has indeed failed, excessive pumping electricity costs are estimated at \$80,000.00 per year.
- We summarize three critical findings from this research with respect to using SCADA data assets for real-time modeling and simulation:
 - Proof of ability to process ordinary/raw SCADA operational data streams to determine accurate hydraulic statuses and metrics for pumps, tanks, and flows in the distribution system
 - Demonstration of succinct methods for the processing of time series data and thereby transforming raw SCADA data streams into useful model outputs
 - Proof of scalability of EPANET-RTX-based real-time modeling tools without custom programming to enable real-time predictions and forecasts for any water system that has a suitable infrastructure network model and sufficient SCADA data assets
- We demonstrate through a detailed analysis of NKWD case study results, a water security application for real-time modeling. We demonstrate the use of real-time water quality simulation results for contamination detection.
- We provide a five-step approach to water utilities for using the EPANET-RTX technologies to develop a real-time model and implement real-time analytics using SCADA data assets.
- We provide a discussion of some potential barriers to real-time model development and implementation for the water community.

1.0 Introduction

Water utilities have invested heavily in data and information infrastructures. Supervisory control and data acquisition (SCADA) systems support operational decisions, and geographic information systems (GIS) and infrastructure models support infrastructure planning. These investments need to be further leveraged to support a wider scope of utility decision making to address the multitude and complexity of problems facing water utilities across the nation. Gigabytes of SCADA data representing years of pressure, flow, tank level, pump status, and water quality time series are stored in a typical historian database, and never accessed. Divorced from these data, infrastructure models are limited in helping to interpret them for useful operational goals.

The fusion of real-time operational data with infrastructure-aware predictive models will yield numerous practical benefits, enabled by the ability to simply and accurately forecast distribution system hydraulics and water quality in real time. Operators should be able to routinely engage in situational response training, and conduct operational analyses to achieve optimization goals related to pressure, leakage, energy, and water quality management just as a pilot uses a flight simulator. Engineers should be able to apply their infrastructure knowledge to these same tasks in a collaborative fashion, while knowing their infrastructure models are continuously updated through a persistent interpretation of the operational record enabling automatic estimation of water usage, operating rules, and pump head-discharge curves. Managers should be able to review automated periodic reports showing trends in unaccounted-for water, energy usage, and water quality, and integrate those with past and future asset management decisions. These capabilities and benefits are not unrealistic. In fact these capabilities and their benefits are already supported by existing investments in SCADA, GIS, modeling, and by network hydraulic theory that is hundreds of years old.

What has been missing is a clear understanding of the methods by which the operational data can be connected with network models, and the resulting accuracy of network simulation models that can be achieved when they are driven by operational data. Absent this understanding, there will continue to be skepticism about the ability of real-time processes to transform raw SCADA data into data streams that can accurately model water demand, as well as the operational control decisions routinely made by system operators (or automatic control algorithms). There will continue to be skepticism about the ability of network models, which were originally developed to support master planning, to provide meaningful predictions that reflect particular system operational decisions.

This report fills the gap in the understanding of the methods that can be used to connect operational data to a network model. This report describes a field-scale evaluation of a real-time hydraulic and water quality model of the Northern Kentucky Water District (NKWD). The data used include both those routinely collected through the District's SCADA system, as well as field test data collected from injecting and tracking a series of salt pulses in a large section of the distribution system obtained during a field tracer test in November 2012. By describing and implementing a real-time network modeling process on the NKWD network model, the resulting hydraulic and water quality model accuracy for the NKWD model is determined.

Tracer tests are the preferred method for calibration and validation of network hydraulic and water quality models. For instance, tracer tests are typically used to calibrate and test models to predict chlorine residual and trihalomethane formation. In addition, the transport of salt pulses can mimic the temporal signatures of contaminants intentionally introduced into a distribution system; thus these data may be used to evaluate whether real-time data analytics and models can enhance water security through support of event detection and emergency response.

For this study, the field tracer test included injecting a calcium chloride (CaCl_2) solution into the distribution system as a series of four pulses over a 12-hour period. The movement of the CaCl_2 pulses was observed by 38 continuous specific conductance monitors located in the distribution system to provide information about the passage of the pulses at high spatial and temporal resolution. Operational data from the SCADA system was used to drive a real-time network hydraulic model, which underlay the real-time water quality model. This report describes the development and accuracy of the real-time hydraulic model first, followed by the development and testing of the water quality model.

The water quality model field experiment described here is one of a few distribution system water quality studies that attempted to follow a large volume of finished water through an extensive portion of the distribution system. Tracer data provide unique information about processes that affect water quality in the distribution system, including water velocity, junction mixing, and flow path-dependent effects, and can be used to evaluate the real-time network hydraulic and water quality models. This report details the first such study to specifically use real-time modeling to drive the tracer simulations, and thus evaluate the fidelity of real-time simulation data processing techniques. The study design represents a challenging test of model accuracy, as 24 of 38 monitors were located on small diameter distribution mains (17) or dead-end mains (7); thus the test is not only evaluating the ability of a real-time model to predict movement through the transmission main infrastructure, but is also evaluating the accuracy at a neighborhood scale.

1.1 Previous Work

The use of SCADA data in water distribution system model development and calibration is not novel. A SCADA database can contain years of data for hundreds to thousands of relevant instruments (such as pumps, valves, and tanks), at sub-minute resolution — and the availability of these data is widely known. Published model calibration studies and the standard practice in the field (e.g., Walski, 1983) is fairly dated and remains focused on sparse, manually collected data sets for a small observation time window, generally combined with a limited range of SCADA data. This typical use of SCADA data requires cumbersome batch workflows involving manual database queries, distinct software packages for data access, transformation, and synthesis, and multiple disparate data formats. It is easy to understand why SCADA data can be viewed by practitioners as “difficult to use,” even if there is clear motivation to leverage a running SCADA system’s abundant data resources. Further, research that relies on a continuous stream of SCADA data, (e.g., testing a state estimation methodology, as in Davidson and Bouchart, 2006; Kang and Lansey, 2009; Shang et al., 2006), is usually based on synthetic data with random noise superimposed to simulate real SCADA measurements.

Examples of SCADA-model fusion do exist, but they are usually burdened by many intermediate steps between data access and synthesis, or by the closed philosophy of proprietary software systems (Bartolin et al., 2006; Johnson et al., 2007). Real-time data fusion promises a persistent connection between network model and SCADA database with automated data transformation and synthesis. Software titles claim the ability to use real-time data, but generally, in reality, support only a batch-oriented import of SCADA information, requiring the export of a dataset as a text file, with off-line processing. Such data connectivity can clearly be useful, but falls far short of the goal for real-time data fusion. Proprietary commercial software programs are also usually derived from design-oriented (i.e., “off-line”) hydraulic modeling software, and seem destined to carry the limitations of those software environments into the real-time realm (e.g., batch-oriented data processing, and complex user interfaces ill-suited for real-time operational analysis). Finally, a significant limitation of all current methods of real-

time network model and data integration is the sole focus on system hydraulics. Water quality issues have not yet been integrated with real-time hydraulic models and SCADA water quality data.

Recognizing that new software systems are needed to support real-time fusion of SCADA data and network models, the U.S. Environmental Protection Agency (USEPA) National Homeland Security Research Center (NHSRC) has developed an object-oriented software library called EPANET-RTX (the EPANET “Real-Time eXtension”), which comprises the core data access, data transformation, and data synthesis (modeling) components of a real-time hydraulic and water quality modeling system (Hatchett et al., 2011; Rossman, 1999). EPANET-RTX was released as an open source software project on September 24, 2012, to support commercialization opportunities. It is intended that EPANET-RTX would become a unifying bridge between data and model, and thus overcome many of the above obstacles, eventually helping to spur the development of real-time modeling software applications for industry.

We define real-time modeling as an integration of network hydraulic and water quality models with operations data collected and stored via SCADA, providing for an automated and routine capability to hind-cast, now-cast, and forecast complete system pressures, flows, and water quality, in support of operational goals, such as emergency response and water system planning goals. The methods for real-time data collection, storage, and analysis can be described as real-time analytics. While EPANET-RTX consists of a set of object libraries to facilitate water distribution analytics (hydraulics and water quality) by providing software to handle database connections, SCADA data cleaning and filtering, data analysis, and real-time distribution system simulations, it more importantly represents a fundamentally new approach to distribution system modeling. EPANET-RTX is termed an extension to EPANET because it utilizes the underlying hydraulic and water quality simulation engines of EPANET, but it is probably more aptly described as “Real-Time EPANET.” The practical benefit of a water distribution system model is its ability to reasonably approximate the behavior of the system in question. Practical applications could include improved operations and water quality management and emergency preparedness. While such benefits are certainly reasonable and intuitively necessary, there has until now been little technological advancement in distribution system modeling tools to enable the water utility engineer to efficiently test and demonstrate that their water distribution system model accurately represents system behavior. “Real-Time EPANET” changes this paradigm by adding to EPANET the necessary capabilities to gather SCADA data (continuous real-time data), analyze it (clean and filter), and provide the means to effectively utilize it for any needed or desired applications.

The real-time modeling results demonstrated here constitute the first full-scale case study of the EPANET-RTX technology. EPANET-RTX promises to be a market-making technology by creating an application development framework enabling researchers, consultants, and commercial software vendors to develop better tools to support real-time analytics in drinking water distribution systems.

1.2 Study Goals

The main goal of this study is to establish the modeling accuracy that can be achieved through real-time network models, using an existing network model developed to support master planning, and an existing SCADA database implemented to support system operations. While generalizations of accuracy cannot be made, the results here do provide a significant benchmark, based on established software systems and data transformation procedures. Further, since decisions about data transformation methods will affect real-time modeling accuracy, this study aims to expose and document those decisions, to give subsequent studies both a starting point and likely opportunities where improvements can be made. In other words, we aim to document our decisions about the real-time model configuration and calibration,

while not making any claim that those decisions are optimal. Indeed, we consider the results presented to be representative of a significant initial effort.

1.3 Document Organization

Section 2 provides a description of the NKWD water utility, its distribution system infrastructure, available real-time (SCADA) data streams, an evaluation of SCADA data, and a description of the calcium chloride tracer test that was conducted.

Section 3 begins with a brief description of real-time modeling using the EPANET-RTX technologies and is followed in Section 3.1 with a description of the “Real-Time Simulation Process” and in Section 3.2 with an in-depth description of how a real-time model is configured in “Real-Time Model Configuration.” The heart of configuring a real-time model using EPANET-RTX is described by the SCADA data transformations, contained in Sections 3.2.1 (Tank Level Data Streams) through 3.2.6 (Reconstruction of Missing Plant Production Data Stream). Section 3.3 describes how district metered area (DMA) real-time water demands are automatically determined using EPANET-RTX. Section 3.3.1 (DMA Demand Time Series Pipeline) describes how EPANET-RTX constructs DMAs using an algorithmic process to identify the boundary pipes associated with each distinct DMA. Section 3.3.2 (DMA Demand Disaggregation) describes how real-time demands are disaggregated to the junctions within each DMA.

Section 4 describes the real-time model calibration process used. Section 4.1 describes the calibration process and Section 4.2 describes the real-time hydraulic simulation process. Section 4.3 is subdivided into five sections (4.3.1 through 4.3.5). Section 4.3.1 describes how the tracer conductivity data were processed for use in the real-time hydraulic and water quality model evaluations. Section 4.3.2 describes the real-time water quality model simulation process performed on the NKWD distribution system model. Section 4.3.3 describes the accuracy metrics employed to evaluate the real-time model’s simulation results; and Section 4.3.4 presents the time series data for the observed and simulated conductivity signals from each of the monitoring sites, by region, that provided conductivity data. Finally, Section 4.3.5 summarizes the results from the comparison of the EPANET-RTX simulations to the conductivity data collected during the field study.

Section 5 demonstrates the use of real-time water quality simulation results in contamination detection. Section 6 provides an itemized discussion of the outcomes that resulted from applying the EPANET-RTX technologies to the NKWD model and system. Section 7 presents the case study’s conclusions. Appendices are provided for supporting information, remaining questions, and suggested recommendations for improving the network infrastructure model as well as the real-time model developed.

One final note, throughout this report we frequently use the term “model” in the context of either describing a “hydraulic model” or a “water quality model,” referring to the hydraulics or water quality, respectively, components of the distribution system model. Similarly, we frequently discuss the EPANET-RTX model in terms of either the EPANET-RTX hydraulic or EPANET-RTX water quality model, again the two components of a complete system model. When we refer to just “model” we are referring to the complete model, i.e., including the hydraulic and water quality components. Finally, we also refer to model in context of the “network model” or water distribution system infrastructure model. The network model is probably best described as the map of the distribution system, i.e., detailing locations and specifications of the distribution system components, e.g., pipes, pumps, and tanks.

2.0 Field Study Description

In Section 2 we describe the Northern Kentucky Water District service area along with the study area, representing a sub-region of the NKWD service area coinciding with the tracer study area conducted in November 2012. Real-time data streams are introduced and described in Section 2.2. The SCADA data quality evaluation process is described in Section 2.3. Section 2.4 provides a detailed description of the tracer field study, and Section 2.5 describes the calcium chloride protocol used to perform the tracer study.

2.1 Study Area Distribution System Infrastructure

The NKWD serves approximately 81,000 customer accounts, or nearly 300,000 people in Campbell and Kenton Counties; portions of Boone, Grant, and Pendleton Counties; and the Cincinnati/Northern Kentucky International Airport (located in Northern Kentucky). It covers over 300 mi² of total service area through 1,282 miles of distribution piping. Three water treatment plants — Fort Thomas Treatment Plant (FTTP), Taylor Mill Treatment Plant (TMTP), and Memorial Parkway Treatment Plant (MPTP) — have a combined capacity of 64 million gallons per day (MGD), and supply water through 16 high service and booster pump stations containing 43 pumps. Average water usage is approximately 28 MGD. Distribution system storage consists of nearly 27 million gallons distributed through 20 elevated storage tanks. Pressure regulation is achieved through the creation of 22 pressure zones by 33 regulating valves. The infrastructure model maintained and used by the utility includes all distribution piping — in excess of 13,500 individual pipes.

Figure 2.1-1 shows the study area, a sub-region of the NKWD service area east of the Licking River with a total demand of 7.48 MGD. The study area was selected to coincide with the tracer study area that was conducted during November 2012. The real-time hydraulic results presented and discussed here are combined with the tracer study results to drive the water quality predictions, providing a rigorous evaluation of real-time model accuracy.

The northern portion of the study area (within the bounding box in Figure 2.1-1) is shown in greater detail in Figure 2.1-2. In both figures, pipeline width is related to pipe diameter; pipes less than 8 in. in diameter are represented by the thinnest lines, while pipes greater than 16 in. in diameter are represented by the thickest lines, and between these limits there is a gradient of line width. The northern portion of the study area has a greater density of infrastructure and instrumentation, and is characterized by older residential and commercial properties. While the real-time hydraulic model was configured, and real-time data were processed, for the entire distribution system, real-time model calibration activities (described in Section 4) have been limited to the pictured sub-region, and thus only the sub-region results are discussed here. The study area consists of three hydraulically distinct regions, referred to as district metered areas (DMAs), and numbered 1 through 3 in Figure 2.1-1. The hydraulic characteristics of each study area DMA will be discussed and described further in Section 3.3; they are introduced here for convenience, and they also serve to define the study area. The boundaries of DMAs 2 and 3 coincide with the boundaries of two pressure zones, indicated in the figure by colored regions separated by white lines. DMA 2 is at a nominal head of 741 ft., and DMA 3 is at a nominal head of 965 ft. DMA 1, by far the largest in geographic area, includes 10 separate pressure zones within its boundary, although two of these zones include the bulk of the infrastructure — one to the extreme north, at 829 ft., and a single large pressure zone that dominates the remainder of DMA 1, at 1,017 ft. From a broad topographic perspective (although not shown in Figure 2.1-1), the study region is bordered by the Ohio River to the north and east and by the Licking River to the west, which drains into the Ohio and represents a ridge with watersheds

that drain into either the Ohio or the Licking. The northwest corner of DMA 2 is at the confluence of the Licking and Ohio Rivers, and is the low point within the study area.

Figure 2.1-2 shows the locations of the two treatment plants within the study area, represented in the network model by reservoirs (head boundaries). Production from the northern treatment plant (TP) supplies District Metered Area (DMA) 2 by gravity from the clearwell (temporary storage to allow mixing or contact time for disinfection), and then DMA 3 through booster pumping. MPTP (Memorial Parkway Treatment Plant, the northern TP) can also supply the northern portion of DMA 1 by high service pumping into the 1017 pressure zone; from the 1017 zone, flow through regulating valves serves lower pressure zones within all three DMAs.

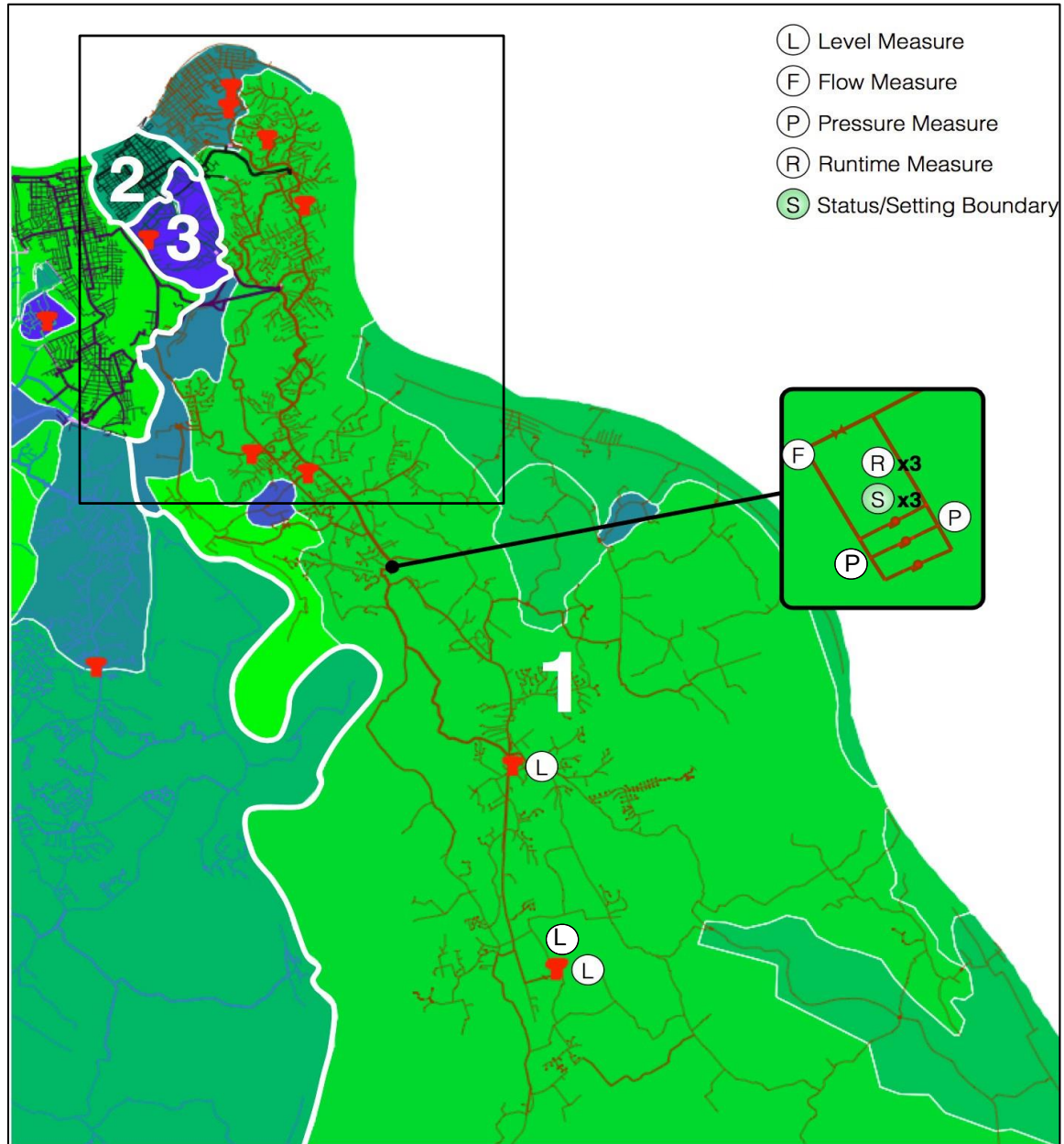


Figure 2.1-1. Distribution system study area map showing supply infrastructure, pressure zones, district metered areas, and categorized real-time data streams. Data streams within the bounding box are shown in Figure 2.1-2.

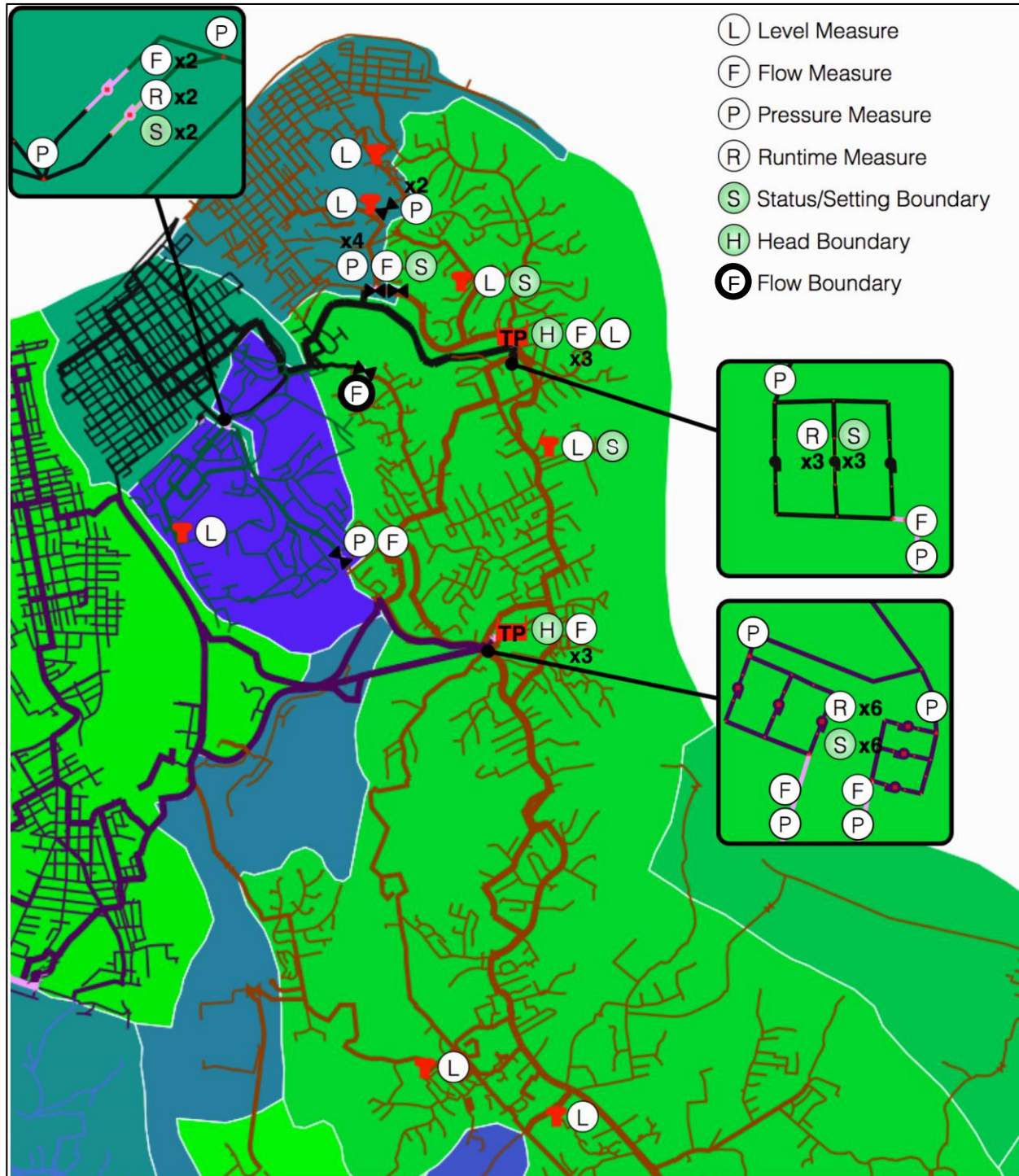


Figure 2.1-2. Northern portion of distribution system study area showing supply infrastructure, pressure zones, and categorized real-time data streams.

“Un-split” depends on the status of three valves near FTTP (Fort Thomas Treatment Plant, the southern TP). When these valves are closed, the 1017 zone is split into a northern and southern region, at approximately the location of the southern TP. In this configuration, the northern portion of the 1017 zone must be supplied by the northern TP, while the southern portion of the zone is supplied by high service pumping at the southern TP. When the valves are open, however, the 1017 zone is un-split, and the entire

1017 zone can be supplied completely by the two banks of high service pumps at the southern TP. Indeed, in the un-split configuration, the entire DMA 1 demand, as well as a portion of the demand in DMAs 2 and 3, is normally supplied by the southern TP and its high service pumps. As shown in Figure 2.1-1, additional booster pumping exists south of the southern TP, to supply water from that TP to a set of three tanks in the southern reaches of DMA 1. For all time periods analyzed here, the 1017 zone was un-split, and the high service pumps at the northern TP were always off (thus the northern TP is only supplying DMA 2 by gravity). The real-time model does not make any assumptions about pump status, instead getting its clues directly from the real-time SCADA information (as discussed in Section 3.2).

2.2 Real-Time Data Streams: Measurements, Boundary Conditions, and Key Assumptions

We distinguish two broad categories of real-time data streams: measurements and boundaries. Measurement data streams are used passively for comparison to simulation results, unlike boundaries, which are used actively to change on/off statuses, or setting values, of their associated model elements. This is a practical way to distinguish data streams according to their purpose for modeling, and not a way to uniquely categorize them. One data stream may serve as either a measurement or a boundary, depending on other factors — such as a pressure sensor downstream of a regulating valve, which could be used with equal justification as a setting boundary for the regulator, or as a measurement to compare with simulated pressure.

Figures 2.1-1 and 2.1-2 show the approximate locations of measurement and boundary data streams within the study area. Measurements are shown using open circles with a single letter indicating the type: level measure (L), flow measure (F), pressure measure (P), and runtime measure (R). Boundaries are similarly shown using filled circles: pipe, valve, or pump on/off status, and status/setting boundary (S), reservoir head (H), and flow boundary (F). The purpose is to illustrate the categories and locations of measurement data streams that are used for assessing simulation results, and of boundary data streams that are used to specify model element statuses and settings. These data streams do not, however, always map directly into the raw SCADA data streams, and they give relatively little information about the various transformation steps required between any one SCADA and measurement or boundary data stream. Raw SCADA data typically require sampling, filtering, and other data transformations to be used as reliable real-time model boundary conditions (which could include pump/pipe status, valve setting, head boundary, flow boundary, or demand). Even SCADA data used purely as measurements can sometimes be resampled and filtered, to reduce noise and to focus on the comparison with the true signal. The data transformations performed on the SCADA data streams in order to render them acceptable for real-time modeling are described in Section 3.2.

In general, each storage tank has a level measurement; each pump has a runtime measurement and status boundary; each pump station has suction pressure measurements, discharge pressure measurements, and a station discharge flow measurement; and each TP has a flow measurement and head boundary. Two storage tanks also include status boundary data streams that are assigned to their inlet pipe. These tanks have altitude valves, and their open/closed status needed to be represented using status boundary data streams.

Five control valves that regulate pressure between the 1017 level and adjacent lower zones are instrumented; four include pressure measurements and four include flow measurements. One control valve regulating between the 1017 zone and the 741 zone (DMA 2) is actively controlled via SCADA, and its downstream pressure measure is also used as a valve setting boundary. In general it is not valid to use a downstream pressure measure alone as a pressure regulating valve setting, as it is necessary to ensure the valve is actively controlling pressure (e.g., through the stem position) before the downstream

pressure can be assumed to represent the setting. In particular, if the valve is closed, then using the downstream pressure as a setting boundary could give erroneous flows through the valve, as it only indicates the downstream zone pressure under the closed condition. Nevertheless there was no way to reliably determine the valve status from the operational record, and it was necessary to assume it to be active; otherwise, without representing the SCADA control of the valve, flow would occur continuously from the 1017 to the 741 pressure zone, such that it reversed flow into the reservoir representing the clearwell of the northern TP. As this simulated behavior clearly contradicts reality, both in terms of the clearwell outflow and the measured flow through the regulating valve, the decision was made to take liberties with setting the boundary for the regulating valve. If the real-time model were put into place for systematic use, it would be recommended that key regulating valves be instrumented for flow, pressure, and valve status.

One flow measure in Figure 2.1-2, associated with a regulating valve, is highlighted by an atypical measure symbol with a heavy border. That flow measure is one of the boundary flows defining DMA 2; without it, DMA 2 would become part of a larger DMA 1, and its real-time demand allocation would be altered accordingly (DMAs and real-time demand computation are discussed in Section 3). Unfortunately, the data for this flow measure exists in SCADA, but those data were missing or of bad quality. It was decided to retain this “flow measure” in the real-time model, and thus to retain DMA 2, by assigning an assumed flow equal to zero to this flow measure. There is no known data to justify a zero flow assumption — it only mimics the assumption made by utility staff, who have assumed zero flow through this regulator by setting its status to close in the hydraulic network model. Indeed, during a field investigation of regulating valve settings and statuses in 2010, the authors noted that this valve was open at the moment when upstream and downstream pressures were recorded, although it was not possible to quantify the flow rate. The rationale for assuming a zero flow measure centers on the importance of retaining DMA 2 for demand computations. DMA 2 contains a dense street grid, its demographics and land use are distinctly urban, and its demand, as well as that of the neighboring DMA 3, dominates the production demand from the northern TP. Retaining DMA 2 thus forces the logical connection between demand in that DMA and flow from the northern TP. Nevertheless, this flow assumption would make the real-time model more sensitive to any disturbances that would affect the true regulator flow, and it would be recommended that such critical flow measure data streams be restored so that all DMA demand computations are supported by valid data streams.

Table 2.2-1 summarizes the measurement and boundary data streams within the study area, organized by data stream category and the associated model element type. For practical reasons, perhaps the most important (at least the most common) flow boundary data streams are not shown in this table, or in the above figures — the nodal demands. Each node (junction, reservoir, or tank) of the network model is assigned to a flow boundary data stream equal to its share of the real-time demand, as computed for the node’s DMA. This bears mentioning only so it does not go unnoticed, as the demand flow boundaries would be a vital component of the data processing for any real-time model.

Table 2.2-1. Summary of measurement and boundary data streams used within the NKWD study area.

Category	Model Element	Number
Level Measure	Tank	10
Flow Measure	Pump Station	6
	PRV	2
	Source (Treatment Plant)	1
Pressure Measure*	Pump Station	9
	PRV	6
Runtime Measure	Pump	14
Status Boundary	Pump	14
	Altimeter Valve	2
Setting Boundary	PRV	1
Head Boundary	Reservoir	2
Flow Boundary		0
Total		67 [†]

PRV, pressure reducing valve which has an associated flow measure.

* One pressure data stream was from the water works pump station (WATER_PI501A) and the other was St. Therese PRV (TAG_NEW1_P1301A).

[†] Note that nodal demand flow boundaries are omitted.

For the study area, two pressure data streams were omitted because no data was available from SCADA during the time frame of interest.

There are actually 6 flow measures associated with the 2 treatment plants of interest. Specifically, 3 are associated with the southern TP and 3 are associated with the northern TP. The 3 flow measures at the northern TP were, however, used to create a flow balance around the clearwell and, thus, provided a calculated flow out of the plant, since there was no flow measure available. (More discussion is provided on the creation of the flow balance in Section 3.2.6.) Thus, these 3 flow measures associated with the northern TP are practically just 1 flow measure. The 3 flow measures at southern TP were, however, not used. This was because the total flow from southern TP splits between the study area, and other demand areas to the west of the Licking River. The amount of flow from the southern TP into the study area was already captured by the 2 flow measures at the southern TP's pump station (specifically pumps 1-3 and 4-6). The remaining flow that would be simulated would be affected by portions of the network that lie outside of the study zone, which were not subject to the same quality assurance procedures. Therefore, these 3 flow measures were eliminated from Table 2.2-1 since they do not affect the study area simulation. This accounts for the one source treatment plant flow measure in Table 2.2-1, which is associated with the northern TP.

For the study area and a thorough analysis of the network model and SCADA data, there could be a total of three pressure reducing valve (PRV) flow measures. However, only two PRV flow measures are indicated in Table 2.2-1. For one of the three potential PRV flow measures we assumed a value for it since no SCADA data was available and, hence removed it from Table 2.2-1. (This PRV flow measure is shown in Figure 2.1-2, as the heavy bordered "F" symbol.) The two PRV flow measures indicated in Table 2.2-1 represent the flow measures at the St. Therese and Memorial Newport regulators.

2.3 SCADA Data Quality

All SCADA data streams were inspected visually for obvious anomalies. Where obvious anomalies were present — including large data gaps or unusual noise characteristics — strategies were considered for addressing them through the data transformation process. Large data volumes, however, make it difficult to develop a straightforward and easily understandable assessment of SCADA data quality. Typical statistical metrics on the data values do not, for example, convey adequate information about the number of data points collected over some period of time. We adopted a visualization approach that allows important features of the data to be inspected and hopefully understood, for a significant time range. This approach has yielded more important and specific insights than relying solely on statistics computed for the various data streams. Our visualization approach was manually performed and the resulting visualization displays were developed outside the EPANET-RTX libraries.

The main data quality concerns are online process control (OPC) data quality indicators, temporal data density and data gaps, and outliers or other obviously false values. OPC data quality is stored along with SCADA point values and timestamps. EPANET-RTX automatically rejects data points that are invalid according to a mapping of OPC quality flags to a valid or invalid point status. Paying attention to the OPC data quality indicators can eliminate many points that otherwise would be labeled “outliers.” While real-time data processing standards do exist for the OPC quality indicators (e.g., OPC quality 192 equates to a good point value), site-specific mappings of these codes to either a good or bad point status may be needed.

Going beyond the OPC data quality indicators, it is useful to understand the character of key data streams in terms of data density and data gaps, and possibly also in terms of outliers. Visualization techniques designed for large data sets are a valuable way to gain insights into overall data quality. For visual analysis of SCADA data quality, the present analysis considered three important categories of SCADA data for real-time modeling: flow, tank level, and pump runtime. These SCADA data streams are required to calculate real-time DMA demands, and to set pump operational status boundaries. Thus they represent critical boundary conditions for the model and are more important than SCADA time series used only for model evaluation. Also, for this analysis we considered all data streams for the entire NKWD service area, deviating from the focus on the study area in order to gain a broader appreciation, perhaps, for overall SCADA data quality.

To visualize large data sets, data must be aggregated. Useful aggregation allows huge data sets — such as all flow SCADA tags over an entire year — to be visualized and compared. Key indicators for each SCADA data category were aggregated on a daily basis and visualized for a 3-month period from October 1, 2012, through January 1, 2013, as a color-mapped image. The key indicators can vary depending on the type of data, but each data stream was examined for measures of data density — specifically the total number of data points per some period of time, the maximum data gap (both computed on a daily basis) — as well as the mean value and inter-quartile range ($IQR = Q_3 - Q_1$) (also computed on a daily basis). The visual data analysis is described in more detail for each of the data categories below. We include only information about the maximum data time gap, as data continuity is a concern for any data stream, whereas indicators related to data value are expected to vary and so must be considered within their physical context.

The maximum data gap is visualized in Figure 2.3-1, for the 27 SCADA flow measures listed in Table 2.3-1. The integer index in Table 2.3-1 is used to identify each data stream in the data quality figure. The maximum time gap between data points was computed for each day, and the entire 3-month span for one data stream is represented by one column of the image. Thus the visual matrix in Figure 2.3-1 has

dimension 27×92; one matrix element for each data stream and each day. Moving from left to right changes the data streams from Indices 1 through 27, while moving from top to bottom changes the time by day from October through December. The color scale represents discretized bins of maximum time gap, ranging from black (0 to 15 minutes) to the lightest grey (exceeding 240 minutes [4 hours]); red indicates that no data were available for that data stream and day. Thus the “ideal” data quality, in this sense, would be uniform black across the entire image.

Table 2.3-1. SCADA flow tags and indices.

Flow SCADA Tag	Description	Index
FI25-3801	FTTP Finished Water Flow 1	1
FI25-3802	FTTP Finished Water Flow 2	2
FI25-3803	FTTP Finished Water Flow 3	3
WALT FI200	Walton Meter Pit Flow	4
BULL FI200A	Bullock Pen Meter Pit Flow 1	5
BULL FI200B	Bullock Pen Meter Pit Flow 2	6
BULL FI200C	Bullock Pen Meter Pit Flow 3	7
PEND FI200A	Pendleton #2 Meter Pit Flow 1	8
PEND FI200B	Pendleton #2 Meter Pit Flow 2	9
MEM FI302	Memorial New Regulator Flow	10
CHES FI200	Chesapeake Regulator Pit Flow	11
NEW1 FI301	St. Therese Regulator Flow	12
US27 FI500	US 27 1–3 Station Flow	13
US27 FI501	US 27 4–6 Station Flow	14
RICH FI500	Richardson Station Flow	15
TMHS FI500	Taylor Mill HS Station Flow	16
RIPP FI500	Ripple Creek Station Flow	17
BRS FI001	Bristow Pump Station Flow	18
LATO FI500	Latonia Station Flow	19
HAND FI500	Hands Pike Station	20
WATER FI500	Waterworks Station Flow	21
COVI FI500	W Covington Station Flow	22
DUD2 FI500	Dudley 1080 Station Flow	23
BROM FI500	Bromley Station Flow	24
DUD1 FI500	Dudley 1040 Station Flow	25
CARO FI500A	Carothers Rd. Pump Flow 1	26
CARO FI500B	Carothers Rd. Pump Flow 2	27

FTTP, Fort Thomas Treatment Plant

The maximum gap data shows that days with no data are to be expected, and there are often extended durations for some data streams when data is absent. Data Streams 4 through 7, as well as 11, are essentially absent from the record and thus were discarded.¹ Data Stream 11 is the flow through the regulating valve previously mentioned that prompted the assumption of zero flow, in order to establish the boundary for DMA 2. Data stream 21 also has significant data gaps, which is the discharge from the high service pumps for the northern TP; this gap, however, is coincident with the 1017 pressure zone being

¹ The significance of missing data on the resulting accuracy of the real-time model is uncertain.

“un-split,” when these high service pumps are not expected to be in service. The data exhibit the characteristics of “delta mode” storage, where new data points are only stored when a significant change in value occurs.

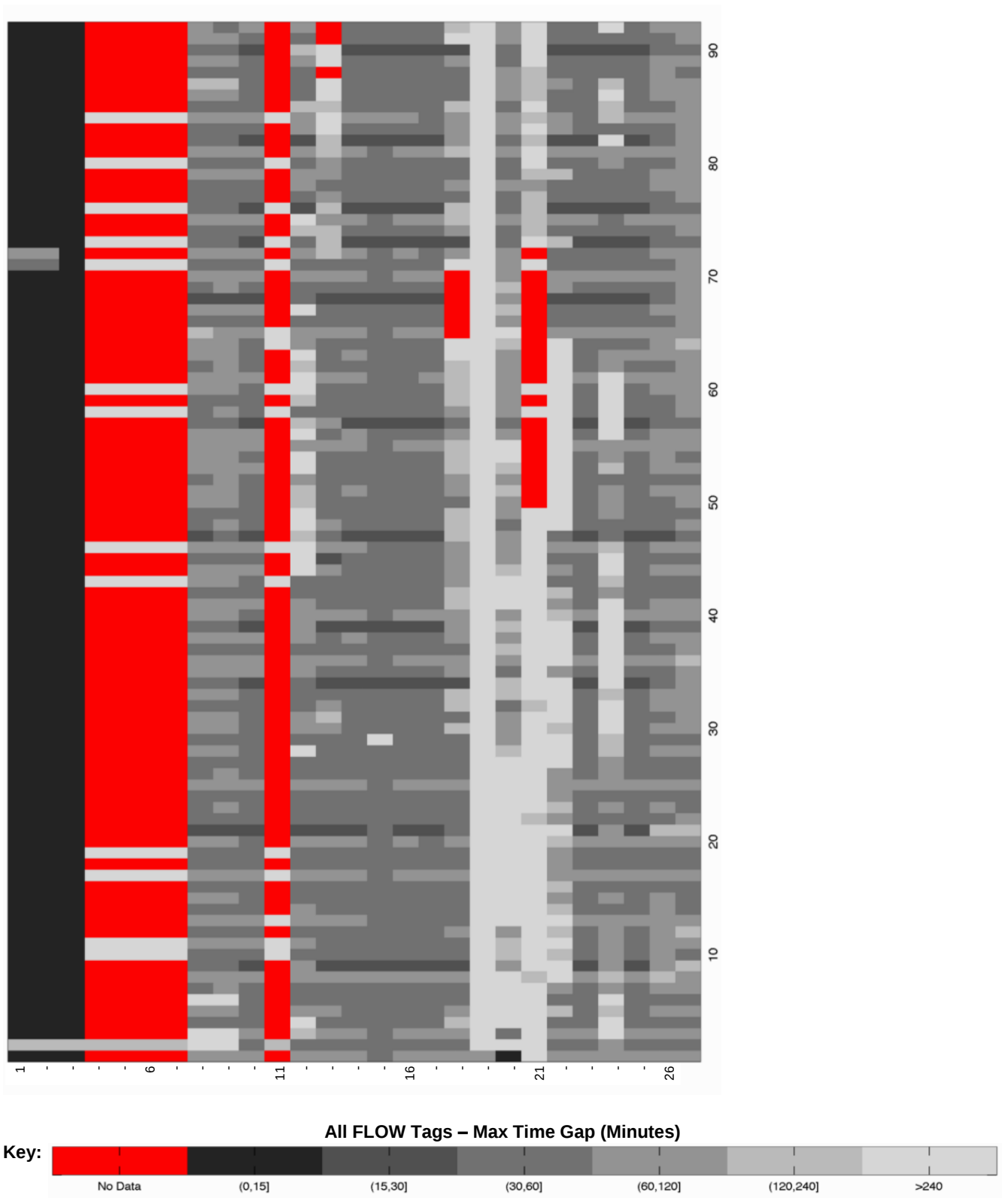


Figure 2.3-1. Maximum time gap (minutes) between valid measurements for all flow measures data streams, for each day from October 1, 2012, through December 31, 2012. Indices refer to SCADA tags in Table 2.3-1.

It seems from the data that pump flow 21 is zero for the entire duration given no data exists. It would be preferable for data quality assurance, however, if the delta-mode SCADA configuration allowed for a minimum data density (e.g., once per day). Aside from the streams with significant periods without any data, the data show that several other data streams exhibit periods when the maximum time gap exceeds 4 hours. Again, this variability in maximum gap size would be expected from a delta-mode storage configuration, and it would be useful to know with certainty the parameters of the data storage scheme (e.g., minimum value change that triggers storage of a new point), and how the parameters varied with the particular data stream. For example, reliable information about such storage characteristics could affect choices about how data points are interpolated. Such information can sometimes be challenging to gather, depending on how and when the SCADA system was configured.

The aggregated maximum time gaps for tank level and pump runtime data streams are shown in Figures 2.3-1 and 2.3-2 for the data stream indices in Tables 2.3-1 and 2.3-2, respectively. As a whole, the data for tank level is good, with relatively small gap sizes. The significant period without data for the Bromley Tank corresponds to a period when it was out of service for painting. A detailed look at this data stream shows that all data points within the period of no data indicate a level of zero. Again, this is consistent with delta mode data storage, although it is unknown why zero valued points are stored at seemingly random times when the value remains zero. The pump runtime data are mostly absent, as shown in Figure 2.3-3, but this is not a cause for concern. Missing data indicate that the pump runtime has not changed in that interval, and thus the particular pump status is off. Significant time gaps during periods when the runtime is changing may indicate the pump to be on during that interval, or during a portion of that interval; the logic of converting these irregular runtime data into pump status information is discussed in Section 4.

Table 2.3-2. SCADA tank level tags and indices.

Level SCADA Tag	Description	Index
AQUA LI100	Aqua Tank Level	1
BARR LI100	Barrington Tank Level	2
HARR LI100	Bellevue Tank Level	3
BROM LI100	Bromley Tank Level	4
CLAR LI200	Campbell County Tank Level	5
DAYT LI100	Dayton Tank Level	6
DEV LI100	Devon Tank Level	7
DUD1 LI100	Dudley 1040 Tank Level	8
DUD2 LI100	Dudley 1080 Tank Level	9
IDA LI100	Ida Spence Tank Level	10
INDE LI100	Independence Tank Level	11
INDU LI100	Industrial Tank Level	12
JOHN LI100	Johns Hill Tank Level	13
KENT LI100	Kenton Lands Tank Level	14
LUML LI100	Lumley Tank Level	15
MAIN LI100	Main Street Tank Level	16
ROSS LI100	Rossford Tank Level	17
STAT LI100	South County Tank Level	18
NEW LI100	South Newport Tank Level	19
TMPIPE LI100	Taylor Mill Standpipe Tank Level	20

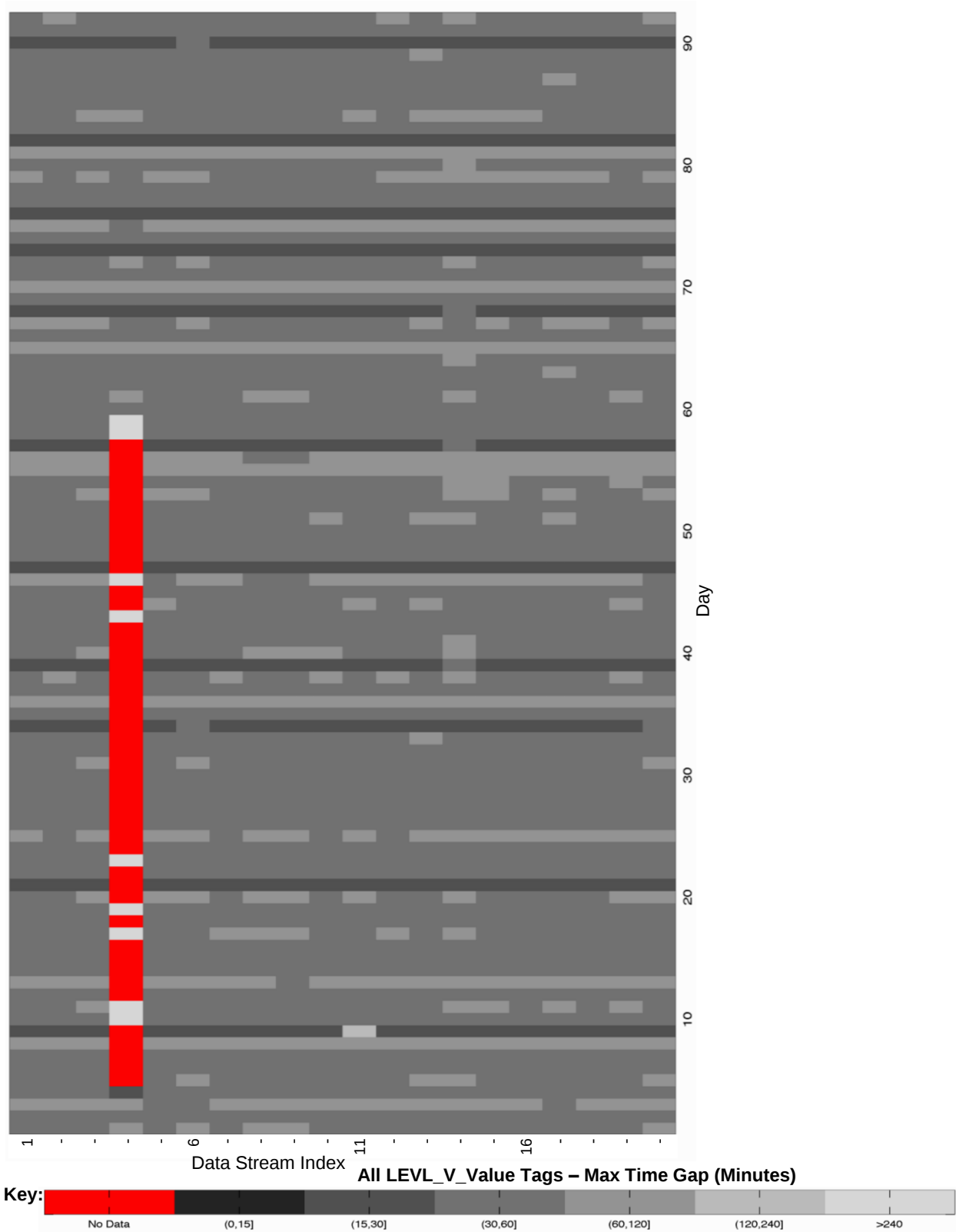


Figure 2.3-2. Maximum time gap (minutes) between valid measurements for all tank level measure data streams, for each day from October 1, 2012, through December 31, 2012. Indices refer to the SCADA tags in Table 2.3-2.

Table 2.3-3. SCADA pump runtime tags and indices.

Runtime SCADA Tag	Description	Index
US27 KQP534BNR	US 27 Pump 4 Status	1
US27 KQP531NR	US 27 Pump 1 Status	2
US27 KQP536NR	US 27 Pump 6 Status	3
US27 KQP532NR	US 27 Pump 2 Status	4
US27 KQP533NR	US 27 Pump 3 Status	5
US27 KQP535NR	US 27 Pump 5 Status	6
LATO KQP532NR	Latonia Pump 2 Status	7
DUD1 KQP534NR	Dudley 1040 Pump 4 Status	8
TMHS KQP536NR	Taylor Mill HS Pump 6 Status	9
DUD1 KQP531NR	Dudley 1040 Pump 1 Status	10
DUD2 KQP537NR	Dudley 1080 Pump 7 Status	11
COVI KQP532NR	W Covington Pump 2 Status	12
CARO KQP531NR	Carothers Rd. Pump 1 Status	13
WATER KQP531NR	Waterworks Pump 1 Status	14
WATER KQP532NR	Waterworks Pump 2 Status	15
BRS KQP2NR	Bristow Pump 2 Status	16
BRS KQP1NR	Bristow Pump 1 Status	17
DUD2 KQP538NR	Dudley 1080 Pump 8 Status	18
RICH KQP531NR	Richardson Rd. Pump 1 Status	19
COVI KQP531NR	W Covington Pump 1 Status	20
RICH KQP532NR	Richardson Rd. Pump 2 Status	21
BRS KQP3NR	Bristow Pump 3 Status	22
TMHS KQP533NR	Taylor Mill HS Pump 3 Status	23
DUD1 KQP532NR	Dudley 1040 Pump 2 Status	24
HAND KQP532NR	Hands Pike Pump 2 Status	25
TMHS KQP535NR	Taylor Mill HS Pump 5 Status	26
RICH KQP533NR	Richardson Rd. Pump 3 Status	27
RIPP KQP531NR	Ripple Creek Pump 1 Status	28
BROM KQP533NR	Bromley Pump 3 Status	29
LATO KQP531NR	Latonia Pump 1 Status	30
TMHS KQP534NR	Taylor Mill HS Pump 4 Status	31
BROM KQP531NR	Bromley Pump 1 Status	32
TMHS KQP532NR	Taylor Mill HS Pump 2 Status	33
WATER KQP533NR	Waterworks Pump 3 Status	34
CARO KQP532NR	Carothers Rd. Pump 2 Status	35
RIPP KQP532NR	Ripple Creek Pump 2 Status	36
DUD2 KQP535NR	Dudley 1080 Pump 5 Status	37
RIPP KQP533NR	Ripple Creek Pump 3 Status	38
HAND KQP531NR	Hands Pike Pump 1 Status	39
DUD2 KQP536NR	Dudley 1080 Pump 6 Status	40

Water Utility Case Study Using EPANET-RTX Libraries

Runtime SCADA Tag	Description	Index
DUD1 KQP533NR	Dudley 1040 Pump 3 Status	41
TMHS KQP531NR	Taylor Mill HS Pump 1 Status	42
BROM KQP532NR	Bromley Pump 2 Status	43

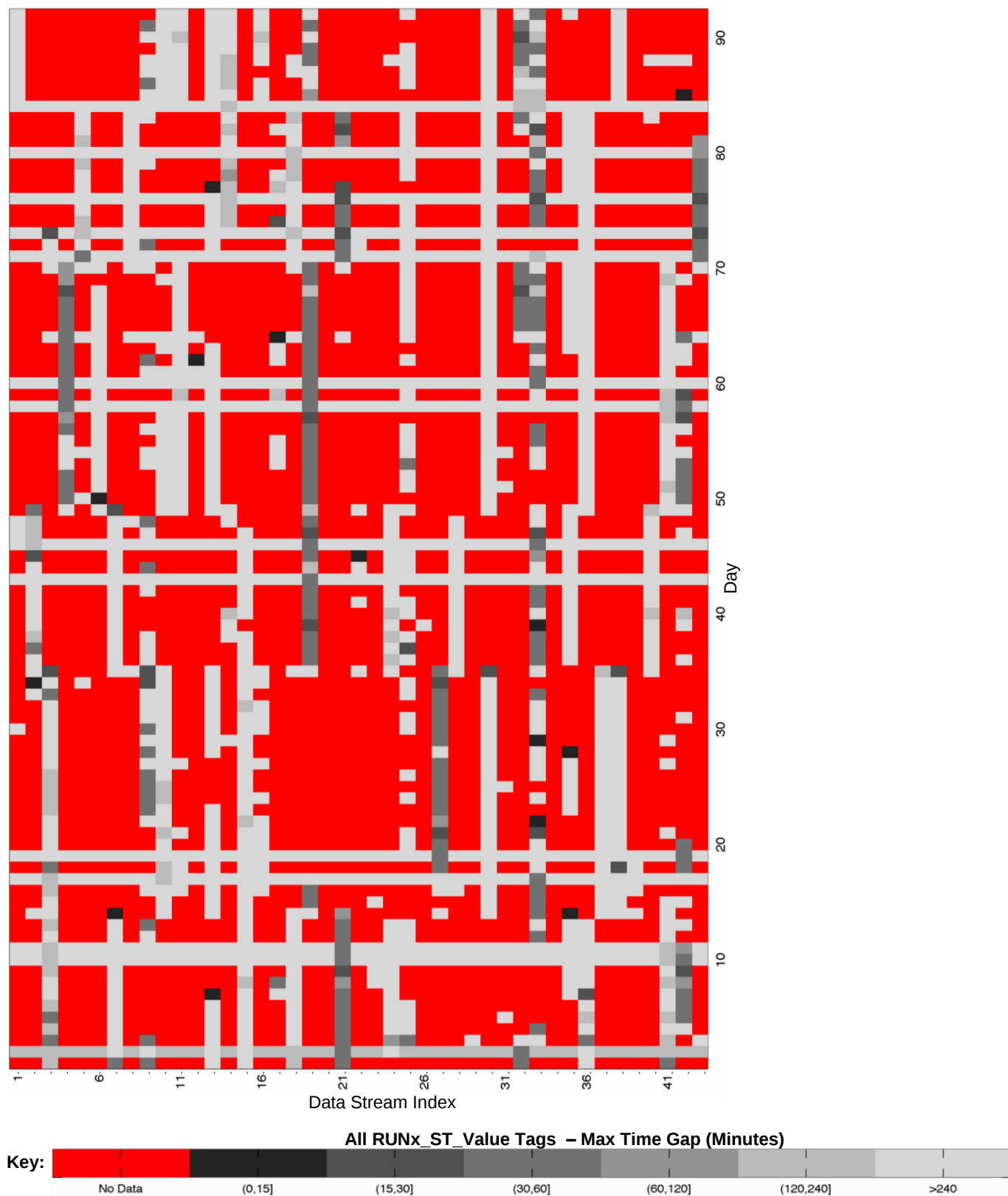


Figure 2.3-3. Maximum time gap (minutes) between valid measurements for all SCADA pump runtime measure data streams, for each day from October 1, 2012, through December 31, 2012. Indices refer to the SCADA tags in Table 2.3-3.

The maximum gap data for all three categories show strong relationships among the individual data streams. For example, in Figure 2.3-3 there are days when data are written for every pump runtime, presumably independent of pump status or runtime change, while in Figures 2.3-1 and 2.3-2 there exist days when the maximum gap is smaller or larger for most data streams. In short, the maximum gap size is not randomly distributed across the data streams, as might be expected, but rather is affected by an internal or external process. The source of these influences is unknown.

2.4 Tracer Field Study Description

An overview map of the portion of the NKWD service area that comprises the tracer study area is shown in Figure 2.4-1 along with monitoring locations. The area is divided into six regions labeled A to F, going from north to south, and 38 monitoring locations was distributed throughout the study area, but also concentrated in specific regions to gather information about spatial variation in tracer transport. The study included 46 monitors, but 8 malfunctioned without storing any data. The problems with the malfunctioning monitors generally included battery or conductivity sensor issues. While the number of monitors that malfunctioned was significant, we do not believe they compromised the goal to demonstrate that the EPANET-RTX technologies can be used to more efficiently calibrate a water quality model. Flow to the study area originates at the southern TP, flowing north and south through transmission lines within a single pressure zone, before descending through regulating valves to a lower zone in the north containing monitoring sites A and B. The northern TP was not operating its high service pumps during the test, so all of the flow to the monitors would be tagged by the brine pulses. This “un-split” mode is one of the normal operating modes for the utility, although the system is also operated in a split mode that requires the north TP to deliver water to regions A-D through its high service pumps.

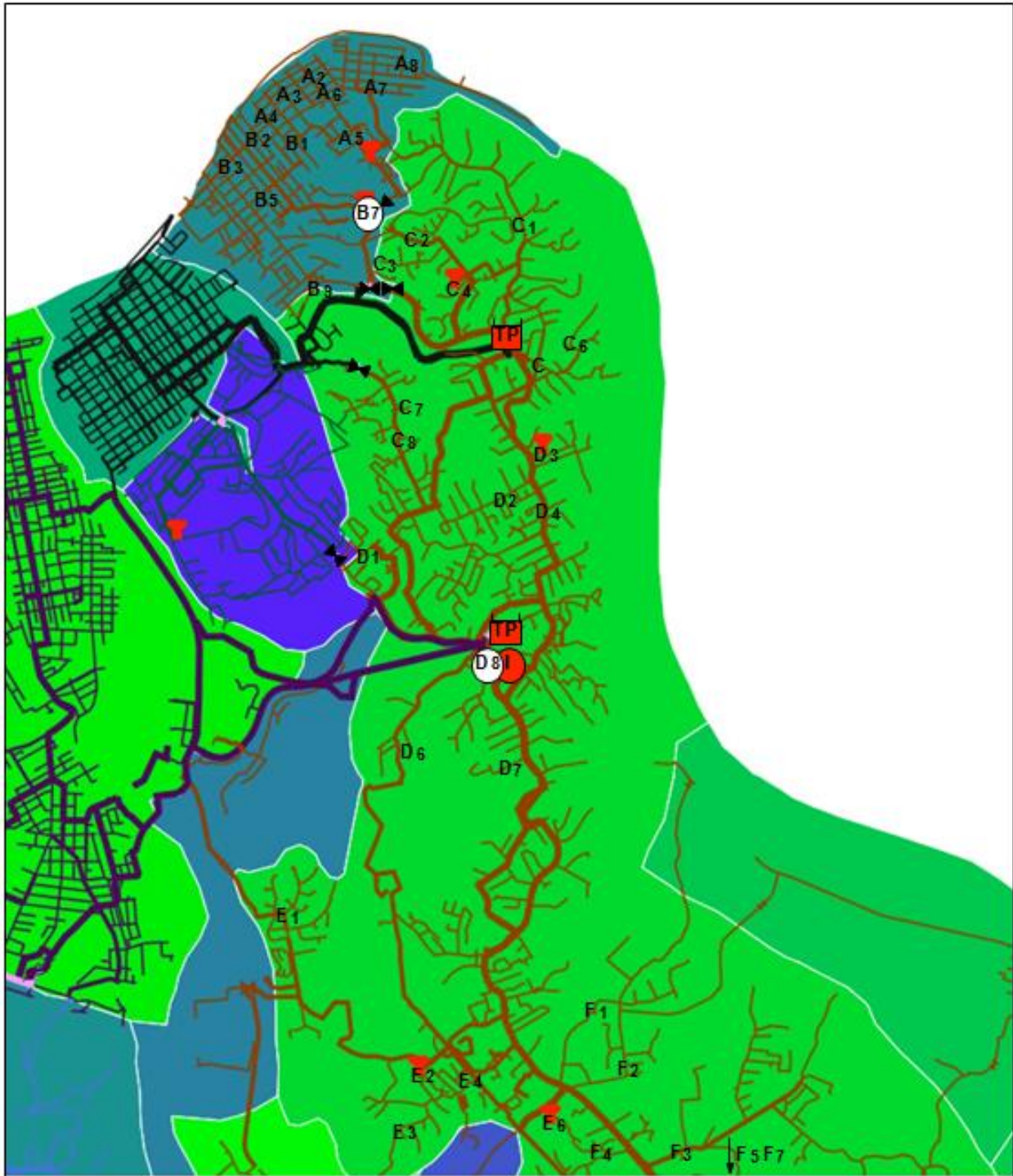


Figure 2.4-1. Tracer study area description illustrating Conductivity Monitoring Areas A, B, C, D, E, and F.

The tracer experiments included a calcium chloride tracer as a series of four pulses over a 12-hour period. The injection pulses were designed to produce a specific conductance of 1,000 $\mu\text{S}/\text{cm}$, more than twice the background level of approximately 350 $\mu\text{S}/\text{cm}$, though the peak conductivity achieved was somewhat less than the design value.

All monitoring locations were located at fire hydrants using standard hydrant adapters, and a continuous flow rate of approximately 1.0 gallons per minute (GPM) was maintained to reduce the residence time in the hydrant barrel to approximately 15 minutes. Each monitoring location included a continuous conductivity sensor housed in a secure container (Figure 2.4-2). Specific conductance data were downloaded from the data loggers periodically. The discharge line from each hydrant was positioned to ensure that it drained to a sewer (if present) or to an area that allowed infiltration.

One objective of the monitoring location selection process was to identify locations that represented the range of hydraulic residence times while being spatially diverse. A secondary objective was to concentrate monitors in one or more densely populated regions, so that variability over small areas could be assessed. Using the distribution system network model provided by the utility, water age and tracer simulations were performed using EPANET (Rossman, 2000) to provide input for the selection of monitor locations.

The selection of locations for monitor placement was performed manually using EPANET. In all, 46 conductivity sensors were put in place. While the overall intent was to place sensors to provide representative monitoring locations with respect to spatial distribution and water age, there were additional key locations that were identified as important regardless of the underlying hydraulic characteristics. Specifically, one monitor was placed downstream of the injection location, and six monitors were placed on the influent/effluent lines of the storage tanks within the study region. For the remaining 39 locations, the intent was to manually identify monitor locations to represent the distribution of water age as well as intensely monitor a more populated “grid” within the network. Thirty of the remaining 39 monitors were placed to have six monitors in each of the water age quintile ranges using visual inspection to spatially distribute the monitors. The remaining nine monitors were placed in the denser, gridded north region of the system (Regions A and B); three locations were intended to capture the influent water quality into this region and the other six locations were selected to capture the potential variability in hydraulic or transport characteristics.

In achieving these design objectives, the monitoring locations represent a strong test of real-time water quality modeling accuracy. Difficult locations off of transmission mains in regions with small or localized demands were not discouraged. Table 2.4-1² shows the monitoring locations and pipe diameters for each of the 38 monitors that produced data for the analysis. In addition, the table identifies whether the monitor was located on a dead-end main or a storage tank. Of the 38 monitors, 24 were located on pipes of 8-in. diameter or less, and 7 were located on dead-end mains.

The electrical conductivity (EC) signals were measured in 1-minute intervals and logged continuously during the study period at each location. Each monitor consists of a conductivity sensor, display unit, data logger, battery, and flow-through piping, as shown in Figure 2.4-3. (A 9v lithium battery and data logger were within the monitor enclosure in these units; other setups used the same conductivity sensor and display but with a different physical configuration.) The conductivity sensor is a 4-electrode conductivity sensor.

² Monitor manufactured by Analytical Technology, Inc., Collegeville, PA (ATI model Q45C4), and can be used to measure specific conductance in the range of 0 to 2,000 $\mu\text{S}/\text{cm}$ with an output voltage ranging from 0 to 2.5 V. All monitors were calibrated against 1,000 $\mu\text{S}/\text{cm}$ standards and tested for variability between the devices by measuring three different lab samples.

Table 2.4-1. Characteristics of selected monitoring locations³.

Location	Pipe Diameter (in.)	Note
A2	6	
A3	6	
A4	8	
A5	6	
A6	4	
A7	8	
A8	8	
B1	6	
B2	8	
B3	10	
B5	4	
B7	12	Tank
B9	8	
C1	8	
C2	12	
C3	8	Dead-end
C4	12	Tank
C6	6	Dead-end
C7	6	Dead-end
C8	12	
D1	12	
D2	8	
D3	8	Tank
D4	16	
D6	6	Dead-end
D7	6	Dead-end
D8	16	
E1	12	
E2	12	Tank
E3	8	Dead-end
E4	8	Dead-end
E6	20	Tank
F1	8	
F2	6	
F3	12	

³ If blank, the location is a connected (not dead-end) pipe.

Water Utility Case Study Using EPANET-RTX Libraries

Location	Pipe Diameter (in.)	Note
F4	8	
F5	12	
F7	8	



Figure 2.4-2. Typical monitor setup.



Figure 2.4-3. Typical components of the monitor (e.g., conductivity sensor, display, logger, piping and tubing).

2.5 Calcium Chloride Injection Protocol

The field activities included injecting a CaCl_2 solution as a series of pulses at one location over a 14-hour period. Multiple pulses were used to provide more information about network flow dynamics, compared to using a single pulse. The high service pump station at the southern TP was selected in consultation with utility staff, based on safety, security, space, and flow control. A series of three valve changes were made prior to the test in order to “un-split” the main pressure zone served by the southern TP. Un-splitting the main pressure zone expanded the service area of this plant and thus the region affected by the brine pulses. A National Sanitation Foundation (NSF) food-grade CaCl_2 solution was added to TP finished water, producing a series of brine pulses of between 1 and 2 hours’ duration. The pulse injection rate was selected to produce a detectable increase in the specific conductance above the background (approximately $350 \mu\text{S}/\text{cm}$), and yet maintain a significant safety factor when compared to the maximum allowable CaCl_2 increase based on applicable Federal and state standards on chloride. In between pulses the CaCl_2 feed was discontinued. For 2 days prior to the start of brine addition, and for a week afterward, the specific conductance was recorded at the monitoring locations.

The USEPA secondary standard on chloride is $250 \text{ mg}/\text{L}$. Based on historical data collected from NKWD over several years, the range in chloride concentrations of finished water was between 16 and $66 \text{ mg}/\text{L}$ (given an analysis of finished water data collected between January 2010 and August 2011). More recent data from the past year for the distribution system and TPs indicated a similar range in chloride concentrations. Assuming a conservative background chloride concentration of $41 \text{ mg}/\text{L}$ during the time period of the test, the applicable standards limit a chloride concentration increase to $(250 - 41) = 209 \text{ mg}/\text{L}$. There were no applicable Federal or primacy agency standards for calcium, and thus it was regulated based on CaCO_3 solubility.

Food grade CaCl_2 was obtained in totes as a pre-mixed 33% (by weight) solution⁴. Assuming a specific gravity of 1.322 @ 60° F , the 33% solution equates to $436.26 \times 10^3 \text{ mg } \text{CaCl}_2/\text{L}$, or $278.71 \times 10^3 \text{ mg Cl}^-/\text{L}$. Reliable controls were placed on the volumetric flow rate of the CaCl_2 solution injection pump, such that the chloride concentration was within the regulatory limits. The maximum CaCl_2 injection flow rate of the food grade stock solution can be calculated from a mass balance at the injection site,

$$Q_{\text{CaCl}_2}^{\text{max}} = \frac{209 \text{ mg Cl}^-/\text{L}}{278.71 \times 10^3 \text{ mg Cl}^-/\text{L}} \times Q_{\text{Prod}} = (0.75 \times 10^{-3}) \times Q_{\text{Prod}}, \quad (1)$$

where $Q_{\text{CaCl}_2}^{\text{max}}$ is the maximum allowable flow rate of the NSF food grade CaCl_2 solution, and Q_{Prod} is the production flow rate (in the force main receiving the injection), with both flow rates expressed in the same units. Knowing the production flow rate Q_{Prod} , obtained from the SCADA system at the time of the injection, Equation 1 was used to calculate the maximum injection flow rate for regulatory purposes. The adopted test protocol limited the maximum addition to 80% of this value.

For an effective tracer test, the injection of brine must create a measurable increase in specific conductance above background. The utility reported that the specific conductance in the distribution system varied between 248 and $637 \mu\text{S}/\text{cm}$. The impact on the specific conductance can be estimated from the relationship between total dissolved solids (TDS, mg/L) and specific conductance (EC, $\mu\text{S}/\text{cm}$),

⁴ Tetra Chemicals™ (West Memphis, AR) NFS grade calcium chloride

$$TDS = k_e EC, \quad (2)$$

or,

$$\Delta EC = \frac{\Delta TDS}{k_e}, \quad (3)$$

where the correlation factor $0.5 < k_e < 0.8$. Assuming the maximum CaCl_2 injection flow rate from Equation 1, the resulting increase in total dissolved solids, ΔTDS , can be calculated,

$$\Delta TDS = 0.8 \times (0.75 \times 10^{-3}) \times (436.26 \times 10^3 \text{ mg/L}) = 261.76 \text{ mg/L}. \quad (4)$$

Assuming a “worst case” $k_e = 0.8$, a corresponding increase in specific conductance was estimated,

$$\Delta EC = \frac{261.76}{0.8} = 327.2 \text{ } \mu\text{S/cm}, \quad (5)$$

which is greater than 80% increase over background; this increase is significant given the accuracy of the specific conductance sensors used.

To determine the appropriate calcium chloride dose on the day of the test, the background concentration of chloride was required on the day of the tracer test. To do so, a relationship between historic chloride and specific conductance was developed for water samples collected in finished water over the previous year. A summary of the analysis performed is attached as Appendix A. The analysis shows that based on a measured specific conductance, it is possible to estimate a range of chloride concentration. A very conservative approach was to assume the high end of this range as the current chloride concentration.

3.0 Real-Time Modeling Using EPANET-RTX

EPANET-RTX is a set of object libraries used for building real-time hydraulic modeling environments. It is a set of building blocks (classes and wrappers), which can be used and extended to create real-time data fusion applications, which could include data acquisition and predictive forecasting. EPANET-RTX provides interoperable access to several different technologies that are foundational to real-time modeling. These technologies involve accessing a SCADA historian database, using filtering, smoothing, and other data transformation methods, and running hydraulic and water quality simulations. EPANET-RTX forms a software scaffolding that interfaces with these technologies to enable the smooth migration of data from the measurement domain into the modeling domain. The typical use of EPANET-RTX libraries could comprise building an application that would connect a water utility's network model and run an extended-period simulation driven by sensor measurements that have been or are being recorded in a SCADA historian. The goal of the EPANET-RTX library is to make the complex task of network model and SCADA data fusion easier for programmers and engineers to use. The user of the EPANET-RTX libraries may choose to incorporate as much of the functionality as desired. For instance, EPANET-RTX could be used only to connect to a SCADA system, clean certain data streams, and provide a predictive forecast of sensor data. Or EPANET-RTX could be used for further EPANET development (e.g., for graphical user interface [GUI] development or other purposes). Additionally, many

processes that would typically be considered part of network model calibration are implemented automatically by an EPANET-RTX-based real-time model.

Here we describe the process for building a real-time model⁵ using EPANET-RTX libraries. Section 3.1 describes the real-time simulation process, after building and using an EPANET-RTX-based application. Section 3.2 describes how an EPANET-RTX-based model is configured using SCADA real-time data streams. Section 3.3 describes how DMAs are defined and how demands are disaggregated to DMA network model junctions.

3.1 Real-Time Simulation Process

The real-time modeling results presented here were obtained using a simple EPANET-RTX application, similar to that shown in Figure 3.1-1. The application uses built-in RTX objects that read the real-time model specification using the libconfig configuration file library (Lindner, 2012). This one configuration file specifies the SCADA databases and how to access their data records; the time series to query in the databases (i.e., the SCADA “tags”) and their properties (e.g., units); the transformations to be applied to each time series; and the connections between the transformed time series and the network model elements.

```
void runSimulationUsingConfig(const string& filePath, time_t start, long dur) {
    // RTX configFactory object
    ConfigFactory config;
    // Pointer to RTX model object Model::sharedPointer model;
    // Process the configuration file and get a pointer to the model
    config.loadConfigFile(filePath); model = config.model();
    // RTX::model knows how to run an EPS with SCADA connectivity model->runExtendedPeriod(start,
    start + dur);
}
```

Figure 3.1-1. Prototype EPANET-RTX application code (C++) for executing real-time simulation using an EPANET-RTX configuration file.

The prototype application runs a single extended period simulation in the following way; all these steps are initiated within the runExtendedPeriod() method of the RTX model class (USEPA 2013):

0. Ignore network model control rules and time patterns. Control rules and time patterns are discarded because they represent static knowledge or assumptions, about particular extreme or average conditions, that are used for planning purposes. Real-time modeling replaces these assumptions with actual knowledge about the system operations for the time period being represented.
1. Access new data from the SCADA database. Queries are constructed to obtain the last known good value for all SCADA time series specified in the RTX configuration file.
2. Transform the measurements, and interpret the statuses and settings for all boundary data streams. Raw SCADA data are transformed according to the time series data pipeline⁶ transformations specified in the EPANET-RTX configuration file (these are described in

⁵ We use the term “model” here to be inclusive of both the hydraulic and water quality aspects of the model. Later we are specific about whether we are referring to just the “hydraulic” or the “water quality” aspects of the model.

⁶ Pipeline in this context should not be confused with a network model pipe which connects nodes.

Section 4). The design of EPANET-RTX elegantly handles the execution of such data transformation pipelines, as each EPANET-RTX data transformation object is responsible for communicating with its upstream data source.

3. Calculate and allocate demand within demand metered areas. DMA demands are calculated by aggregating boundary flows along with flows into storage tanks. These DMA demands are disaggregated according to the modeled base demand at each node.
4. Advance the simulation and store results; go to Step 1.

For this case study, the above steps were executed for a particular historical time frame (i.e., corresponding to “start” and “dur” in Figure 3.1-1). In a true real-time simulation, the EPANET-RTX application software would periodically wake up from an idle state, perform the above Steps 1 through 4, and then go to sleep for a specified interval. Such a persistent real-time simulation would provide a constantly updated view of system status and model performance.

Also, as a practical matter, the results presented in the following sections were not obtained through a live connection to the SCADA historian database (Wonderware® [Invensys software, Lake Forest, CA] SCADA historian based on Microsoft SQL [structured query language] Server). To avoid the need to be on-site (the NKWD SCADA historian database was not connected to the internet), a copy of the SCADA historian server was created, so that a virtual SCADA historian could be run off-site. The EPANET-RTX software libraries and application program used were no different from that which would connect to the live SCADA historian, and in fact the only difference with a live connection was that queries were limited to data that existed when the copy was created. Hence, the real-time process described here was not a “batch process”.

3.2 EPANET-RTX Real-Time Model Configuration and Data Transformations

Here we describe the process for configuring a real-time, EPANET-RTX-based network model, a process involving a series of data transformations.

The time series data transformation pipelines represented in the following sections form the foundation of an accurate real-time hydraulic model. They are templates that can be applied to different data streams within the same category, and were devised through experimentation. The EPANET-RTX libraries were designed with such experimentation in mind, acknowledging the important role that connecting real data streams to the network model have in determining model accuracy. Rather than develop a separate program to query the database, implement a set of serial transformations, and store the results in some fashion, it is simpler and more reliable to configure a time series data transformation pipeline composed of EPANET-RTX objects, and simply request the data points. Such requests are propagated backward through the time series pipeline (as needed — some points may be already available from prior requests), and results may be automatically persisted in a database.

In the following sections, we describe data transformation time series pipelines for the following data stream categories: tank level, pressure, pump status, flow, and altitude valve status. We also describe a time series data transformation pipeline used to construct key missing flow data from one of the TPs, without which real-time demand for DMA 2 could not be estimated.

3.2.1 Tank Level Data Streams

Tank level data are used for two purposes: DMA demand estimation (see Section 4.2) and comparing with real-time model predictions. Figure 3.2-1 shows representative raw tank level data for the NKWD system. There is obvious noise present in the level data, including sudden spikes of several feet — associated with changes in pump status or demand, and consequent switching from a fill to drain cycle, or

vice versa. These large spikes, as well as some low level noise, are consistent with measuring tank level using pressure transducers on the inlet/outlet line — rather than on a static pressure line, or within the tank itself. In this case the pressure reading and the tank level indicator are affected by minor losses associated with tank piping and valving. When the tank is filling, the hydraulic grade overestimates the true level, and when the tank is draining, it underestimates the level, due to head loss between the transducer and the point of discharge within the tank.

Both low level noise and sudden spikes must be adequately filtered before tank level data can be used for DMA demand estimation. The time series data transformation requires converting tank level data into net tank inflow — a process that requires differentiating the tank level signal. The interaction between data smoothing, or filtering, and differentiation has been studied for some time because of its practical importance in a wide variety of applications (see Wood, 1982, or a practical online introduction by O’Haver, 2013). If smoothing is not performed on a signal prior to differentiation, the signal-to-noise ratio is reduced. A practical rule-of-thumb for smoothing prior to differentiation is to use $n+1$ applications of a simple rectangular weighted moving average filter when computing the n^{th} derivative; thus for a first derivative it is often sufficient to use two applications of a moving average (equivalent to a single pass of a triangular weighted moving average — see O’Haver, 2013).

Figure 3.2-2 represents the EPANET-RTX time series pipeline implemented for smoothing all tank level data. The data transformation pipeline begins with a Time Series object, named “Tank Level” in this generic representation, but assigned the SCADA database identifier in a particular instance. This object knows what database holds the associated data stream, and how to connect to it. Asking this object for data points within a time range will retrieve raw SCADA values. Points from the Time Series object are input to the Resampler object (more accurately, the Resampler fetches its points from the Time Series object). Resampling produces regularly spaced points (in time) by interpolating at intervals specified by its clock. Interpolation could be done in a number of ways, but simple linear interpolation is used here.

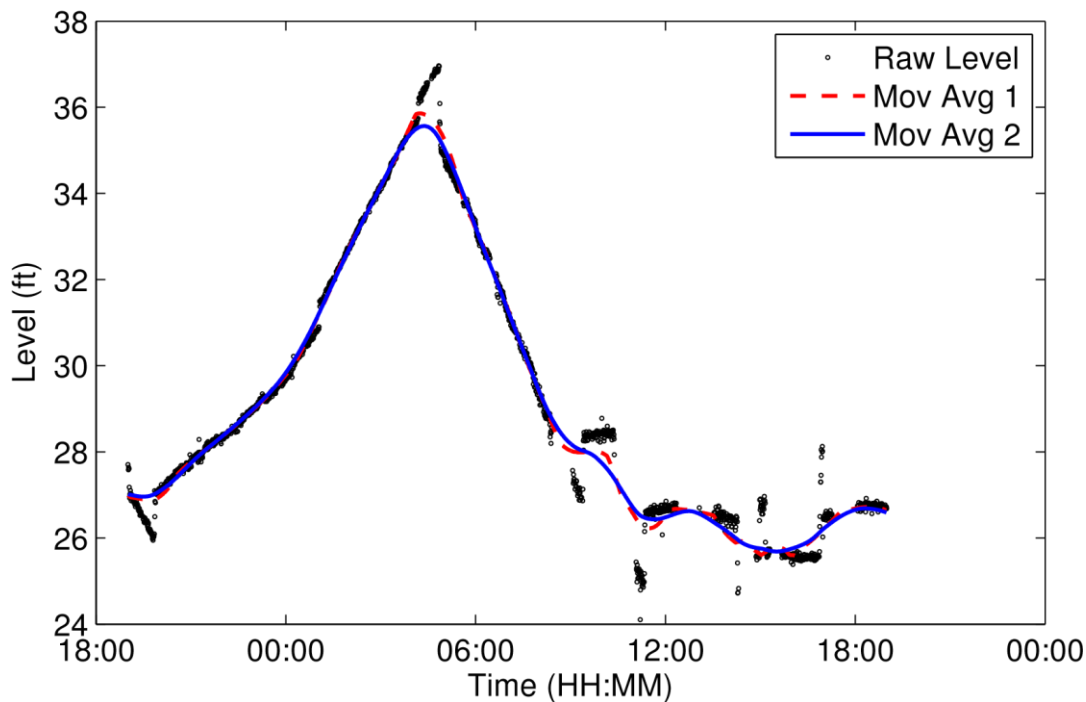


Figure 3.2-1. Representative raw storage tank water level data with 1st and 2nd moving average filters. Note signal noise and spikes separating fill and drain cycles.⁷

The tank level Resampler uses a 1-minute clock, so interpolated points are produced at 1-minute intervals. These regularly spaced points are input to the MovingAverage object, which implements a uniform (rectangular) weighted moving average. The moving average requires a window width, specified as a number of points. Here the window width is 91 points, so the filtered point at time t averages its source values in the 90-minute time interval $[t-45, t+45]$ (recall the Resampler clock is 1 minute). A subsequent identical MovingAverage object performs the identical function as the first, which, as mentioned, is equivalent to a single pass of a 90-minute window triangular weighted moving average filter. The last object in the time series pipeline represents the association with a model element — in this case, to the element associated with the SCADA “Tank Level.”

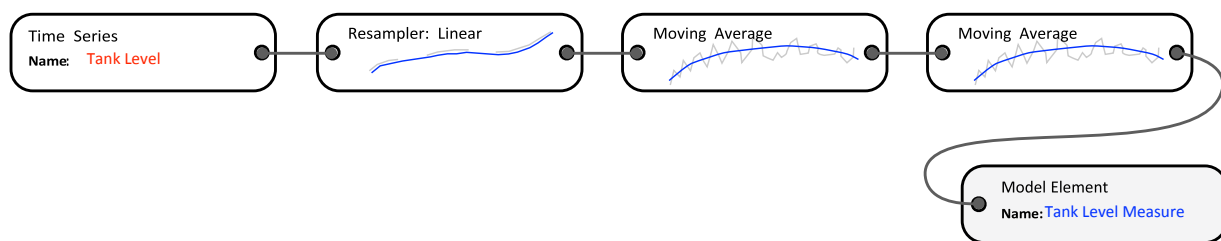


Figure 3.2-2. Time series data transformation pipeline for resampling and smoothing raw SCADA tank level data. The Resampler uses a 1-minute clock with linear interpolation. Sequential moving average filters provide the smoothing and each use a 91-data point, or 90-minute, window.

Figure 3.2-1 shows representative results from the tank level data transformation pipeline, including data points produced by both moving averages. As with any filtering process, there will be a loss of signal along with the decrease in noise. This particular smoothing process is not claimed to be optimal for purposes of real-time modeling, and the width of the smoothing interval (90 minutes) may be adjusted or subject to further scrutiny by studying its influence on simulation results. This data transformation scheme has, however, yielded good results for the NKWD case study.

3.2.2 Pressure Data Streams

Figure 3.2-3 shows typical pressure measurement data — in this case the discharge pressure at a pump station. The signal is noisy, as expected for data generated directly by an inline pressure transducer. The significant jumps in the signal correlate with hydraulic events occurring in the distribution system, in particular with changes in pump status. The data transformation seeks to reduce low level noise without eliminating signals that have operational causes. These raw pressure data also show uneven polling, or artifacts of other downstream data management processes, as the time gaps between successive points range from seconds to tens of minutes (and can be several hours). This behavior is observed across all the analog data streams, and as observed in Figures 2.3-1 and 2.3-2, there are unexplained relationships between the maximum daily time gaps across different data streams.

⁷ HH:MM represents hours and minutes.

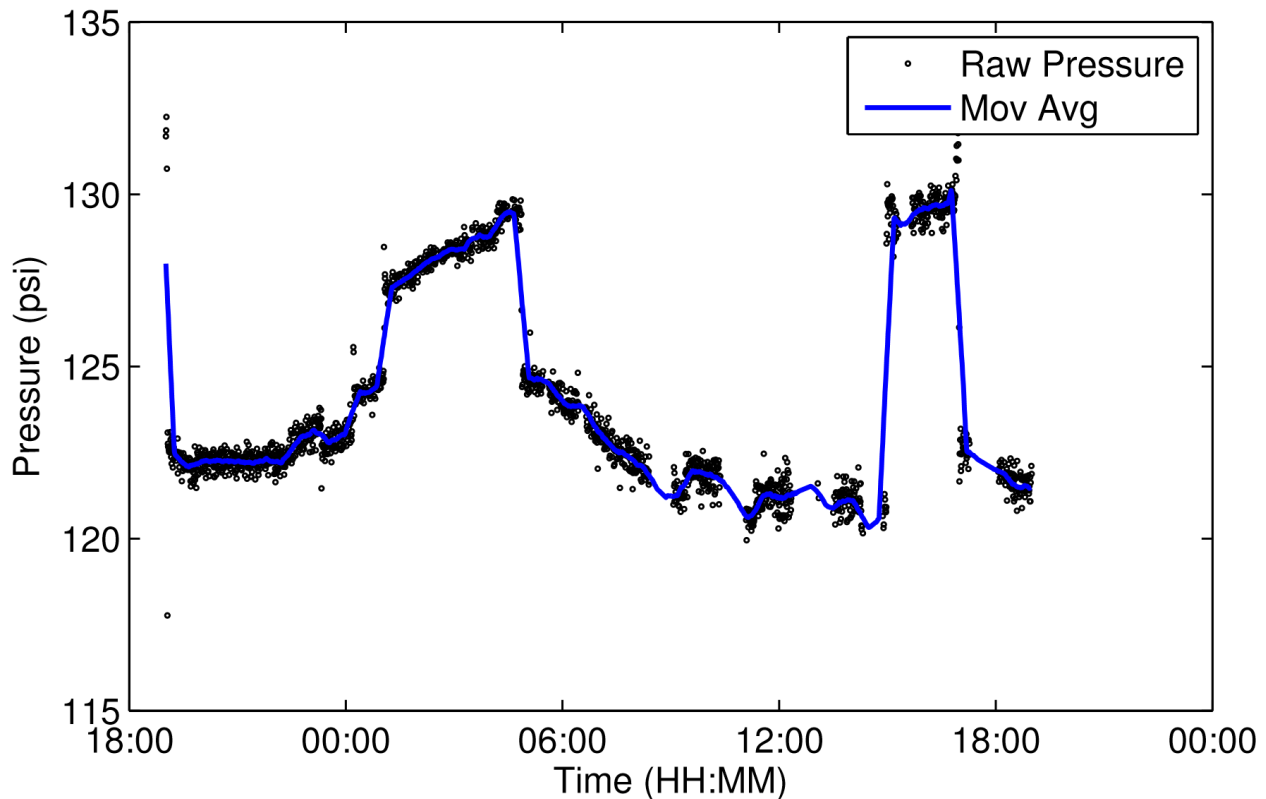


Figure 3.2-3. Representative raw pressure data with moving average filter. Note the presence of cycles of intense data polling activity with interspersed data gaps. This behavior is observed in many raw SCADA time series.

The time series data transformation pipeline used for all pressure data is shown in Figure 3.2-4. This pipeline represents perhaps the simplest possible set of transformation steps, consisting of resampling with linear interpolation, and a single pass of a rectangular moving average filter. The Resampler clock is again 1 minute, and the moving average window is 25 data points, or 24 minutes, wide. Results from applying this time series pipeline to the representative pressure data are shown in Figure 3.2-3.

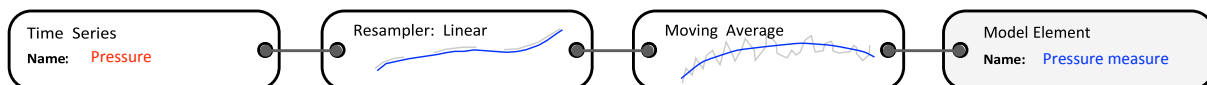


Figure 3.2-4. Time series data transformation pipeline for resampling and smoothing raw SCADA pressure data. The Resampler uses 1-minute clock with linear interpolation. Sequential moving average filter uses a 25-data point, or 24-minute, window.

3.2.3 Pump Status Data Streams

High service and booster pump operation is recorded in SCADA using non-reset runtime meters. These are digital data — the reading from a clock, in hours, equal to the cumulative time that the pump has been in a running state⁸. These data streams are processed in real time to produce the binary pump status data streams that will represent pump operation in the real-time model. A data transformation approach using EPANET-RTX objects is represented by the time series data transformation pipeline in

⁸ There remain details that are unknown — whether the runtime meters reference the time a discharge valve was opened, or when the pump motor starts and stops, or a relevant switch.

Figure 3.2-5. The runtime data stream is first resampled using a 1-minute clock and assigned as the input to a First Derivative object, which differentiates its input data stream. Since runtime has time units, the derivative data stream is dimensionless. If there were no errors in time stamp or value, and no significant data gaps, the derivative value would equal the fraction of time the pump was on in any sampling interval; its value would lie in the interval $[0,1]$ – 0 if off, 1 if on, and fractional on the boundaries of a pump cycle. The derivative data stream may then be assigned as the input to a Threshold object, which compares its input value at time t , $x(t)$, to a threshold value, $\bar{x}$, and assigns a value of 0 if $x < \bar{x}$, and 1 otherwise.

Representative results using this derivative pump status time series data transformation pipeline are shown in Figure 3.2-6. The left figure shows 4 weeks of cumulative pump runtime for a single pump. Several different data streams are processed. The SCADA data is the true solution, obtained by accumulating pump runtime directly from the SCADA record. The “Deriv w/Resamp” data are obtained by implementing the time series data transformation pipeline in Figure 3.2-5, and then accumulating pump runtime from this new status data stream. (Given the time scale in the left figure, these data appear to lie on top of the true solution.) The solution labeled “Deriv” is obtained using the time series pipeline in Figure 3.2-5 skipping the resampling step. Reflecting on the transformation process, there is no logical requirement for resampling; indeed, resampling would seem to only add uncertainty and potential errors, depending on the size of the resampling clock. Yet the data in Figure 3.2-6 show large errors in cumulative runtime when differentiating the raw data. The source of these errors turns out to be seemingly random errors in the data point timestamps. Polling of runtime data (and other data streams as well) produces a time spacing on the order of 10 seconds. While the digital runtime clock values are accurate enough, the timestamps may be off by several seconds, leading to significant errors in the derivative values, and frequent false pump starts. Resampling is a useful remedy, simply because the timestamp error magnitude is relatively small compared to the resampling interval (clock).

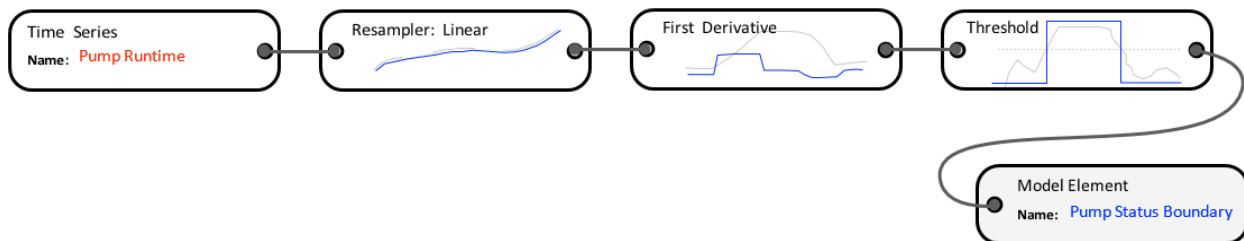


Figure 3.2-5. Time series data transformation pipeline using derivative and threshold to derive binary pump status from raw SCADA pump runtime data.

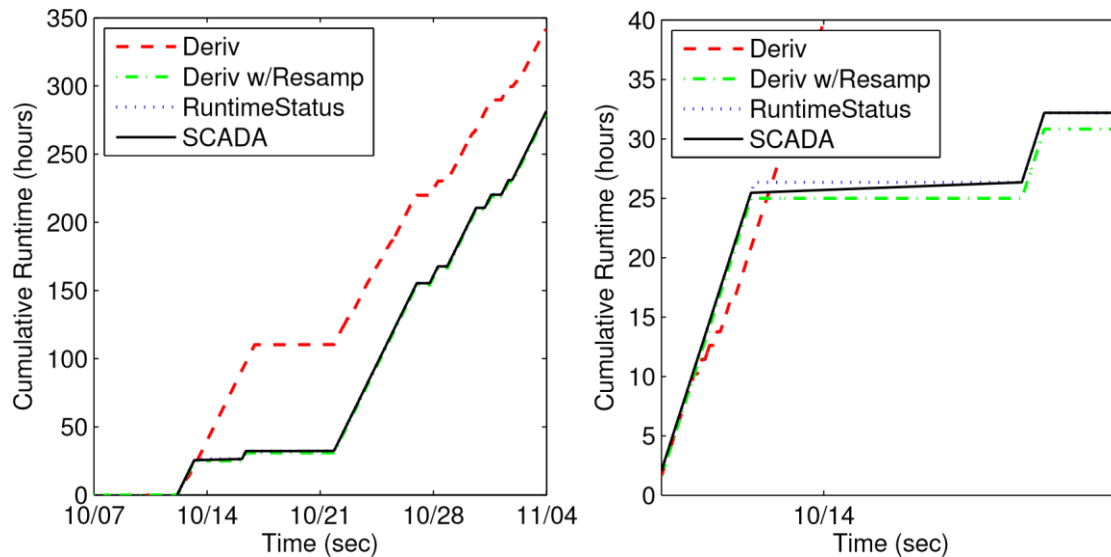


Figure 3.2-6. Representative cumulative pump runtime calculated from raw SCADA pump runtime data, as well as three pump status data streams derived from runtime data. Right figure shows detail around two separate pump status changes, illustrating pump status errors introduced when differentiating runtime data, which are resolved by the EPANET-RTX RuntimeStatus class.

Errors in the cumulative runtime can still occur when differentiating runtime to produce the status data stream due to large time gaps between points. The right plot in Figure 3.2-6 provides some detail over several days surrounding two pump cycles. While the SCADA data *should* yield a cumulative runtime with a slope of either zero or one⁹, the results show a time span exceeding 2 days between a pump off and on status, where the slope is distinctly greater than 0 (but less than 1). Within this time frame the pump has run for about 1 hour, but the derivative pump status with resampling does not register that runtime; lowering the slope threshold for turning on the pump will not help, for as soon as that threshold is reached, the pump would be turned on for the entire 2-day time gap — a much greater error than the 1 hour of lost runtime. Errors originating in large time gaps are common for this SCADA system so they should be expected in each runtime data stream.

The perception, at least, is that significant errors in pump status could lead to significant errors in real-time model results. Moreover, it is disappointing to process inherently high quality SCADA values — the digital runtime clocks — and derive pump status data streams that do not preserve actual pump runtime. This motivated the development of a specialized EPANET-RTX class named RuntimeStatus that processes runtime clock data and accurately detects the status changes; the new time series data transformation pipeline for pump status, which was used for all the real-time modeling results, is shown in Figure 3.2-7. It is expected that this EPANET-RTX class will have wide applicability and use.

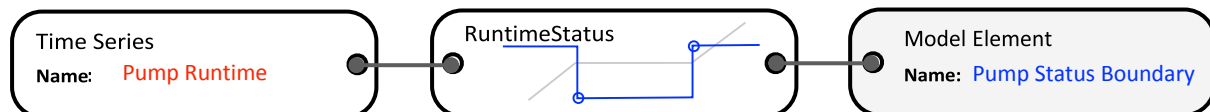


Figure 3.2-7. Time series data transformation pipeline using special-purpose RuntimeStatus class to derive binary pump status directly from raw SCADA pump runtime data.

⁹ Or very close to 1. Errors in the timestamp mean that the slope will not equal exactly 1, but these errors are not cumulative, so that as time progresses during a pump on cycle, the slope approaches unity.

In summary, a `RuntimeStatus` object processes raw runtime data, in order to identify the time when a pump status changes from on to off, or off to on. It does this accurately because it is looking explicitly for those status changes in the time record, as opposed to a general EPANET-RTX derivative object that is limited by its local perspective. The data in Figure 3.2-6 is a case in point — a threshold object will repeatedly leave the pump in an off state because its derivative source is too low, even if a simple difference of successive runtime values proves the pump was on for about an hour during the time gap. The `RuntimeStatus` object is able to see this runtime difference because it is looking for it, and ensure that the pump is run for a time that obeys the SCADA record.¹⁰ This behavior is illustrated in Figure 3.2-6, which shows that the `RuntimeStatus` preserves the cumulative SCADA runtime by delaying the pump off status change. Alternatively, the algorithm could advance the beginning of the next pump on status change, but it is impossible to know where to assign the needed runtime, within the data gap. Nevertheless, at least the total runtime is preserved for each of the 43 high service and booster pumps.

3.2.4 Pipe Flow Data Streams

In general, flow measurements that could participate in DMA demand calculations were transformed like the tank level data: resampled on a 1-minute clock using linear interpolation, and used as input for two sequential `MovingAverage` objects, each with a 91-data point averaging window. The two moving averages used for tank level data were driven by the need to differentiate those data streams when they enter the DMA demand aggregation. There is no proven need for consistency in the treatment of all data streams that participate in DMA demands, and perhaps the justification for such consistency is mostly aesthetic, at this point. Still, it seems undesirable to aggregate data streams that have been filtered to very different degrees, and so in anticipation of that aggregation through the DMA demands, the flows are filtered the same as the tank levels. The effect otherwise would be to add, say, pump station flows with sharp boundaries at the pump status changes, to tank flows where the changes from filling to draining have been more heavily filtered. Also, since only linear filters are used, their use does not affect the mean values.

For most pump stations, additional processing of the data streams was performed prior to moving average smoothing to force the flow to zero when all pumps were off¹¹. The motivation for this additional processing was the presence of significant and regular time gaps between points in the flow data streams. Figure 3.2-8 shows illustrative raw pump station flow data (SCADA) along with the *station* status (`Status`) — equal to 1 if at least one station pump is on, and zero if all pumps are off. Large data gaps appear regularly in this flow record — indeed no data are present when station pumps are off — but gaps are present to some degree in all flow data streams. These gaps create significant errors in the processed flow measure (and in any related DMA demand calculations) if performed by the typical resampler and moving average time series data transformation pipeline (`Smoothed`); the flow measure when the pumps are off is significantly greater than zero. It is not possible to remove these errors through a different raw data resampling and interpolation method, because of the sparsity of the data points.

Data transformation strategies were developed for “trimming” pump station flows so that the data gaps would be managed effectively in real time. While such problems could be dealt with manually in a fairly simple manner, in real-time the data processing must be automatic and robust. The core idea is to

¹⁰ The `RuntimeStatus` class is able to handle the “normal” non-reset runtime, as well as runtime clocks that reset periodically at a certain time or when a threshold is reached. One NKWD runtime data stream is reset, while the others are non-reset. The one reset runtime seems likely due to a SCADA programming error or omission.

¹¹ A zero flow assumption is valid when the flow sensor does not measure station bypass flow — true for all by two NKWD pump stations.

generate a pump station status data stream and use that to insert zero-valued points into the data stream, when they logically should be present. The data transformation pipeline that accomplishes this is shown in Figure 3.2-9. While this pipeline appears significantly more complex than those examined previously, each data transformation component is represented by an existing EPANET-RTX object, which does all the data processing. The upper portion of the time series data transformation pipeline constructs the station status data stream, by using an Aggregator object to sum the individual pump statuses, and then thresholding that at zero — so that if at least one pump is on, the result will be 1, and if all pumps are off, the result will be zero. This data stream is then multiplied with the resampled and linearly interpolated raw pump station flow, producing a data stream that equals zero whenever all pumps are off, and equals the resampled flow measure when at least one pump is on. This latter data stream is then filtered, producing the trimmed and smoothed flow measure. Figure 3.2-8 shows the Trimmed/Smoothed flow measure, which better represents the true station flow.

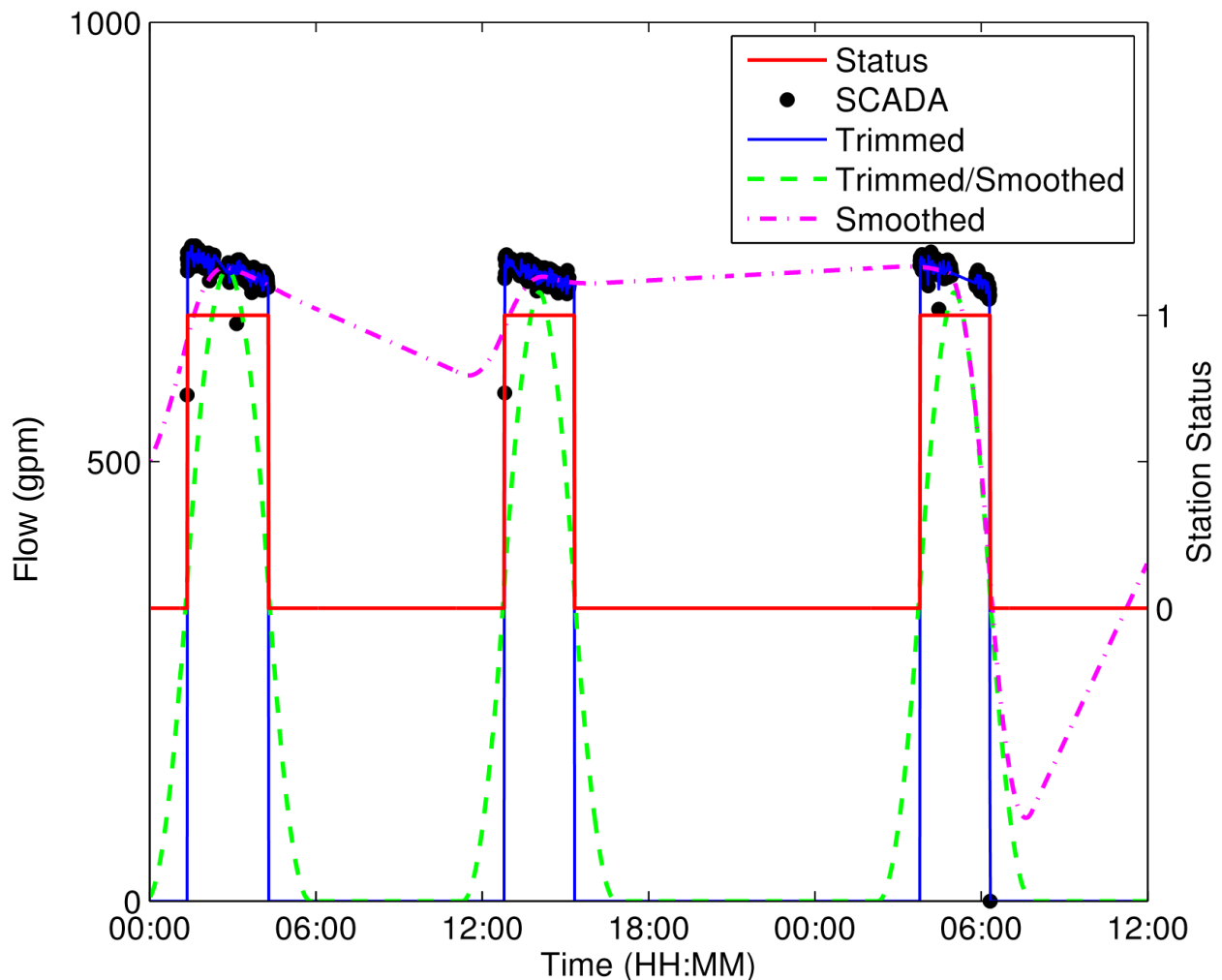


Figure 3.2-8. Raw pump station flow SCADA data (gallons per minute) and trimmed/smoothed station flow (left axis), along with station status (right axis) used to produce the trimmed data stream. Between pump on cycles, flow data gaps make it difficult to use simpler interpolation methods, which could lead to significant non-zero flows when all pumps are off.

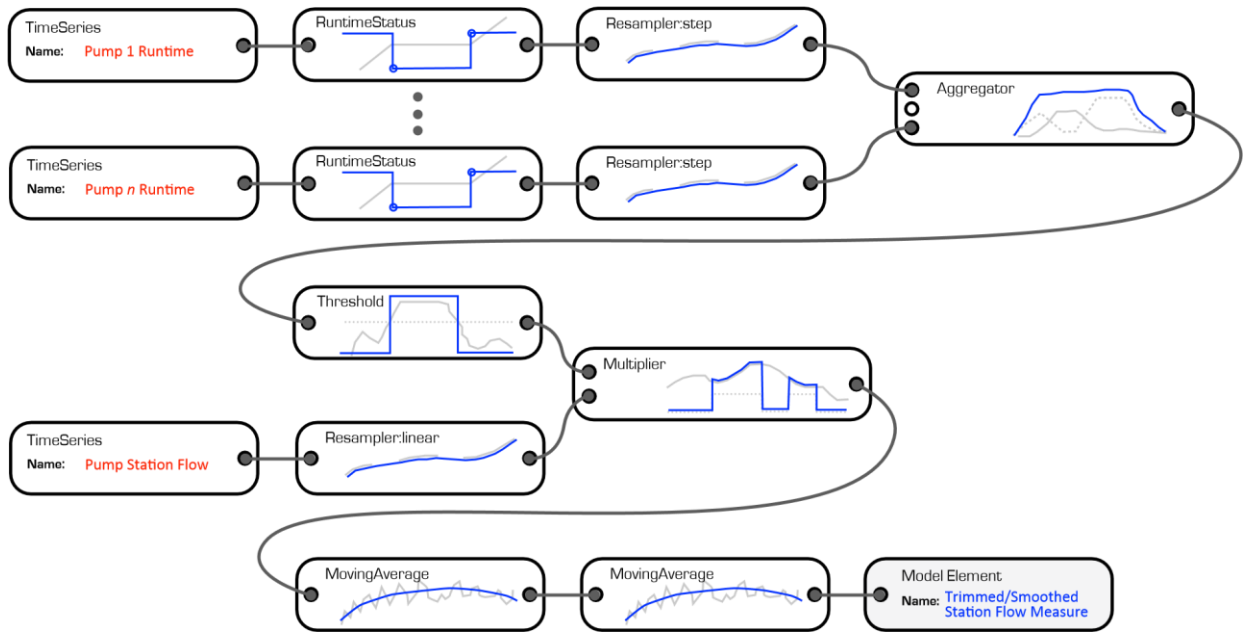


Figure 3.2-9. Time series pipeline for trimmed and smoothed pump station flow measures.

Trimming eliminates flow out of the pump station when all pumps are off; the resampled flow data stream is multiplied by the pump station status (the output data stream from the threshold transformation).

3.2.5 Altitude Valve Data Streams

Several tanks are equipped with altitude valves on their inlet/outlet pipes. A valve that closes whenever a set hydraulic grade within the tank is exceeded, is modeled simply by setting the tank maximum elevation appropriately in the EPANET model input data. More sophisticated valves, however, will open after closing only once the hydraulic grade drops below another, lower, set level. Under these conditions, the tank level can only drop (through a bypass check valve around the altitude valve).¹² Consider, for example, the SCADA tank level data in Figure 3.2-10. These data show extended periods during which the tank level either is not changing or dropping, consistent with the presence of a controlling valve and bypass, as described above. If the operation of such altitude valves is ignored, there is little chance that the real-time model will match observed behavior.

Unfortunately, no SCADA data streams record the status of these altitude valves directly. The implemented approach was to reconstruct these “missing” SCADA status streams by inference from the tank level data. The time series data transformation pipeline is shown in Figure 3.2-11; this pipeline is similar to that used to calculate pump status with a First Derivative object, but this time we filter the data series first, as was done for the tank levels. The Resampler has a clock of 1 minute, the MovingAverage objects each use a window size of 19, and the Threshold object sets the altitude valve status to close if the rate of change in tank level drops below 0.2 ft/hr. Representative results from this time series pipeline is shown in Figure 3.2-10 (Status). When this data stream is assigned as a status boundary for the tank inlet/outlet pipe, it effectively shuts off flow to or from that tank, consistent with the SCADA record.

¹² Currently, the hydraulic model does not include such bypass piping. It could be argued that it should.

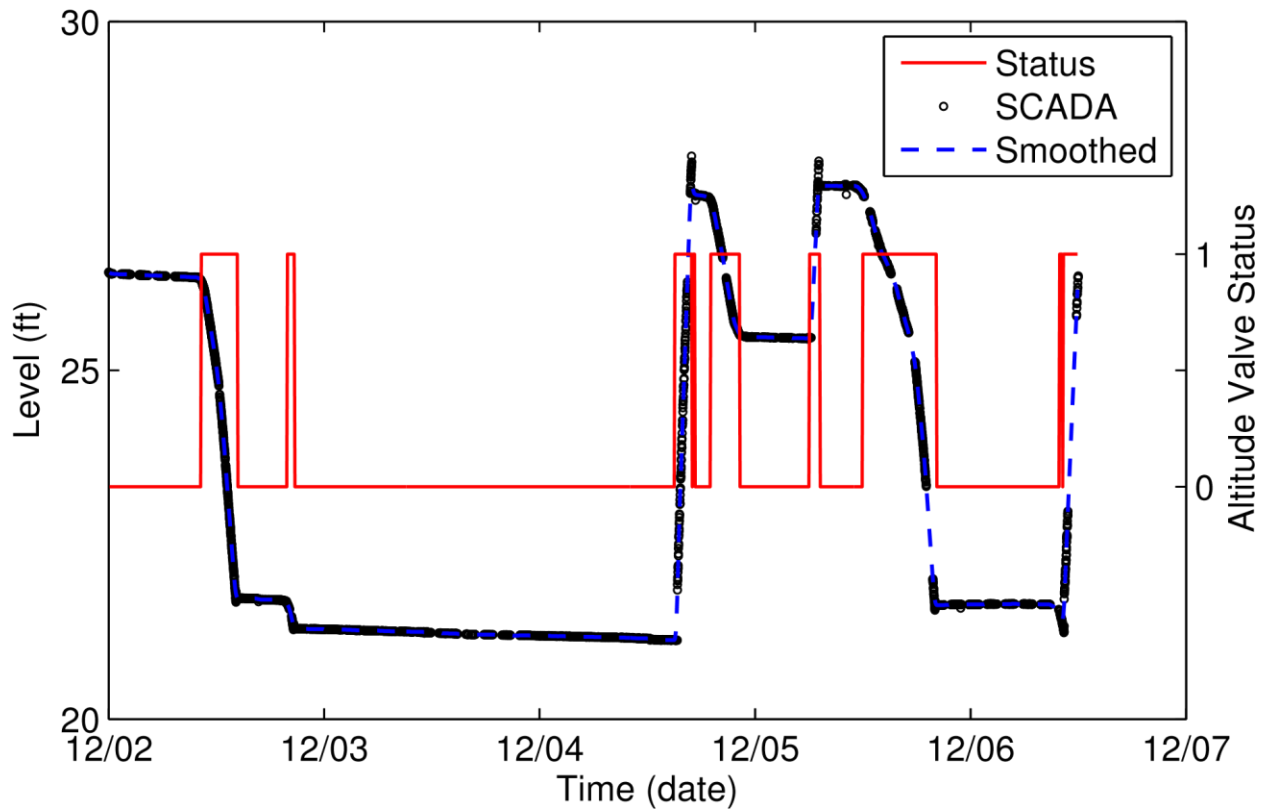


Figure 3.2-10. Raw and smoothed tank level data (left axis), along with altitude valve status (right axis) used to model control action of altitude valves on specific tank inlet/outlet lines.

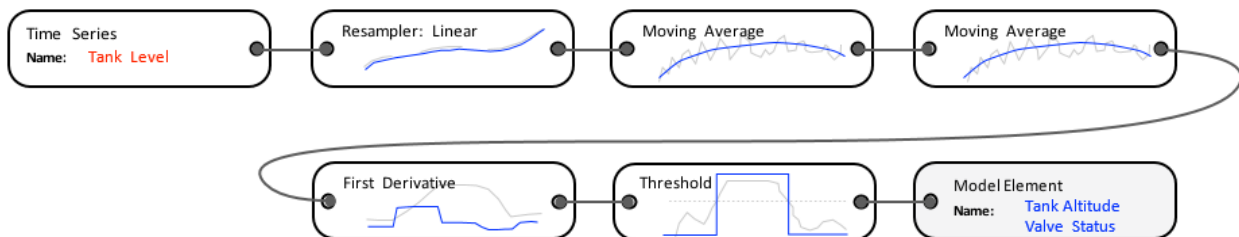


Figure 3.2-11. Time series data transformation pipeline to determine status of tank inlet/outlet pipe with altitude control valve.

3.2.6 Reconstruction of Missing Plant Production Data Stream

The northern TP feeds DMA 2 by gravity from its clearwell, yet there is no flow sensor that monitors flow out of the clearwell. This flow is a critical component of the DMA 2 real-time demand calculations, so effort was made to re-create that flow from other available data sources. Figure 3.2-12 is a schematic of the essential northern TP infrastructure¹³. Actiflo® (Veolia, L'Aquarène, Saint Maurice, France) is a process in which water is flocculated with seeded particles in specialized tubes. The missing flow

¹³ There are filters in between the Actiflo units and the clearwell, but filter flow data were mostly missing from SCADA. Thus filters were omitted from the diagram, showing only the Actiflo units where flow data were available.

measure out of the clearwell is indicated on the schematic (F), as are the available data for Actiflo flow rates (F), and clearwell level (L).

Given the available data, as well as the clearwell geometry, a flow balance on the clearwell can be used to calculate a replacement for the missing flow,

$$\frac{dV}{dt} = F_a + F_b - F, \quad (1)$$

or,

$$F = (F_a + F_b) - \frac{dV}{dt}. \quad (2)$$

The Actiflo flows F_a and F_b are both available in SCADA, and the rate of clearwell volume change, dV/dt , can be estimated from the clearwell level. The EPANET-RTX time series data transformation pipeline that implements this strategy is shown in Figure 3.2-13. The bottom half of the figure constructs the total Actiflo flow rate by adding the resampled individual Actiflo flows, and filtering them with a single moving average. Typically, the Resampler clock was 1 minute, and the moving average window for the summed flow was 91 data points. The top half of the figure constructs the rate of volume change in the clearwell — or the net clearwell inflow — which is identical to how tank level will be converted into flow for DMA demand computation. The clearwell level is resampled and filtered, and assigned as the source to a CurveFunction object, which uses the clearwell geometry to convert (smoothed) level into volume. The FirstDerivative object then calculates the slope of the smoothed volume versus time data stream, approximating dV/dt . These two data streams are then aggregated as in Equation 2 to produce the estimate of supply flow leaving the clearwell. Figure 3.2-14 shows representative results from the above time series data transformation pipeline, including the total Actiflo flow ($F_a + F_b$), the clearwell net inflow (dV/dt), and the resulting supply flow (F). The smoothed clearwell level is also shown on a separate axis.

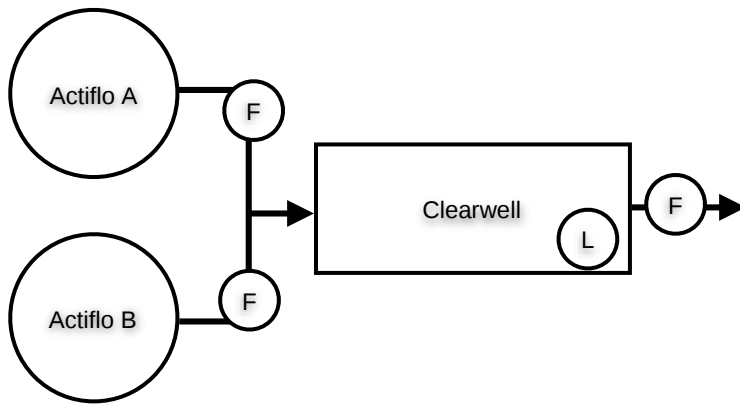


Figure 3.2-12. Treatment plant clearwell schematic showing flow and level measures used for construction of boundary flow data stream.

Water Utility Case Study Using EPANET-RTX Libraries

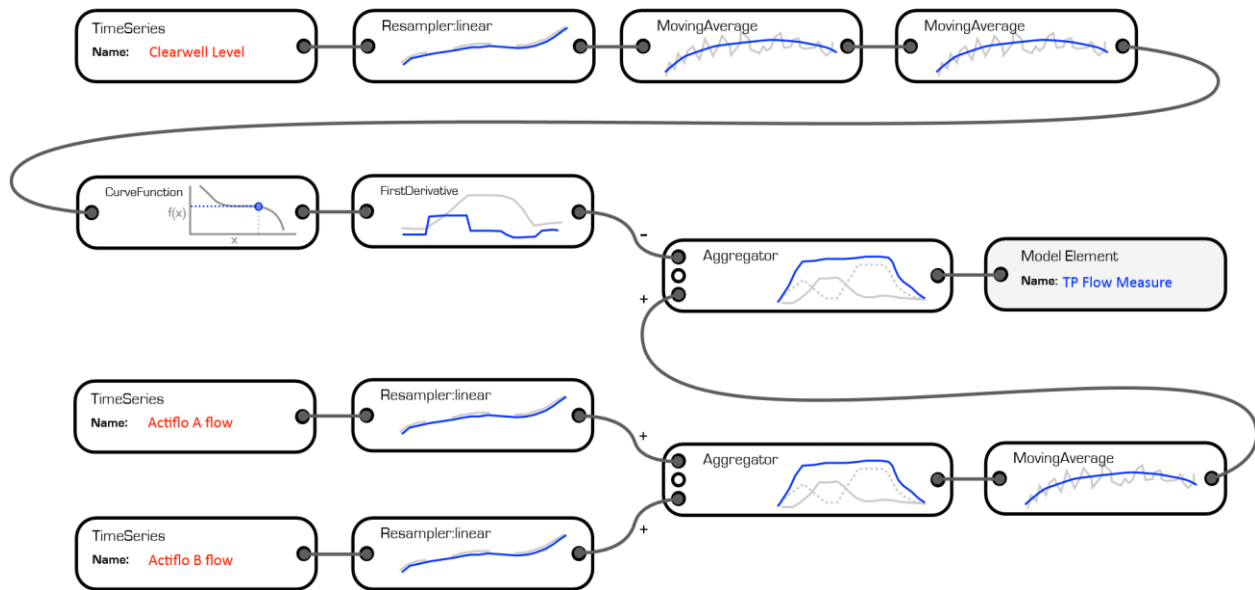


Figure 3.2-13. Time series data transformation pipeline to determine flow measure at northern treatment plant, from conservation of fluid volume within clearwell.

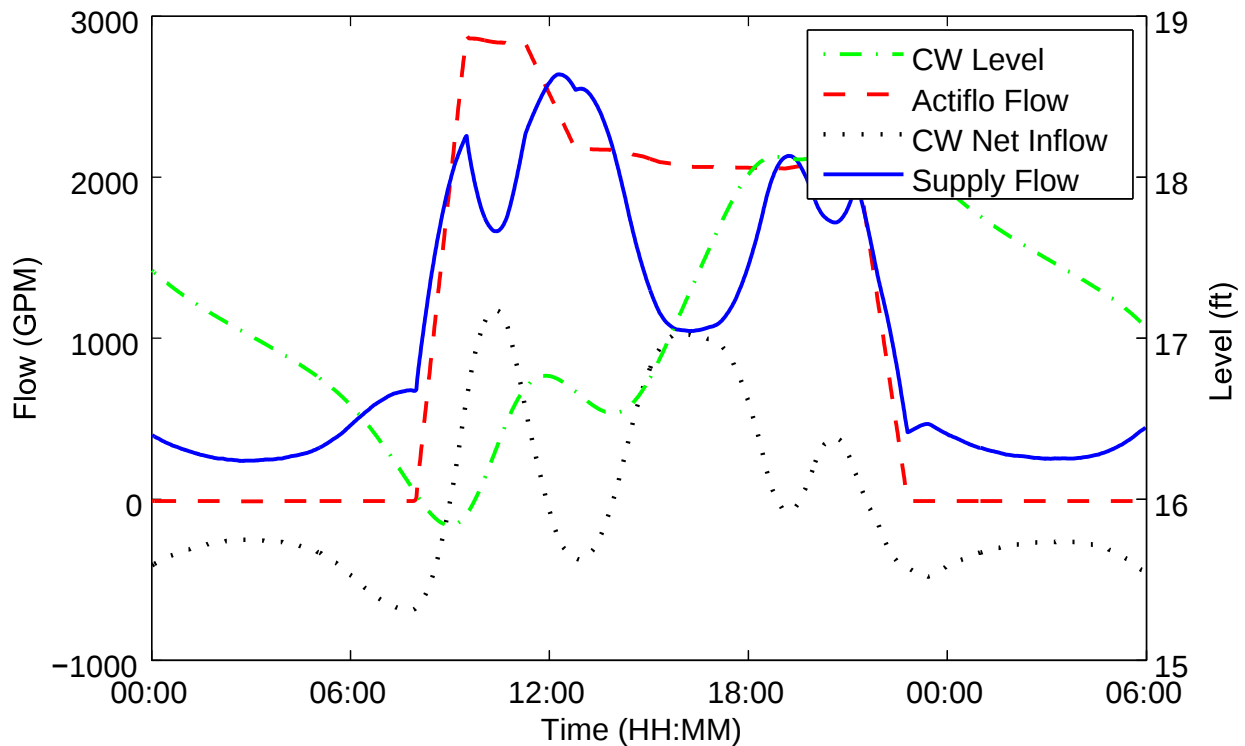


Figure 3.2-14. Representative results from the associated time series data transformation pipeline, including the total Actiflo flow ($F_a + F_b$), the clearwell net inflow (dV/dt), and the resulting supply flow (F). The smoothed clearwell level is also shown on a separate axis. (GPM, gallons per minute.)

3.3 District Metered Area Real-Time Demands

The district metered area, or DMA, is a demand management concept introduced in the United Kingdom in the early 1980s. The United Kingdom (UK) Report 26 (Association, 1980) defined a DMA as an area of a distribution system that is specifically defined, e.g., by the closure of valves, and for which the quantities of water entering and leaving the district are metered. DMAs are an essential component of demand management in the UK and Republic of Ireland, Dublin for example, historically because of the lack of domestic customer metering. Not only do DMAs allow the utility to understand the spatial and temporal pattern of demand, they are also used to estimate and control leakage. Leakage control is implemented by focusing on statistical analysis of minimum nightly usage rates within each DMA. It is assumed that the night usage is composed of relatively stable customer usage plus leakage. Thus, as infrastructure improvements are implemented, one expects to see consequent reductions in night usage rates, attributed to reductions in leakage. Relatively rapid increases in night usage rates indicate new bursts or continued deterioration of existing bursts; these incidents then initiate an intensified focus on leak identification and repair. For example, Dublin, serving 1.5 million customers with a supply rate of 143 MGD, developed an infrastructure monitoring strategy that relies on DMAs consisting of between 1,000 and 2,000 customer connections, and a demand of approximately 0.7 MGD; as a result the Dublin distribution system is divided into approximately 200 DMAs. An equivalent subdivision of the NKWD study area would require 10 DMAs instead of three. Such an increase in flow instrumentation density would likely have a positive impact on the accuracy of real-time demands and model predictions.

It is possible to confuse a DMA with a pressure zone; their boundaries may often be similar, simply because pressure zone boundaries are often defined by pump stations, which are often points of measured flow. Yet there is little fundamental to link these two ways of organizing network elements. Pressure zones regionalize locations based on hydraulic head, and DMAs regionalize locations based on a common set of water sources and sinks. And as is clear from Figure 2.1-1, a single DMA can contain multiple pressure zones (and vice versa).

3.3.1 DMA Demand Time Series Data Transformation Pipeline

Each DMA is described completely by its set of boundary pipes, limited to those with a valid flow measure, a status set to closed (effectively, a measure of zero flow). Construction of the complete set of DMAs for a network is an algorithmic process defined by the infrastructure topology, flow measure locations, and pipe statuses. Each DMA is constructed in a straightforward procedure that involves traversing the network in a methodical manner (e.g., depth-first or breadth-first graph search) and recording the junctions that have been visited, including storage tanks. The network search stops at all boundary pipes (measured flows, or closed statuses), and continues until all possible paths from DMA junctions have been explored. At the conclusion of this process, the DMA junctions and storage tanks are known, as are DMA's closed and measured boundary pipes. Boundary pipes that are either closed or measured will in general belong to multiple DMAs, but most often it would be either one or two. In that case a single boundary pipe serves to separate the DMA by virtue of knowing the flow measurement. Nodes will only belong to one DMA.

As an illustration of DMA construction, the boundary elements describing DMA 3 are shown in Table 3.3-1. This DMA is defined by five boundary elements, including one closed pipe, one tank, and three flow measures. The tank belonging to the DMA is considered a boundary element because it, too, serves as a possible water source or sink. The data include the "Model Element," the associated "Flow Measure Time Series" data stream, a "Multiplier" for the DMA demand aggregation, and a brief "Description" note. Note in particular the boundary element for the South Newport Tank. The listed flow

measure “SNEWPORT flow” is not a physical flow measure, but rather a calculated flow measure based on tank water level and geometry; the flow measure name was, for convenience, constructed from the tank identifier “SNEWPORT” prepended to the string “flow.” The tank flow measure is assigned to the tank element instead of inlet/outlet piping, as that assignment more accurately represents the tank as a source/sink for the DMA. All tank flow measures, by definition, have a multiplier of -1 , because a positive rate of tank volume change represents removal of water from the DMA.

Table 3.3-1. Summary of boundary elements for DMA 3 demand time series aggregation.

Model Element	Flow Measure Time Series	Multiplier	Description
16004			Closed Pipe
CAROTHERP1	Carothers Rd. Pump 1 Flow	+1	Pump 1 @ Carothers Station
CAROTHERP2	Carothers Rd. Pump 2 Flow	+1	Pump 2 @ Carothers Station
SNEWPORT	SNEWPORT Flow	-1	South Newport Tank
ST THER REG	St. Therese Regulator Flow	+1	St. Therese PRV

DMA, district metered area; PRV, pressure reducing valve

The data for DMA 3 in Table 3.3-1 can be expressed more usefully as a DMA demand time series data transformation pipeline, through a flow balance on the DMA. While the detailed boundary elements will vary from one DMA to the next, the template for the demand time series data transformation pipeline does not, and so can be automated for any utility’s network model. The data transformation pipeline for DMA 3 is shown in Figure 3.3-1. The four flow measures are represented by their model element connections: CAROTHERP1, CAROTHERP2, ST_THER_REG, and SNEWPORT. Each of these model elements has a time series data transformation pipeline that is responsible for generating it — but for clarity most of these details are omitted from Figure 3.3-1. What is shown, however, is the piece of the time series data transformation pipeline that converts tank level to net inflow, using the Curve Function and First Derivative objects, prior to aggregating the boundary flows.

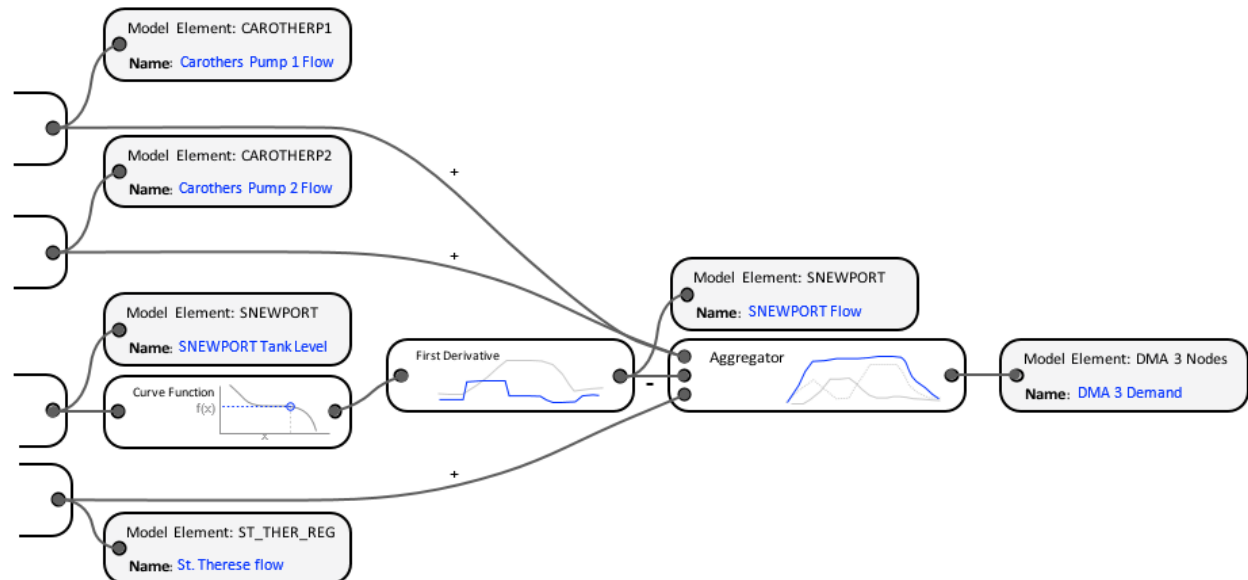


Figure 3.3-1. Time series data transformation pipeline constructed automatically by EPANET-RTX to aggregate boundary flows (DMA 3 demand). Time series data transformation pipelines for boundary model elements (not shown) are specified as part of real-time model configuration.

The general process of producing real-time DMA demands, including the identification of DMAs and their boundary elements, and the construction of the demand time series data transformation pipelines for each DMA (e.g., the objects in Figure 3.3-1 for DMA 3), are automated by EPANET-RTX algorithms. Figure 3.3-2 shows representative calculated demands for all three DMAs; these demands drive the real-time simulation results for the 1-week period examined in Section 4.2.

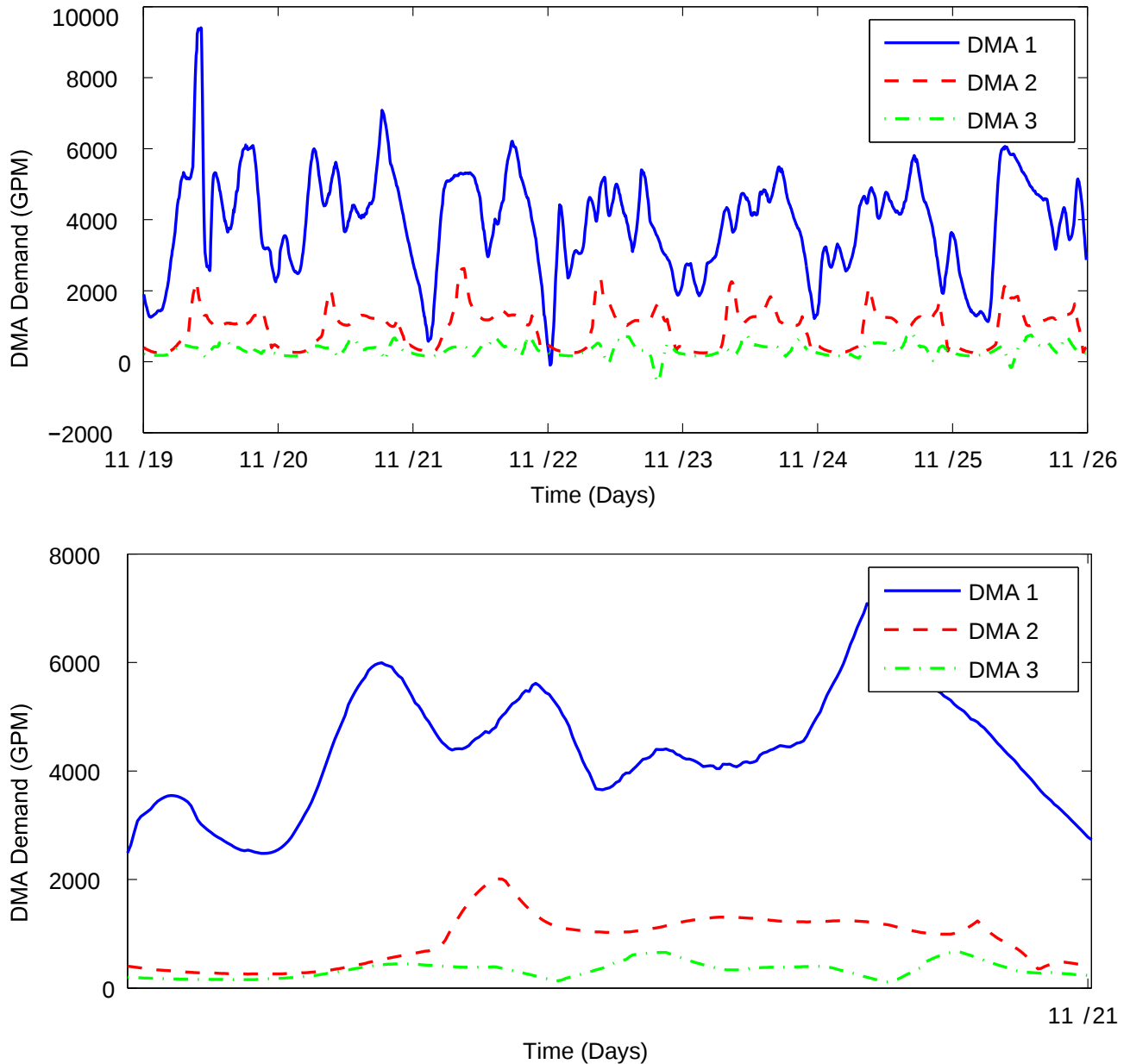


Figure 3.3-2. Real-time aggregate demand (gallons per minute) for DMAs 1 through 3 for November 19 through 26, 2012 (top), and expanded for a 1-day period within the same time frame (bottom).

3.3.2 DMA Demand Disaggregation

Real-time DMA demands are disaggregated to DMA junctions according to their modeled average demand. The modeled average demand for junction j , $\bar{d}_j$ is defined,

$$\bar{d}_j = \sum_{q=1}^m \left(\frac{b_j^q}{n^q} \sum_{k=1}^{n^q} p_{jk}^q \right), \quad (3)$$

where m is the number of demand categories; b_j^q is the base demand for category q and junction j ; p_{jk}^q is the k^{th} demand pattern factor assigned to junction j and demand category q ; and n^q is the number of factors in the category q demand pattern. Given these average demands for each junction belonging to DMA i , the real-time junction demands at time t are assigned,

$$d_j(t) = \frac{\bar{d}_j}{\sum_{j \in S_i - B_i} \bar{d}_j} (D_i(t) - \sum_{j \in B_i} b_j(t)), \quad j \in S_i - B_i \quad (4)$$

$$d_j(t) = b_j(t), \quad j \in B_i, \quad (5)$$

where S_i is the set of all junctions belonging to DMA i , B_i is the set of boundary flow junctions belonging to DMA i (these are junctions with measured demand, such as a wholesale master meter tied into SCADA), $D_i(t)$ is the calculated DMA demand at time t , from the DMA demand time series data transformation pipeline, and $b_j(t)$ is the measured boundary flow from the time series data transformation pipeline associated with junction j . Expressed in words, the measured junction demands are assigned their measured values, and the remainder of the DMA demand $D_i(t)$ is distributed to the non-measured junctions in proportion to their average model demand.

4.0 Real-Time Model Calibration

In Section 4 we describe the real-time model calibration process used on the NKWD model. Section 4.1 describes the calibration process. Section 4.2 describes the real-time hydraulic simulation process that was performed on the NKWD model, along with a quality evaluation of the simulation results using Pearson's correlation coefficient. Section 4.3 uses an EPANET-RTX-based application to simulate the tracer movement in the NKWD study area and compares the simulation results to the conductivity measurement results obtained from the field tracer study.

We also provide a broad summary of modifications that were made to the NKWD network (infrastructure) model in support of the case study. These modifications were, in general, not specific to a real-time simulation capability. A catalog of the modifications that were made is provided in Appendix B. Recommendations for further model improvement are provided in Appendix C.

4.1 Calibration Process

Many processes that would typically be considered part of network model calibration are implemented automatically by an EPANET-RTX-based real-time model. The statuses and settings of controllable model elements are all determined by the real-time data transformations, and the real-time demands are determined by aggregating boundary element flows for each DMA; these processes were described in Sections 3.2 and 3.3.

The calibration process described here is consistent with macro-calibration, as it is known in the field (Ormsbee and Lingireddy, 1997). Macro-calibration deals with inconsistencies in infrastructure

representation (e.g., pump characteristic curves, tank geometry, valve statuses and settings, pipe diameters, or reservoir elevations) that can lead to large simulation errors. The macro-calibration process is typical of an engineering investigation, and follows a path of identification of errors, generating hypotheses about their causes, gathering and analyzing relevant data, and reassessing model results. We performed no micro-calibration activities as part of this case study — we did not seek to optimize pipe roughness or node base demands, in order to maximize or minimize an error criterion. Such activities can be useful, but care must be taken not to over-parameterize the process, and in so doing jeopardize the physical validity of the parameter estimates. Because this was the first large-scale effort to calibrate a real-time model and to describe its accuracy, it was decided not to “fine-tune” model parameters in ways that may make it more difficult to interpret results.

Several model calibration activities were initiated prior to any real-time simulation results being generated — indeed, prior to configuring the real-time model. These activities were implemented across the network model and summarized as follows:

1. Updates to pressure reducing valve (PRV) settings and elevations based on field measurements
2. Updates to tank geometry and elevations, based on utility light detection and ranging (LIDAR) elevation data, field measurements, and SCADA data obtained while purposefully filling the tanks to overflow
3. Updates to pump characteristic curves based on analysis of SCADA-derived total dynamic head, pump station flow, and pump status

The macro-calibration process then was driven by real-time simulation results. While Appendix B catalogs the model modifications that were ultimately required, the calibration process was structured according to the DMAs. Simpler DMAs such as DMA 3 were considered first, as it had one main source of supply, no downstream DMAs to interact with, and a single storage tank. Attempts were made to correct problems where the error was clearly within the DMA, before moving to the next. DMA boundary flows were generally examined first, to verify the status of pumps and station discharge. If flows were off significantly, consideration was given to adjustment of pump characteristic curves. Once modeled boundary flows were judged to be reasonable given the data, storage tank elevations were considered. As a quality assurance step on EPANET-RTX data processing, the DMA demand aggregations were constructed separately using the simulated flows, rather than the SCADA flow and level data (as performed by EPANET-RTX). The two computations were expected to be identical — and were observed to be — as EPANET-RTX is setting the real-time demands based on the SCADA data, and the hydraulic simulations must balance the flows within each DMA.

4.2 Real-Time Hydraulic Simulation

A real-time extended period simulation model was run in a continuous retrospective mode for a 1-week evaluation period, from midnight, November 19, 2012, through midnight, November 26, 2012¹⁴. For the real-time model configuration described above, the results are identical to what they would have been, if propagated in real-time during that 1-week period in 2012. No special data processing was performed, beyond the data transformations described in Section 3.2 and its associated subsections for the real-time model configuration. Initial reservoir heads and tank levels were reset to their transformed SCADA values at the beginning of the evaluation period, but after that time they evolved with the

¹⁴ This 1-week period includes the Thanksgiving holiday in the United States, Thursday, November 22.

extended period hydraulic solution. The 1-week simulation results were obtained using an EPANET-RTX application, and a real-time configuration as specified in an EPANET-RTX configuration file; the simulation was then driven automatically by EPANET-RTX.

The data used for evaluation are those available in the SCADA record for the study area over the 1-week evaluation period, as presented in Sections 2.2 and 2.3. These include the data streams for 15 pressure measures, 8 flow measures, and 10 tank levels, within the three DMAs. In addition, time series plots of pump station flows contain useful visual information about the actual versus simulated pump statuses.

Figure 4.2-1 summarizes the quality of real-time simulation results, for all individual measurements, using Pearson's correlation coefficient, $0 \leq \rho \leq 1$. The correlation coefficient measures the linear relationship between the simulated and measured values. Another interpretation is the fraction of variability in the measured signal that is explained by the simulated signal. Since explaining variability is in some sense the purpose of a dynamic simulation, the correlation coefficient is a useful measure of simulation accuracy. Average correlation coefficients for pressure, flow, and tank level data streams were 0.83, 0.79, and 0.81, respectively. Flow measurements 4 and 7 do not have zero correlation coefficients, as would seem to be indicated by Figure 4.2-1; these correlation coefficients are mathematically undefined and thus left out of the average computation, because the associated SCADA flow streams (both flows through regulating valves) were constant and equal to zero. In both cases, the simulated flows closely approximate those measured flows.

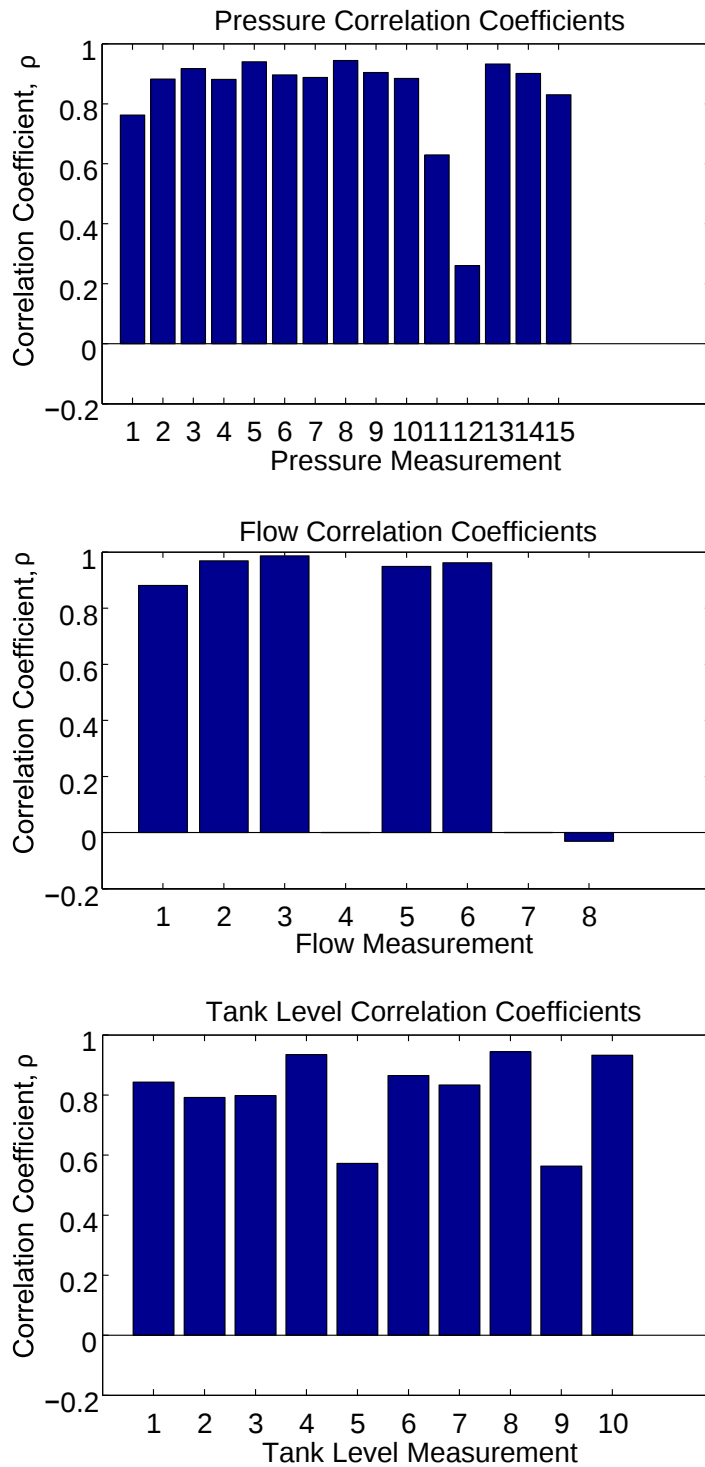


Figure 4.2-1. Pearson's correlation coefficients between measured and real-time simulated heads, flows, and tank levels.

A useful way to interpret the real-time model accuracy is through time series plots of measured and simulated values. These show the time variation in the measured and simulated signals, and also give an easy visual indication of bias, or difference in mean values. Results for the 15 pressure data streams, converted into hydraulic head, are in Figures 4.2-2 through 4.2-6; for the seven¹⁵ flow data streams, in Figures 4.2-7 through 4.2-9; and for the 10 tank levels, in Figures 4.2-10 through 4.2-13. In each figure, the data points are the red circles, and the EPANET-RTX simulated values are the blue solid lines. The title of each figure identifies the model element identifier, as well as the value of the associated correlation coefficient. For each graph, the data range was allowed to adapt to data stream characteristics, to provide better resolution of the variability over the 1-week period; thus it is important to note the scale when comparing results across the different measurements.

¹⁵ There should be two additional PRV flow measures compared as indicated in Table 2.2-1 (St. Therese and Memorial Newport regulators) for a total of 9 flow data streams. The data collection and generation of figures for this report were generated from a special purpose computer code specifically developed for this demonstration project (actually an EPANET-RTX-based application) and, unfortunately, the St. Therese and Memorial Newport regulator data were not obtained and processed correctly and, therefore, were not included in the analyses that follow. In the special purpose EPANET-RTX application developed, the software program included a search routine to examine each of the link types: valves, pipes, and pumps, and identify whether any were associated with a flow measure. If a flow measure was found, then the measured data as well as simulated data were collected and written to a file for plotting. For reasons still being investigated, the St. Therese and Memorial-Newport measured flow data were corrupted and not used in the analyses presented here.

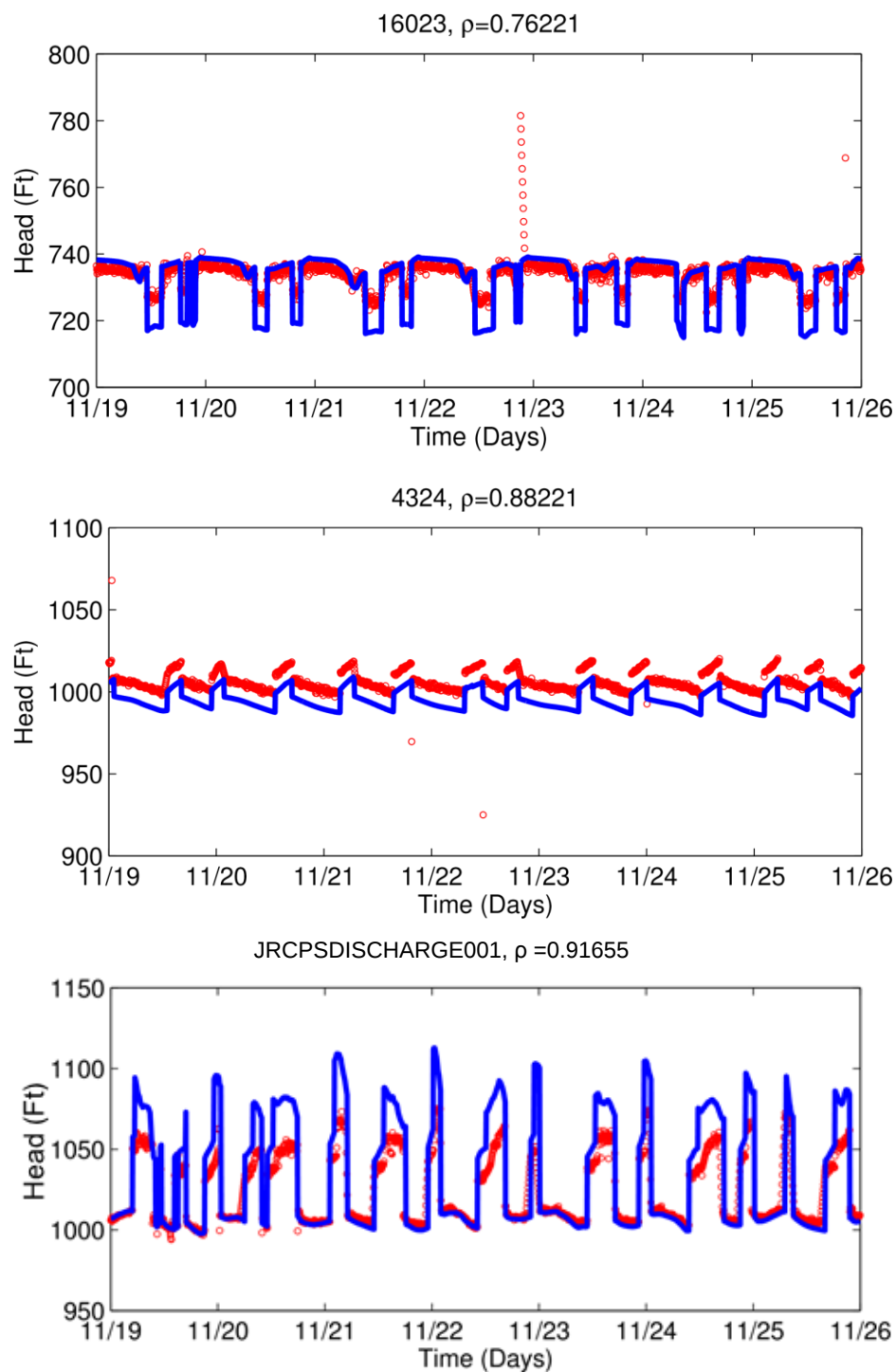


Figure 4.2-2. Measured and real-time model simulated heads.

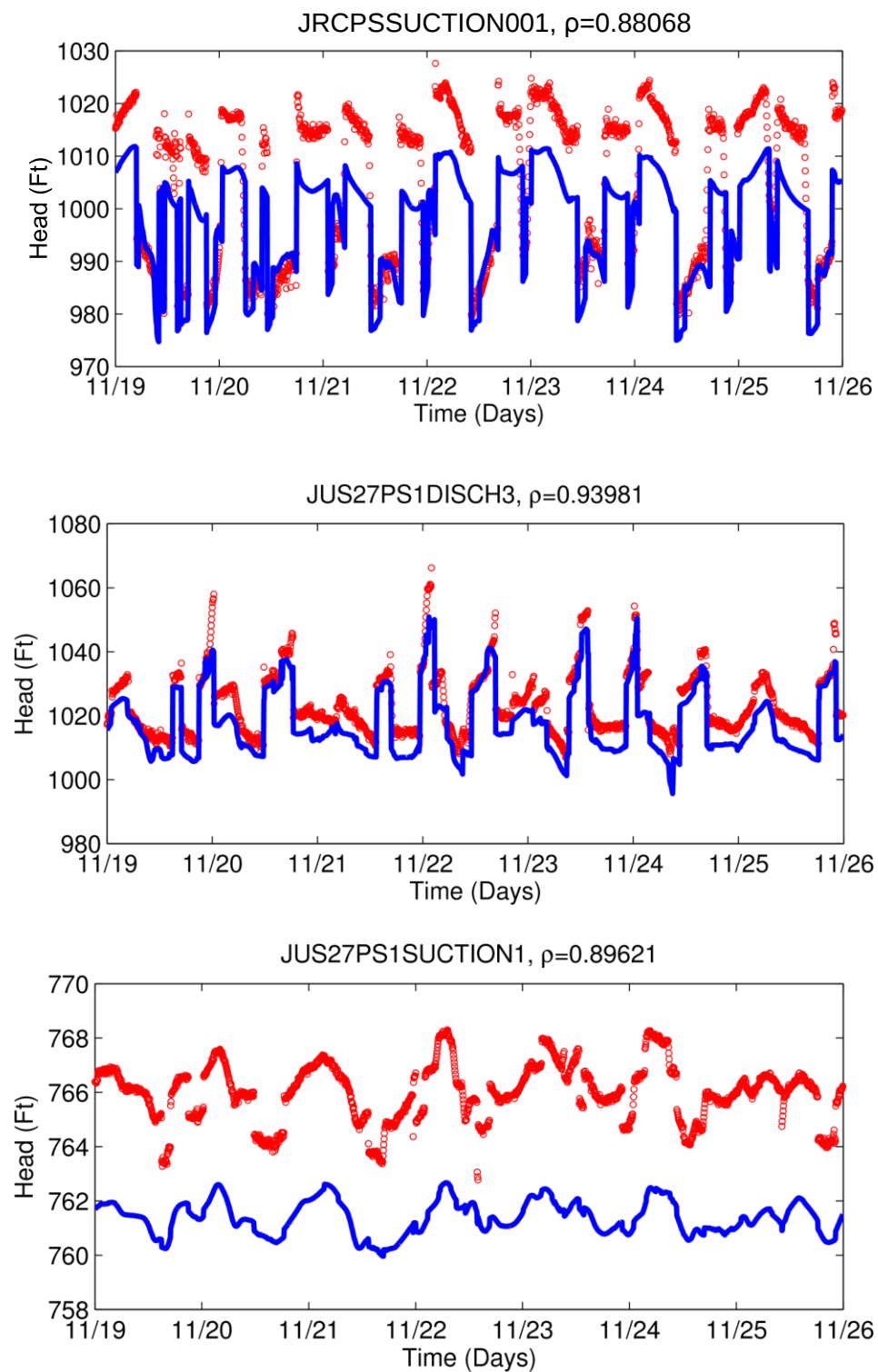


Figure 4.2-3. Measured and real-time model simulated heads.

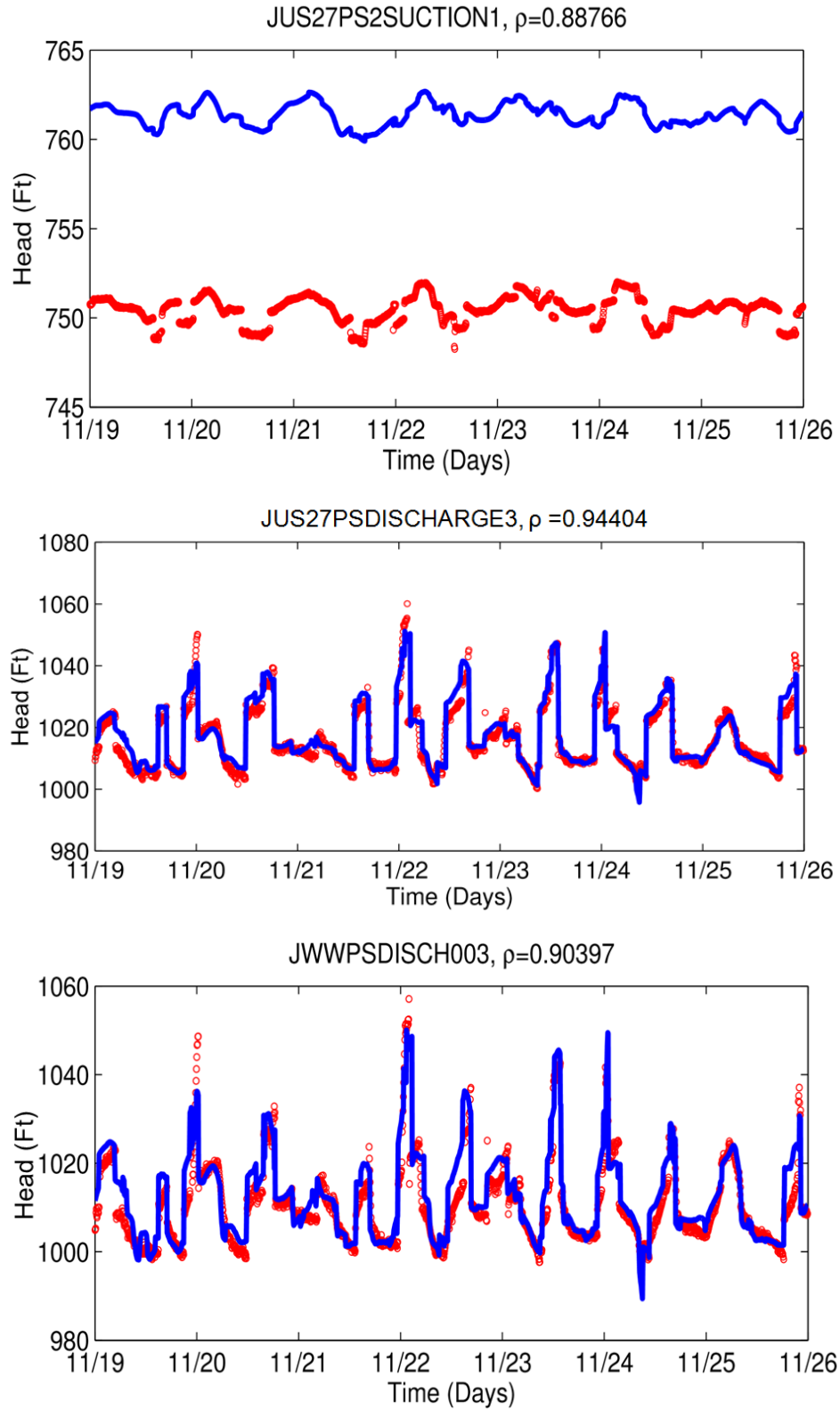


Figure 4.2-4. Measured and real-time model simulated heads.

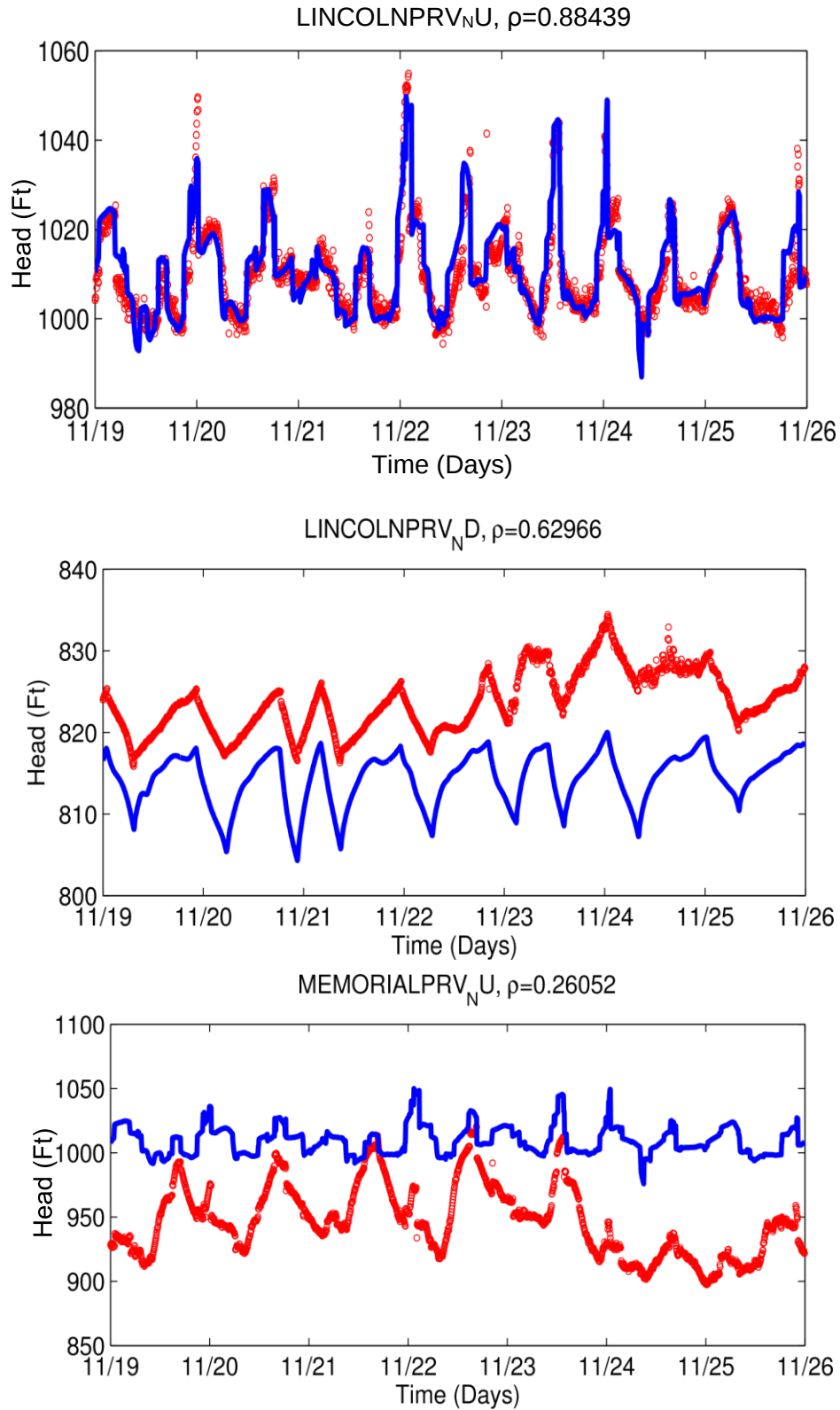


Figure 4.2-5. Measured and real-time model simulated heads.

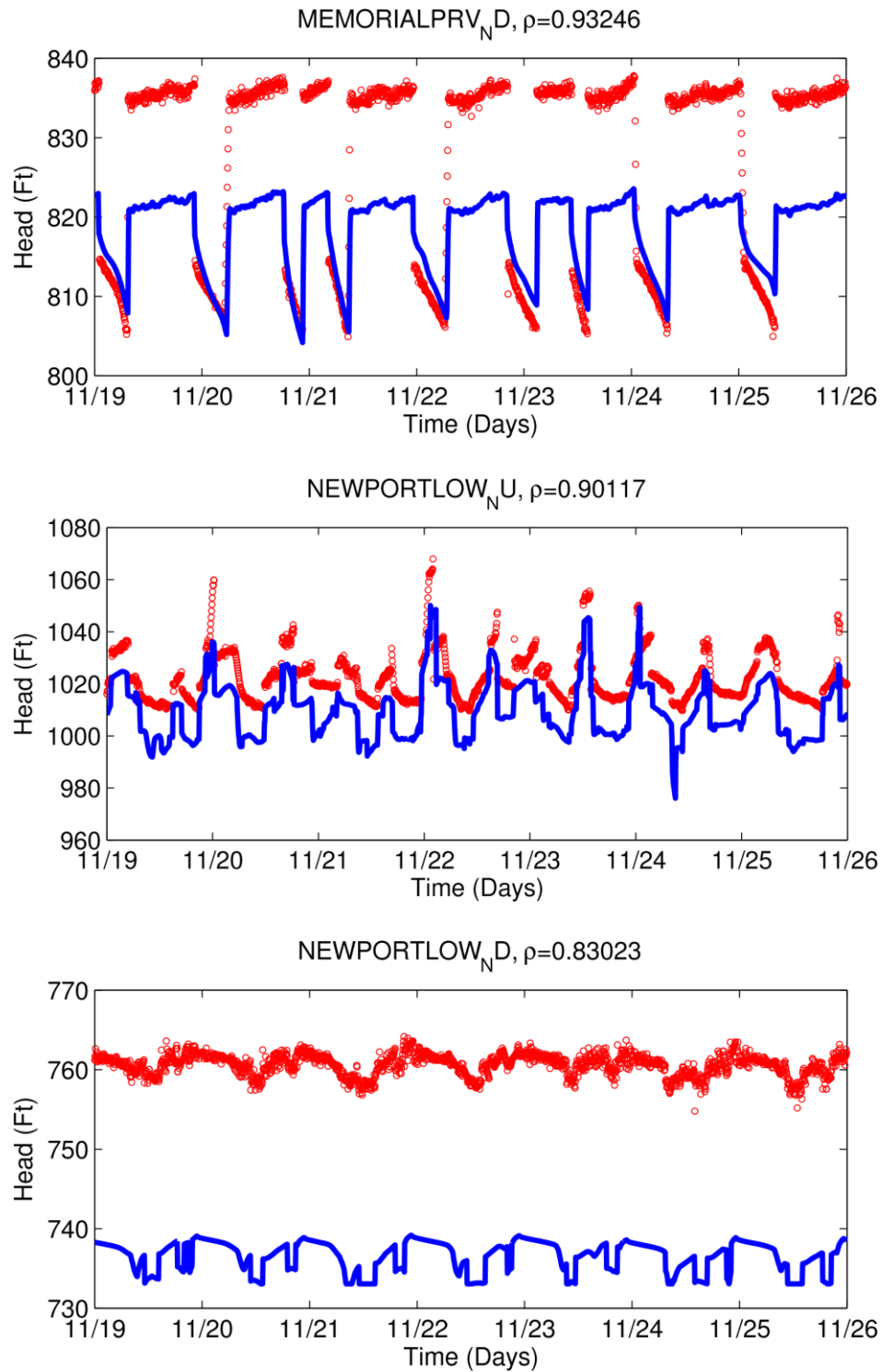


Figure 4.2-6. Measured and real-time model simulated heads.

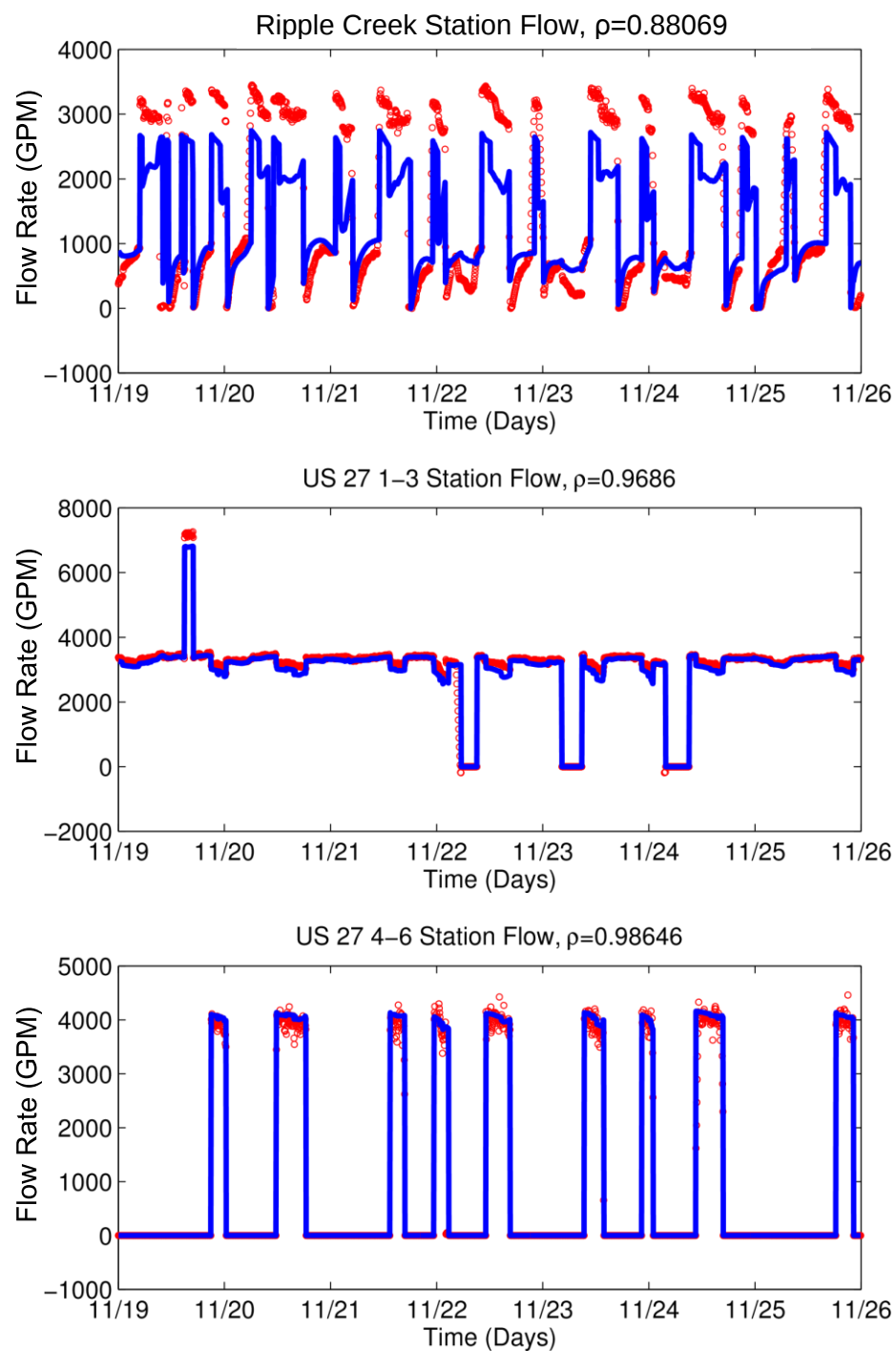


Figure 4.2-7. Measured and real-time model simulated flows.

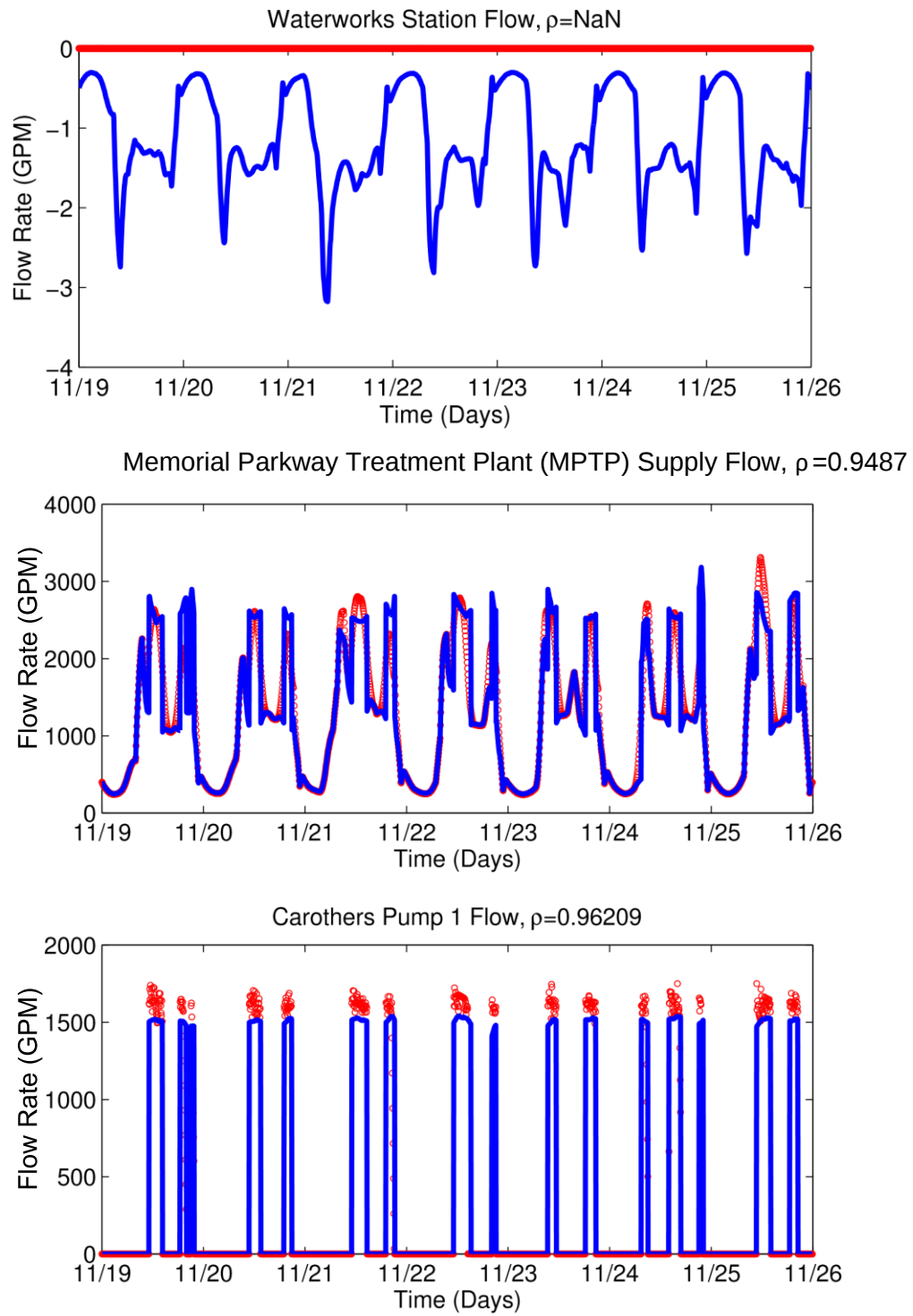


Figure 4.2-8. Measured and real-time model simulated flows.

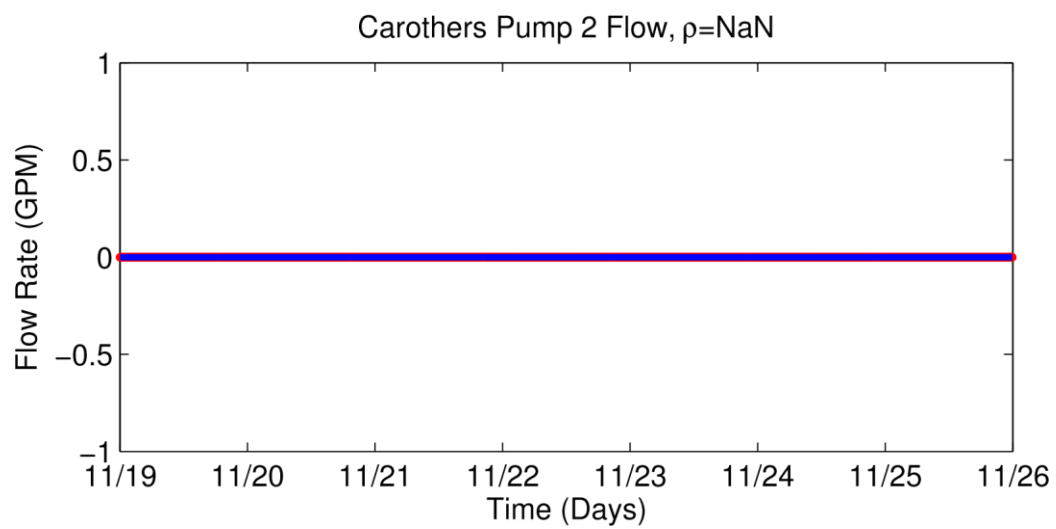


Figure 4.2-9. Measured and real-time model simulated flows.

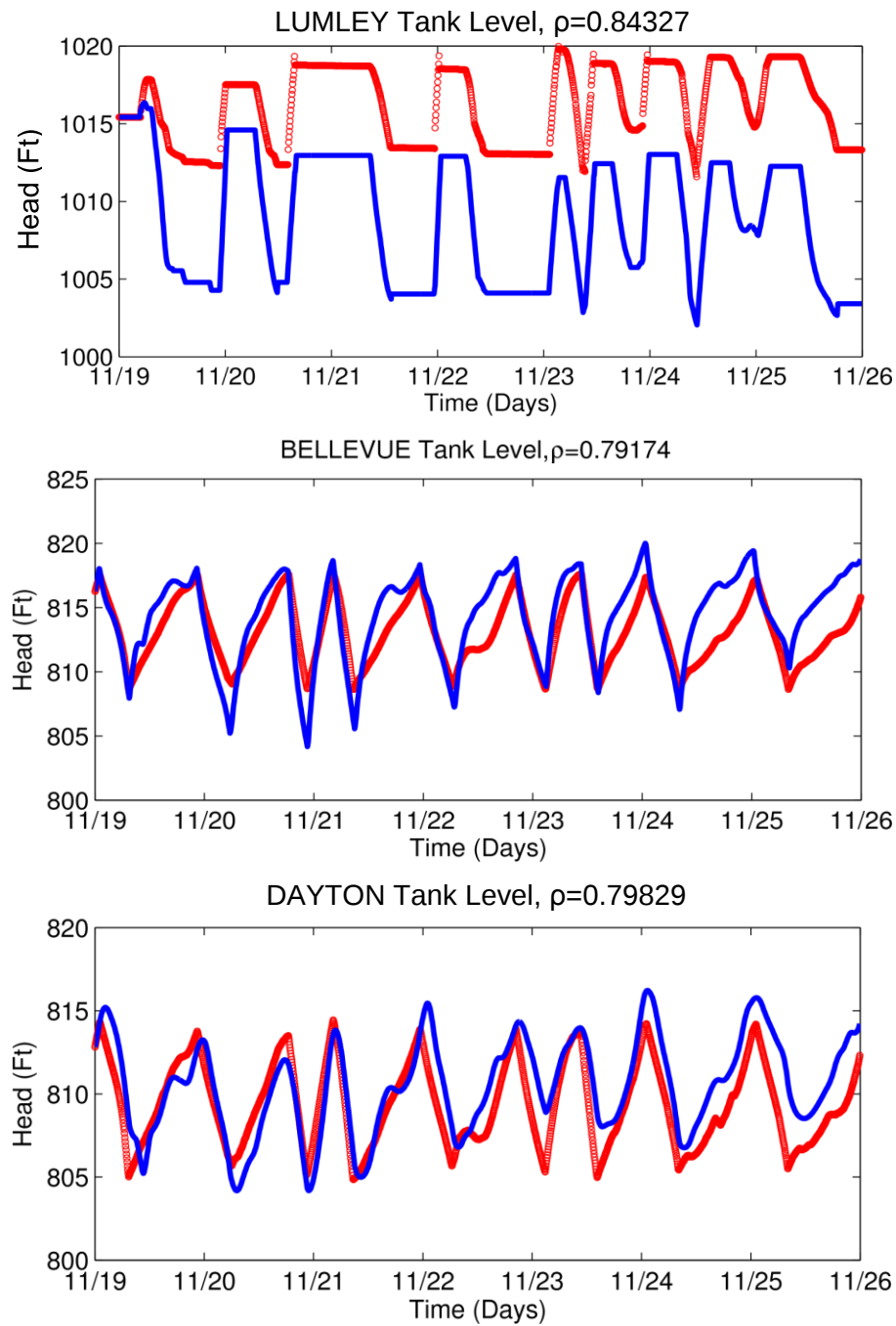


Figure 4.2-10. Measured and real-time model simulated tank levels.

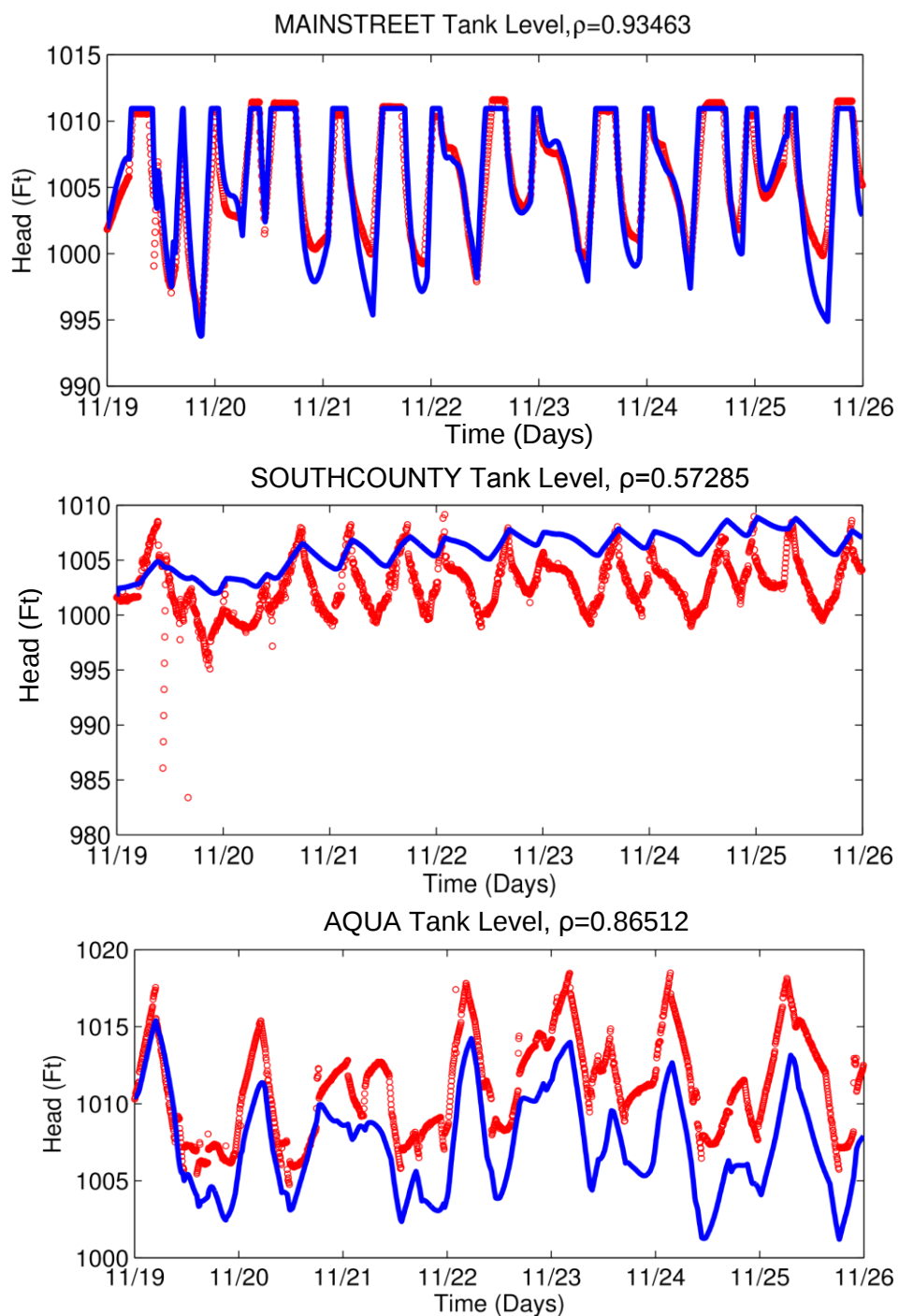


Figure 4.2-11. Measured and real-time model simulated tank levels.

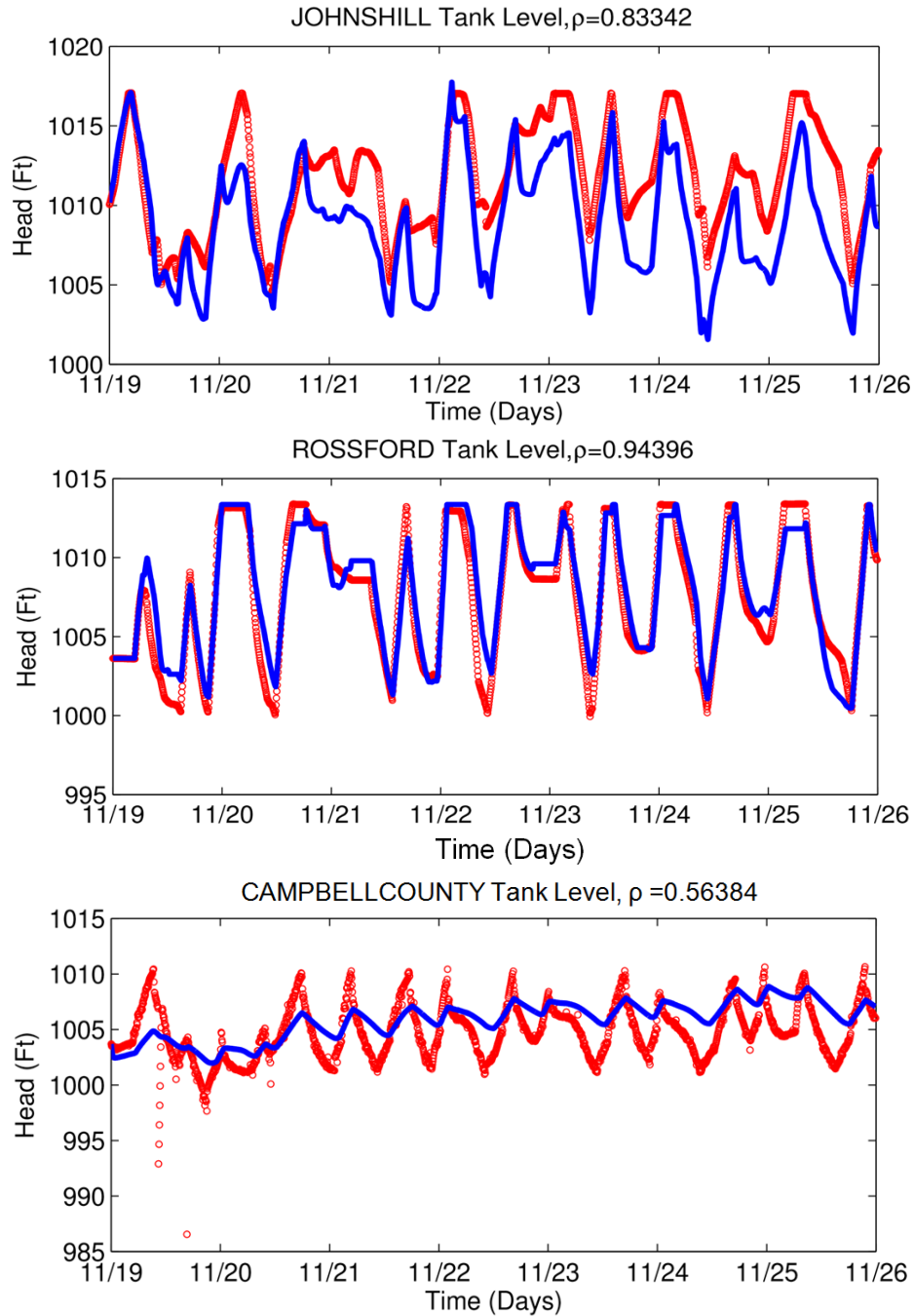


Figure 4.2-12. Measured and real-time model simulated tank levels.

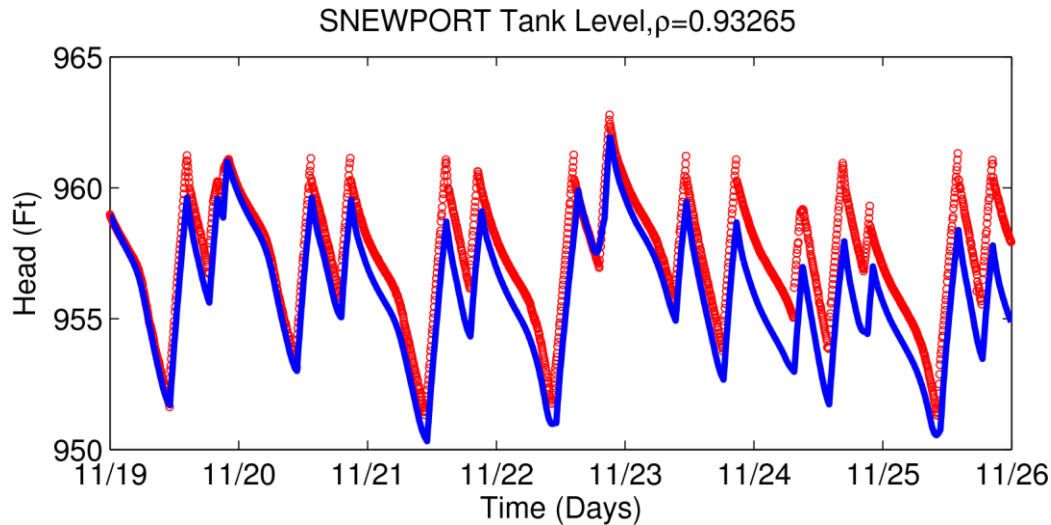


Figure 4.2-13. Measured and real-time model simulated tank levels.

In general, the real-time simulation results accurately reproduced the hydraulic behavior of the distribution system, as described by this set of SCADA measurements. This does not mean that the real-time model is validated, as we would rather have a denser grid of data points, as well as performing similar evaluations using the same network model at different times of the year, or in different operational modes. The results show some areas where improvements are needed, and we have outlined some suggested macro-calibration issues to be looked at further in Appendix C. Some of the tanks exhibit significant errors in the mean values between modeled and measured results. Two tanks — the Campbell County and South County tanks — exhibit relatively poor correlation between modeled and measured results. These latter two tanks are adjacent to each other and can be described as sluggish in terms of model performance. Compared to measurements, the tanks are not responding to either booster pumping or demand in a way that closely mimics reality. Indeed, these two modeled tanks are filled by the Ripple Creek pump station, shown in Figure 4.2-7, and the Ripple Creek modeled pumps are significantly undersupplying flow when they are on. These pieces of evidence taken together are indicative of a modeled system curve that is too steep, and there are clearly additional macro-calibration activities necessary to determine the cause, whether that be incorrect data on pipe diameters, incorrect data on valve statuses, or another cause to be determined.

4.3 Real-Time Simulation of Tracer Movement

This section describes the real-time simulation of the calcium chloride tracer using the collected conductivity monitoring results and the EPANET-RTX real-time model. The processing of the conductivity data is described in Section 4.3.1, and the development of the real-time water quality model using the conductivity data is described in Section 4.3.2. The accuracy metrics for evaluating the real-time water quality model are described in Section 4.3.3. The results of the accuracy evaluation are summarized in Section 4.3.4.

4.3.1 Tracer Data Processing

The specific conductance monitors typically produced data with a noise level that was well below the signal difference introduced with the tracer pulse injections. Each raw data stream was visually inspected for obvious anomalies, including sudden and persistent changes in conductivity to levels that were noticeably below background or above the maximum conductivity injection peak, or sudden and persistent increases in noise levels that indicated a sensor instability. Based on such visual inspection, specific data ranges were removed from the following monitor locations: A7 (data before 11/19/12 00:00); B5 (data before 11/19/12 06:00); D2 (data before 11/19/12 13:20 and after 11/20/12 07:30); D4 (data before 11/19/12 11:50 and after 11/19/12 22:50); and F4 (data after 11/20/12 18:00).

Beyond the visual inspection, the tracer data was processed prior to analysis by ranging, interpolating, and smoothing. Specifically, each data stream was processed first by passing through a ranging filter that excluded points below 200 and above 1,000 $\mu\text{S}/\text{cm}$. These data were then resampled on a common clock once per minute, using linear interpolation between adjacent data points. Finally, the ranged and resampled/interpolated data was automatically passed through a moving average filter with a 14-minute averaging window (7 minutes before and after each resampled point). This processing did not affect the underlying signal at time scales of interest. The processing did simplify the subsequent data analysis procedures by removing small anomalies and, in the process, simplifying the visual comparison between observed and simulated time series. An example of the results from these data processing steps is shown in Figure 4.3-1 for a location with a higher than normal level of sensor noise.

A final data processing step adjusted the time stamps of all tracer data streams to account for residence time in the hydrant barrel. The residence time varied from one hydrant to the next, due to variations in depth of main and sample flow rate (which was set to approximately 1 GPM at each location). The hydrant nearest the injection site was used to measure the approximate delay due to the hydrant barrel residence time; at that location, the conductivity pulse arrived approximately 14 minutes after the start of the injection. Each of the tracer data streams was adjusted backward by 14 minutes to better approximate the conductivity values in the main, and compare to tracer simulation values that would not include the hydrant barrel residence times.

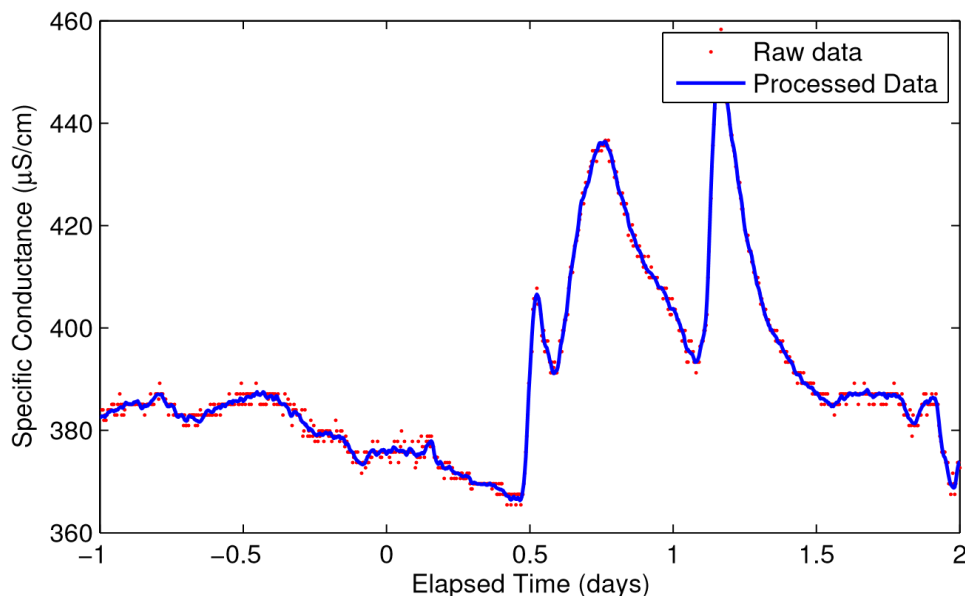


Figure 4.3-1. Illustrative results from data processing applied to raw tracer data.

4.3.2 Real-Time Water Quality Model

The tracer simulations presented in Sections 4.3-4 and 4.3-5 were developed using CitiLogics (Covington, KY) Polaris™ software, a real-time data analytics environment based on the EPANET-RTX real-time extension for the EPANET programmer's toolkit (USEPA, 2013). Due to practical constraints, simulation results were created after the tracer event, as the real-time Polaris software was not installed at the water utility, and the conductivity monitors were not configured to collect data in a real-time database. Subsequent to the tracer test event, the water utility SCADA historian database was copied and brought off-site for connection to EPANET-RTX objects, through the Polaris interface. The tracer data was stored in a MySQL database using EPANET-RTX time series objects. (MySQL is the open source, community edition of structured query language (SQL) database.) Polaris was also used to construct the EPANET-RTX time series data transformation pipelines that performed the conductivity data processing steps described in Section 3.2.

The data access, transformation, and simulations were conducted as they would have been in a real-time computation; the results are identical to those which would have been computed in real-time, automatically and without the intervention of an analyst, had that been a possibility during the conduct of the tracer test. The hydraulic behavior is defined by status, setting, and head boundary conditions, and district metered area demand computations, as defined for the real-time hydraulic model and described in Sections 3.2 and 3.3. No additional adjustments to assumptions or model parameters were made. Beyond those real-time hydraulic calculations, the real-time tracer simulation only requires initial and boundary conditions, as the tracer is non-reactive. The only relevant simulation parameter is the EPANET water quality time step, which was 15 seconds.

The boundary condition at the site of injection (see Figure 2.1-2) was constructed from continuous conductivity monitor observations, snapshot observations downstream of the injection location, and records of the start and stop time of the injection pump. The resulting tracer concentration is shown in Figure 4.3-2. Beyond the characteristics of the four salt pulses that were injected, the background conductivity at the injection site averaged 365.5 $\mu\text{S}/\text{cm}$, and was stable with a standard deviation of 4.2 $\mu\text{S}/\text{cm}$. The positive displacement injection pump discharged into a force main on the suction side of three parallel high service pumps. Coordination with operations staff ensured that the high service pumping status did not change while the tracer was being injected. Thus, once the injection pump was turned on and the controller was set, the resulting conductivity was relatively stable. The changes in conductivity at the start of the first pulse in Figure 4.3-2 were due to experimentation with the injection rate. The injection was stopped for 5 minutes after 14 minutes had elapsed, because the first conductivity pulse had not been seen at a nearby downstream hydrant, creating uncertainty about the initial pulse characteristics. This uncertainty was resolved, and the pulse arrival time delay (estimated to be between 14 and 19 minutes) was attributed to residence time in the hydrant barrel.

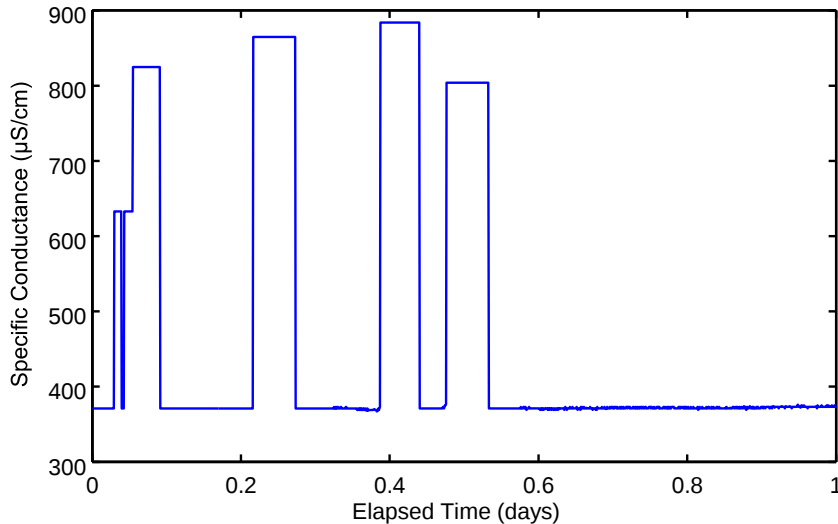


Figure 4.3-2. Conductivity boundary condition at the injection site used for tracer simulations.

One additional boundary condition was required at the northern TP (see Figure 2.1-2). Under the operating conditions during the tracer test period, this TP operated part of the day, and all plant production flowed by gravity from the clearwell to lower pressure zones located to the west and southwest. These latter pressure zones were excluded from the tracer test monitoring area, because only a small fraction of their demand was satisfied via a regulator from the southern TP that carried the tracer injection. Thus, in principle, the conductivity at the northern TP was not important for simulating tracer movement within the study area, because it only fed the lower pressure zones which were excluded from monitoring. As a result, a conductivity monitor was not placed at the northern TP boundary, and the initial plan was to arbitrarily set that tracer boundary condition to zero.

Under some operating conditions, water is delivered from the northern TP through high service pumps to the upper zone that interacts with the study area. Under the operating conditions in effect during the test, these high service pumps were off, and all demand in the study area was satisfied by the southern TP high service pumping. The north TP high service pumps were used, however, up until the start of the tracer injection. Thus water from the northern TP is expected to be within the study area, at the start of the tracer injection. There were two methods of dealing with this issue. The first would be to ignore it and start the simulation with the tracer addition, and represent as accurately as possible the initial conditions at the start of the test. The other would be to model the conductivity at the northern TP, and back up the simulation start to include a time period prior to the tracer addition, during which the background conductivity from both plants would be represented as tracer boundary conditions. This latter approach was chosen because of evidence that the background conductivity at the two plants were significantly different.

Given the lack of a monitor at the northern TP, two nearby monitors were used to develop an assumption of the north TP's background conductivity. Figure 4.3-3 (top) shows a nearby storage tank level along with the status of the north TP high service pumps (if the status is 1, one pump is running; if 0, all pumps are off). The time scale is in days relative to the start of the tracer injection. More than 3 days before the start, the north TP was running continuously and the high service pumps delivered water to a significant portion of the study area. During this operation mode, the record shows that the nearby storage tank filled only when the north TP high service pumps were on. The time period from days -4 to -3 in Figure 4.3-3 (top) is indicative of this behavior. Through day -2, this tank was still filled only when the

north TP high service pumps were on. A reasonable assumption, due to spatial proximity and pumping operations, was that the conductivity within this nearby storage tank represented an integrated measure of the background conductivity at the northern TP. Figure 4.3-3 (middle) shows the specific conductance measured at this tank along with its level, for a time period prior to the start of the test. The conductivity signal prior to day -2 occurred at the end of a prolonged drain cycle, and averaged 445 $\mu\text{S}/\text{cm}$; this was the best estimate given the data of the conductivity within the tank when it is being filled from high service pumping at the north TP.

The conductivity monitor nearest to the northern TP was also used to infer the background conductivity (see Figure 2.1-2; the monitor is to the southeast of the TP). Data from this monitor was lost soon after the salt injection, but valid data was harvested for the preceding 2 days, as shown in Figure 4.3-3 (bottom). During the 2 days preceding the test, the area that included this monitor was fed primarily by the southern TP. There were clear signatures, however, of higher conductivity pulses entering from the northern TP, associated with high service pumping activity. Thus these data provided additional evidence that the background conductivity at the north TP was significantly higher than the 365 $\mu\text{S}/\text{cm}$ at the south boundary. The conductivity pulses at the nearby monitor were more likely, however, to be influenced by mixing with the south TP water. For that reason, the north TP boundary condition was set at 445 $\mu\text{S}/\text{cm}$ — the value measured in the nearby storage tank prior to the change in operations.

The tracer simulation was divided into two parts: an initial period 53 hours prior to the test start at 11/19/2012 08:00; and a simulation period of 9 days. The start of the initial period was determined based on data availability. Most monitors were set to turn on at 11/16/2012 23:00, or 57 hours prior to the injection start. Delaying the start of the initial period by 4 hours allowed more monitors to be used to specify initial conditions, because of various data problems. Initial conditions were specified using data from the conductivity monitors distributed throughout the study area, at the start of the initial period. A nearest neighbor spatial interpolation was used to distribute those data to each network node. Initial conditions for storage tanks within the study area were treated separately. For each tank, its observed conductivity was plotted along with its level, and the initial conductivity was estimated from that observed during a suitable drain period. At the start of the simulation period, all simulated tank levels were reset to observed levels. Thus at the start of the simulation period, the water quality initial conditions reflect the propagated background conductivity from both TPs over a 53-hour period, and the hydraulic initial conditions reflect measurements from SCADA.

Water Utility Case Study Using EPANET-RTX Libraries

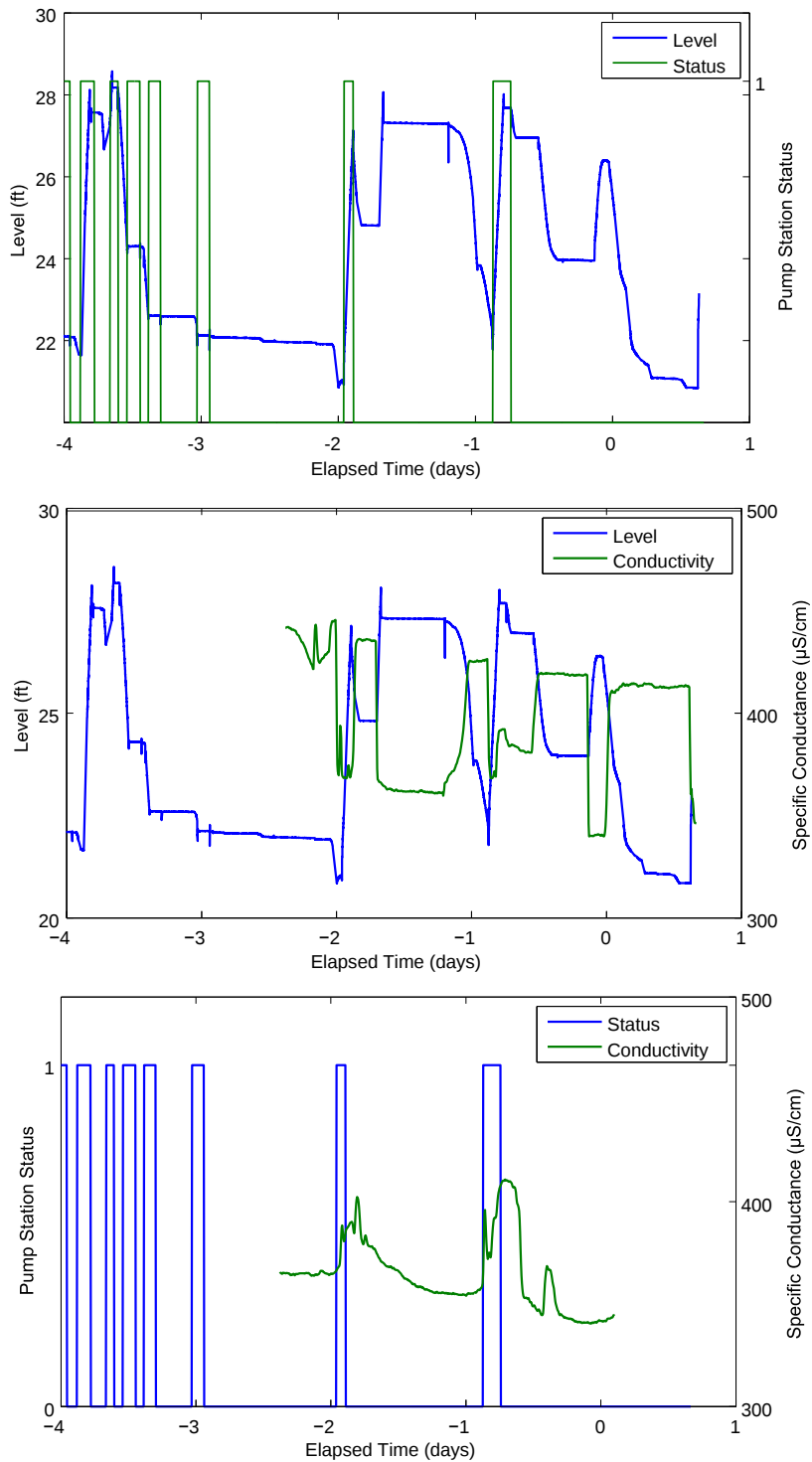


Figure 4.3-3. Evidence used to determine elevated background conductivity at northern treatment plant. Water level at nearby storage tank and north treatment plant high service pump status (top); same water level and conductivity on inlet/outlet line (middle); and high service pump status and nearby conductivity measure (bottom).

4.3.3 Accuracy Metrics

It is difficult to assess the quantitative similarity between two time series. This is especially true of the simulated versus observed tracer time series from the current field study — the four tracer pulses were designed to produce unusual fluctuations in tracer signals, and time-shift errors between simulated and observed series could produce large disparities in traditional “goodness of fit” metrics, such as the root-mean-squared error (RMSE). The RMSE is not expected to discriminate between locations where the signals are time shifted, and those where the simulated signal bears little qualitative resemblance to the measurements. One alternative approach could use the Pearson’s correlation coefficients at different time lags to measure the goodness of fit, as well as the lag that produces the largest correlation between simulated and observed. Another approach could use dynamic time warping (DTW) to compare the paired series, which allows for variable and non-linear time shifts. While both of these approaches are expected to improve upon a simple RMSE (or similar metrics), we opted for an approach that compares the quantiles of the conductivity area above background, or CAAB. This approach relies on traditional statistical concepts, and naturally accommodates variable time-shifting behavior, while being simpler than either the lag-correlation or DTW methods.

Figure 4.3-4 illustrates the quantified pulse characteristics for one particular tracer time series. Time is shown as elapsed time relative to the start of the tracer test at 08:00 on November 19, 2012. The first analysis step was to subtract the conductivity background individually for each series, to expose the conductivity pulse signals as the key features to be measured. The background exhibited some degree of variability, so background subtraction relied on an operational definition. We estimated the background for each series as the average conductivity during the 24 hours immediately preceding the tracer injection (either measured or simulated), and subtracted this from the entire time series. The time series in Figure 4.3-4 shows the conductivity signal after the background is subtracted. In cases where the data record did not extend the full 24 hours prior to the injection, the background was estimated from available data. If no measured data existed prior to the start of tracer injection, the background of the observed signal was assumed equal to the simulated series background.

The key characteristics of the tracer pulses were estimated from integrating the area under the conductivity signal (numerically), above the background, as a function of the elapsed time. This area is plotted in Figure 4.3-4 on the right axis. Note the integral was computed between the limits of 0 and 2 days’ elapsed time — an operational assumption reflecting that the tracer pulse signal had passed all of the monitor locations within the first 2 days after injection. Given the CAAB versus elapsed time curve, the first, second, and third quantiles of the area (Q_1 , Q_2 , and Q_3 , each with units of time) are identified and used as key characteristics of the simulated and observed pulse signals at each location. More specifically, we compared the simulated and observed median, Q_2 , or the time for 50% of the pulse area to pass the monitor, and the simulated and observed IQR , as a measure of the time spread of the pulse. Since the simulated and observed medians and IQR could match exactly, even if the simulated pulse was attenuated or amplified relative to the observed, the total CAAB at 2 days elapsed time, labeled C_{TOT} in Figure 4.3-4, was also used to compare simulated and observed time series.

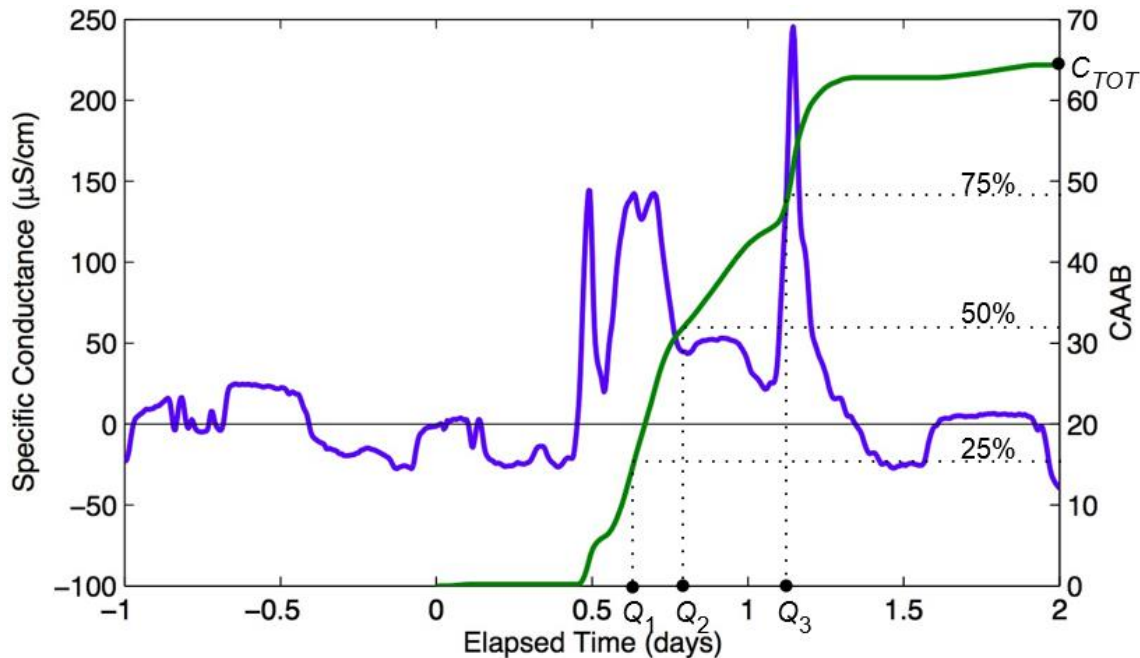


Figure 4.3-4. Tracer time series characteristics used for error analysis. Quartiles Q_1 , Q_2 , and Q_3 mark the times when 25, 50, and 75% of the tracer pulse signature (not necessarily tracer mass) has passed the sensor, allowing use of the median (Q_2) and *IQR* as comparative characteristics. The CAAB at the 2-day mark provides an integrated measure of signal strength and temporal signature.

4.3.4 Observed and Simulated Tracer Signals

Time series data are presented here for the observed and simulated conductivity signals from 37 monitoring sites. The sites are grouped and plotted as regions A through F. The selection of these regions was based upon physical proximity for implementing the field study, and not necessarily because they span a pressure zone or demand metered area, or any other reason related to infrastructure or hydraulic behavior. Each location plot shows the simulated and observed conductivity signal above background, and includes information about the simulation accuracy metrics discussed in Section 3.3. Inset plots are included that show the CAAB versus time for both simulated and observed time series, allowing the conductivity pulse areas to be compared visually as they evolve over time. The CAAB quantiles are shown symbolically on each graph adjacent to the time axis, in the manner of a box plot. The outer box represents the first and third quartiles, the difference between them the interquartile range, and the line within the box the second quartile, or median.

Region A

Figures 4.3-5 through 4.3-7 show the observed and simulated conductivity signals over a 3-day period. This region was densely monitored, providing an unusual spatial-temporal picture of tracer evolution in an older, more densely populated, urbanized area. No data is included for Location A1 because the conductivity sensor malfunctioned. With the exception of A5, which was just outside of the “gridded” portion of Region A, the conductivity signals for Locations A2 through A8 showed similar conductivity signals. These results are surprising given the gridded pipe connectivity within this region. Although the observed signals for A2, A3, and A4 are very similar and these locations are within several blocks of each other, the simulated signal for A3 shows distinct characteristics that are not reflected in the data.

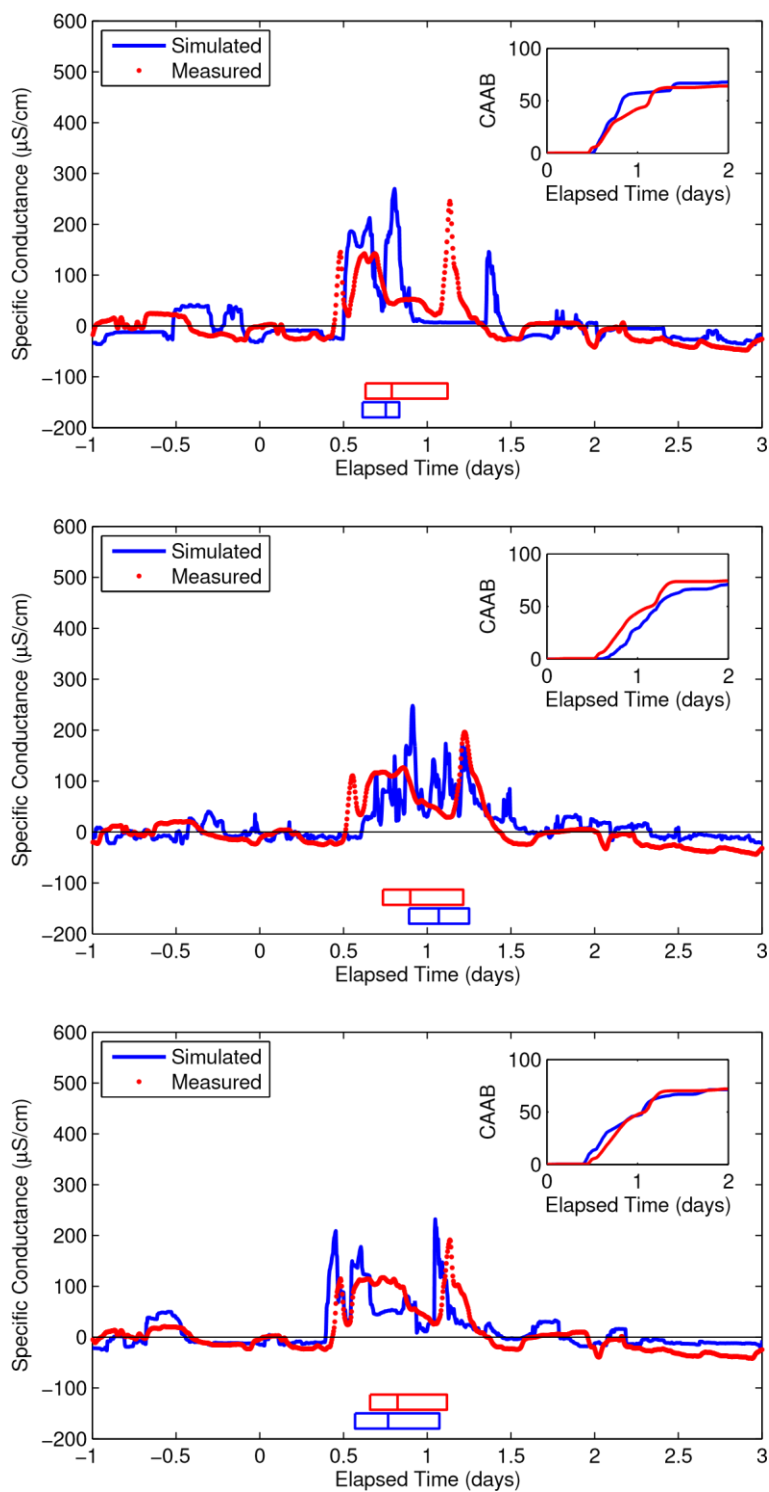


Figure 4-5. Observed and simulated tracer movement, Locations A2 (top), A3 (middle), and A4 (bottom). Inset figure shows the CAAB over a 2-day period commencing with the start of tracer injection, for both simulated and observed time series. The symbols along the x-axis show the IQRs (outer box boundaries) and medians (lines within boxes) for each time series.

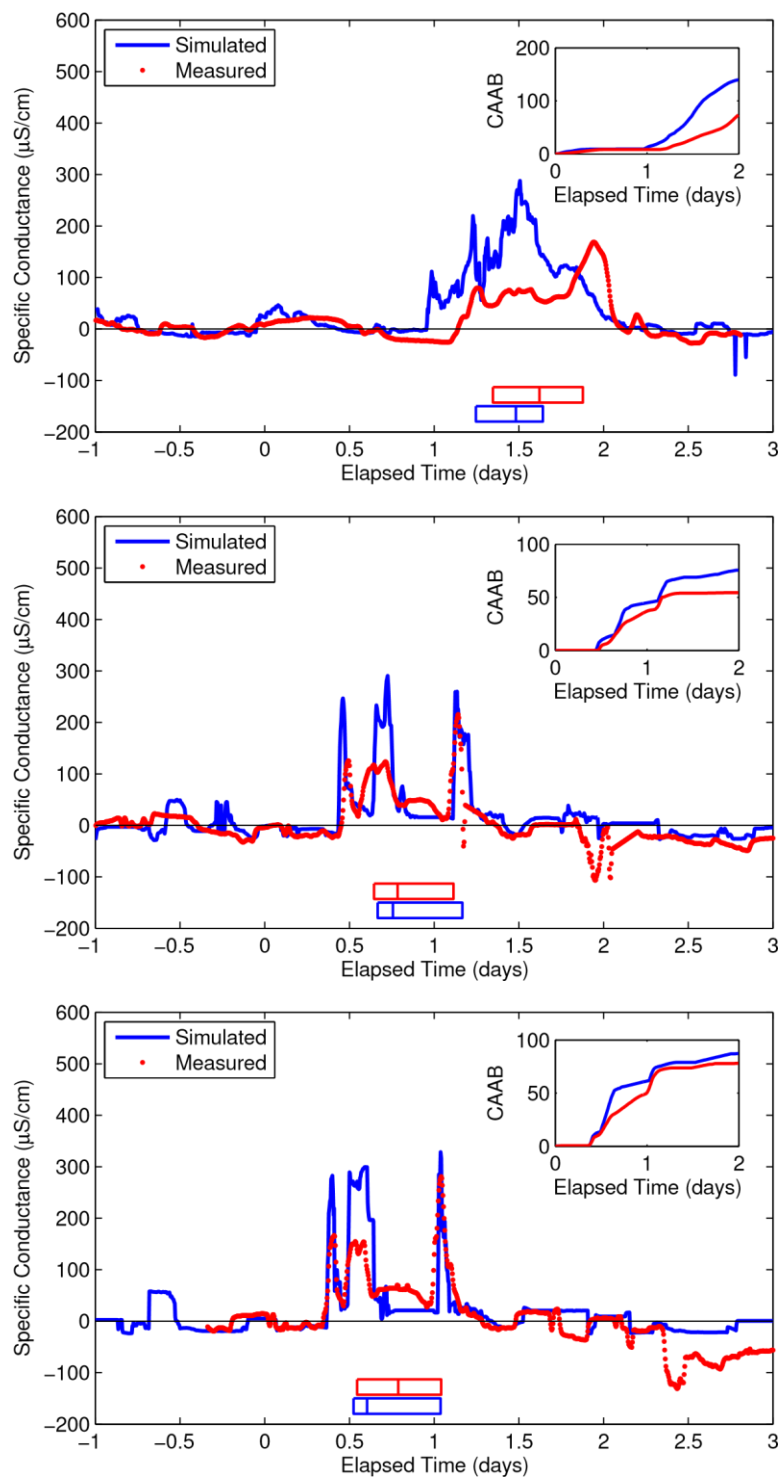


Figure 4.3-6. Observed and simulated tracer movement, Locations A5 (top), A6 (middle), and A7 (bottom). Inset figure shows the CAAB over a 2-day period commencing with the start of tracer injection, for both simulated and observed time series. The symbols along the x-axis show the *IQRs* (outer box boundaries) and medians (lines within boxes) for each time series.

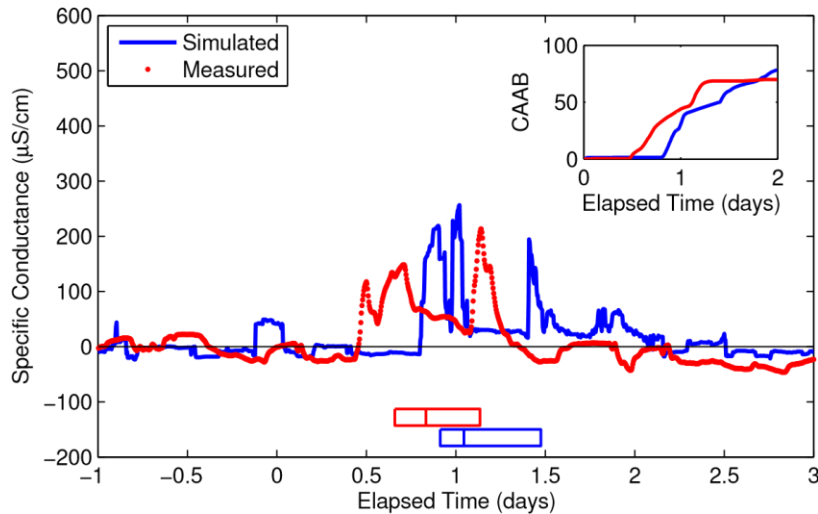


Figure 4.3-7. Observed and simulated tracer movement for Location A8. Inset figure shows the CAAB over a 2-day period commencing with the start of tracer injection, for both simulated and observed time series. The symbols along the x-axis show the *IQRs* (outer box boundaries) and medians (lines within boxes) for each time series.

Region B

Figures 4.3-8 and 4.3-9 show the simulated and observed conductivity results for the six monitoring stations in Region B, which were also located in a densely populated region of the distribution system. No data was collected from Locations B4, B6, B8, and B10 due to monitor malfunction. While Locations B1 through B5 show similarity in the observed data, the degree of similarity is not as great as the monitors located within Region A. These observations also showed significant pulse attenuation compared to Region A observations, and also compared to the simulated time series. Location B7 was located at a storage tank; the square pulses are generated by the tank drain and fill cycles. The signal from Location B9 was similar to locations in Region A. The signal at B7 shows how visually similar observed and simulated time series can exhibit large errors in the CAAB median, suggesting that perhaps quantile ranges are a better, and more stable, metric of simulation accuracy.

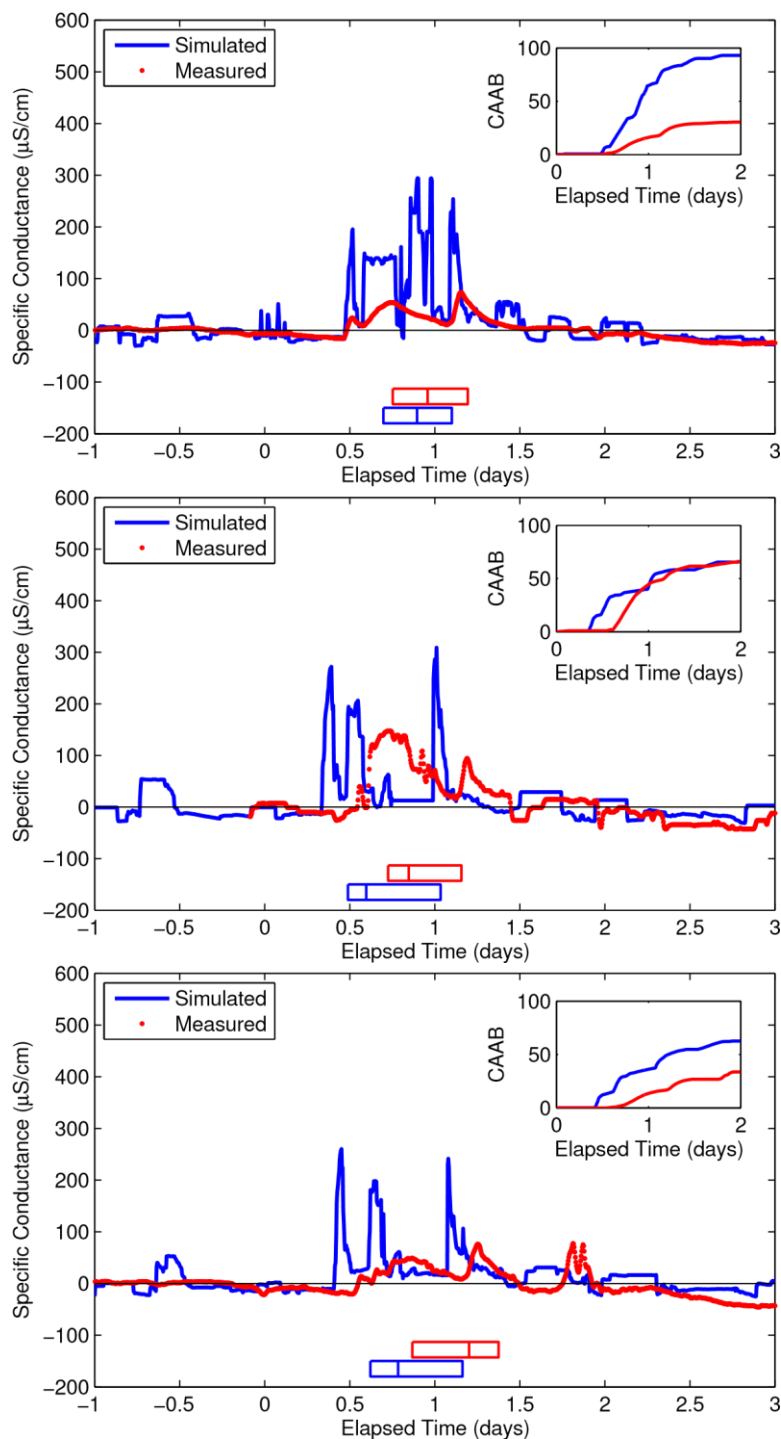


Figure 4.3-8. Observed and simulated tracer movement, Locations B1 (top), B2 (middle), and B3 (bottom). Inset figure shows the CAAB over a 2-day period commencing with the start of tracer injection, for both simulated and observed time series. The symbols along the x-axis show the *IQRs* (outer box boundaries) and medians (lines within boxes) for each time series.

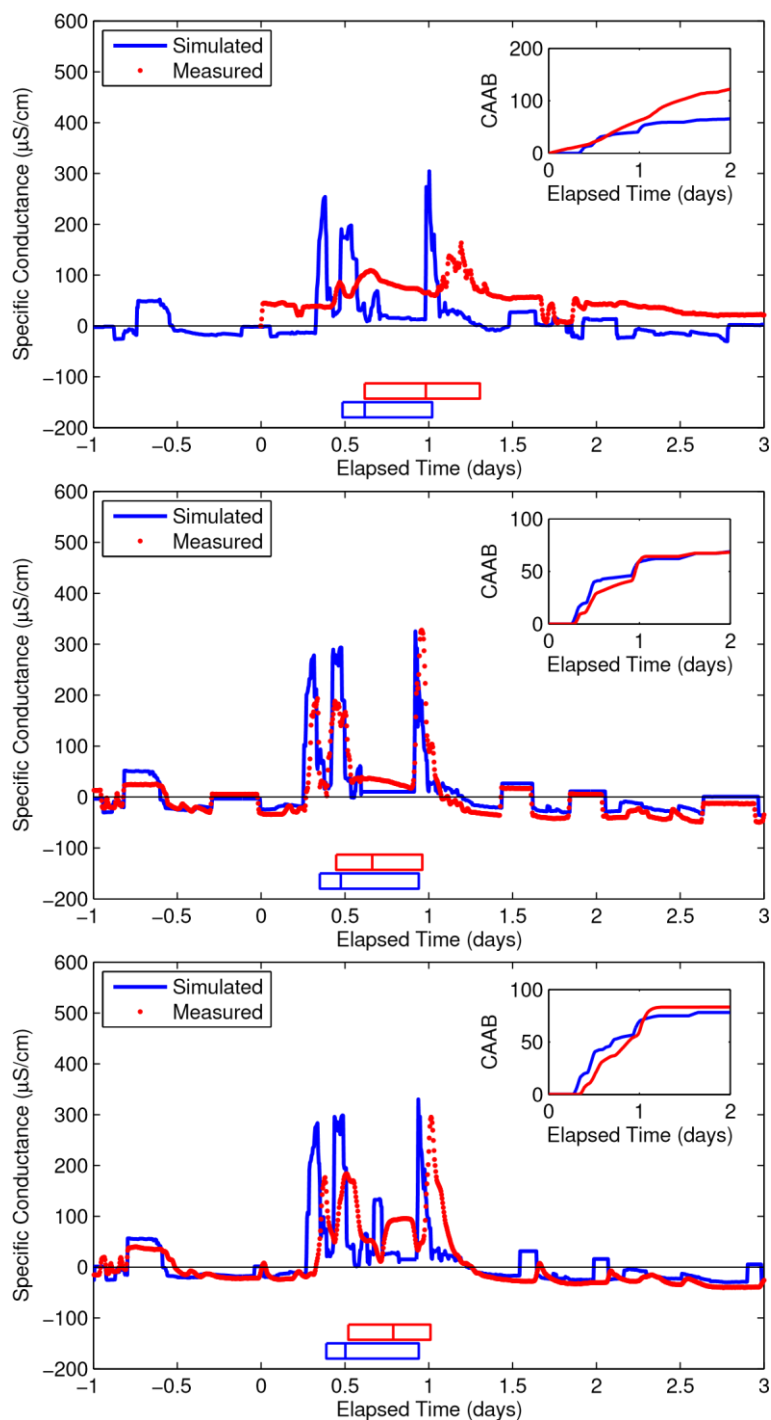


Figure 4.3-9. Observed and simulated tracer movement, Locations B5 (top), B7 (middle), and B9 (bottom). Inset figure shows the CAAB over a 2-day period commencing with the start of tracer injection, for both simulated and observed time series. The symbols along the x-axis show the *IQRs* (outer box boundaries) and medians (lines within boxes) for each time series.

Region C

Figures 4.3-10 through 4.3-12 present the simulated and observed conductivity signals for the seven monitoring locations in Region C, adjacent to and south of Regions A and B. No data for Location C5 is presented due to monitor malfunction. Location C4 represents the conductivity signal adjacent to a storage tank. Locations C2 and C3 were under the influence of this tank as illustrated by the similarity in signal characteristics. Although the simulated conductivity signal at C4 is reasonable (e.g., *IQR* = 2.33, compared to the median *IQR* of 1.46 hrs.), the median CAAB at nearby signals at C2 and C3 are significantly too early and too late, respectively. Both of these sites predicted conductivity signal peaks resulting from the injection of the tracer at the source location that are not supported by the observed data. Interesting, while Location C3 was at the end of a long dead-end main, C2 was within what could be called a dead-end loop, due to the presence of a downstream regulator that is likely to be closed, according to utility personnel. Locations C7 and C8 presented an interesting study on the impact of demands, and possibly transport mechanisms, within dead-end mains. Location C8 was on a 12-in. main and C7 was just downstream on a 6-in. dead end. While C8 provided results that had visually and quantitatively good error characteristics, C7 provided results that were visually a much poorer fit to the simulated signal, which has qualitatively different characteristics even though it is just a short distance downstream. Presumably, these different characteristics were due to dead-end demands as well as dispersion processes that may be dominant within the dead end. Finally, the comparison of C7 and C8 points out the challenges inherent in comparing two time series; while these locations had roughly equal quantitative error characteristics, the visual fit of the C8 observed results to the simulated signal data is noticeably superior to that of Location C7. What is happening in this case is the *IQR* is relatively good yet the times associated with passage of the first and third quantiles are significantly in error; it may be preferable to examine a metric focused more on these particular times, in contrast to using the *IQR*.

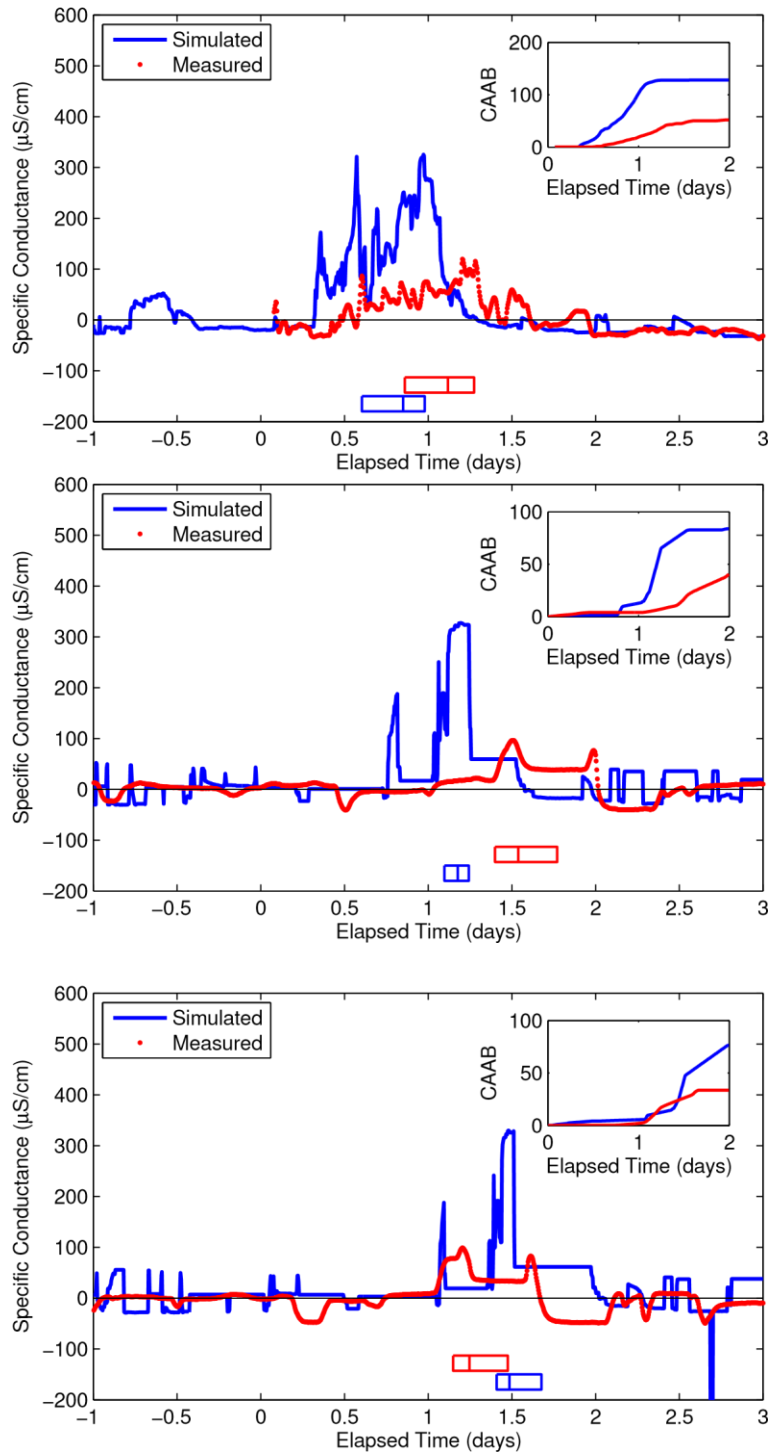


Figure 4.3-10. Observed and simulated tracer movement, Locations C1 (top), C2 (middle), and C3 (bottom). Inset figure shows the CAAB over a 2-day period commencing with the start of tracer injection, for both simulated and observed time series. The symbols along the x-axis show the IQRs (outer box boundaries) and medians (lines within boxes) for each time series.

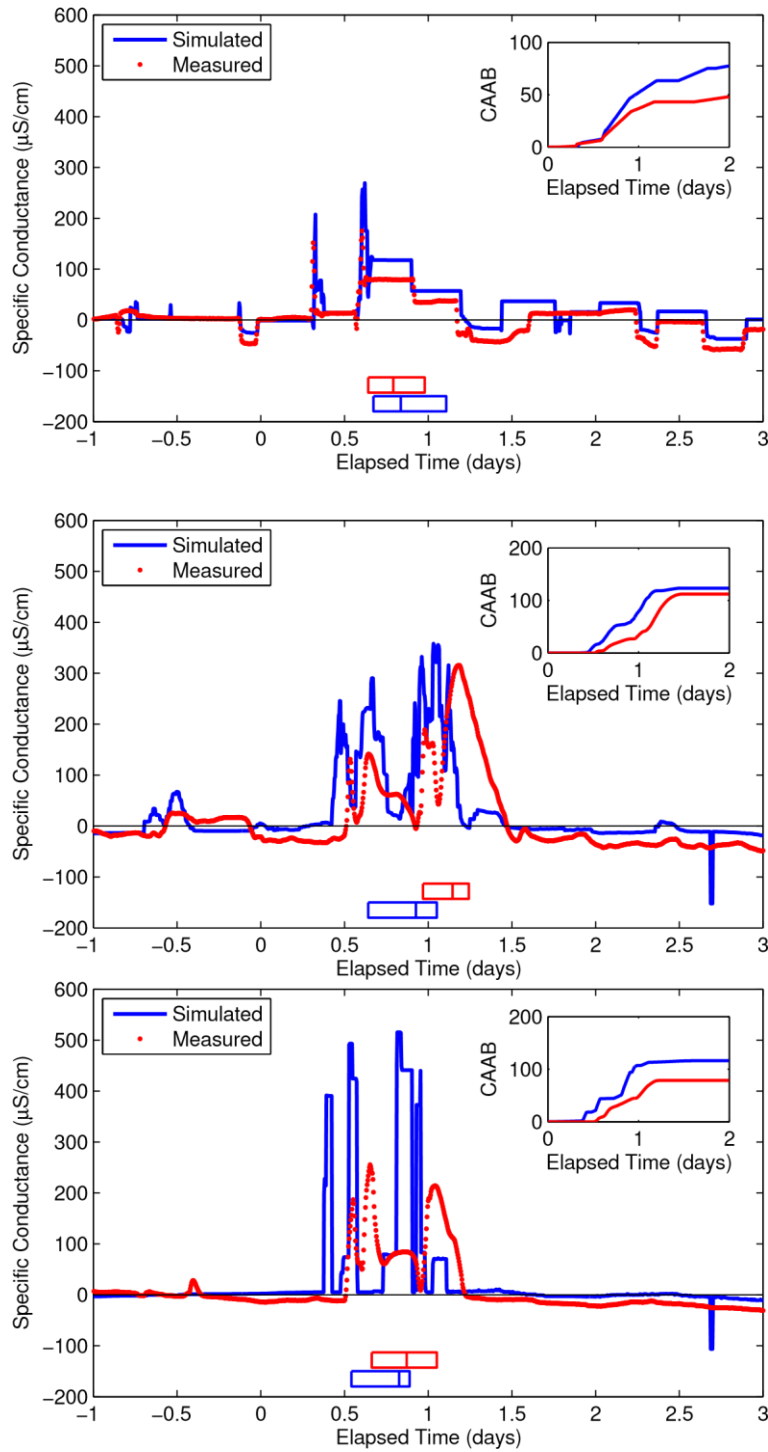


Figure 4.3-11. Observed and simulated tracer movement, Locations C4 (top), C6 (middle), and C7 (bottom). Inset figure shows the CAAB over a 2-day period commencing with the start of tracer injection, for both simulated and observed time series. The symbols along the x-axis show the IQRs (outer box boundaries) and medians (lines within boxes) for each time series.

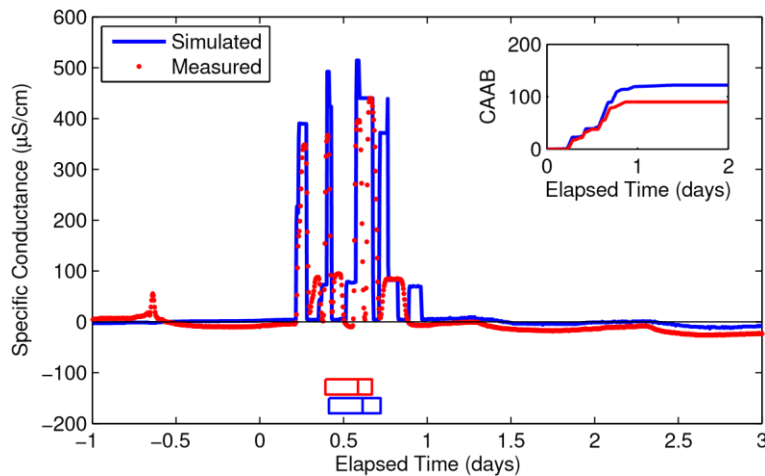


Figure 4.3-12. Observed and simulated tracer movement, Location C8. Inset figure shows the CAAB over a 2-day period commencing with the start of tracer injection, for both simulated and observed time series. The symbols along the x-axis show the *IQRs* (outer box boundaries) and medians (lines within boxes) for each time series.

Region D

Figures 4.3-13 through 4.3-15 show the conductivity signals for seven monitoring stations within Region D. Region D is south of Region A and includes the injection location, D8, at the southern TP in Figure 2.1-2. The monitor at Location D8 started recording data after the first salt pulse, because it was moved to replace a malfunctioning monitor. The close proximity of D8 to the injection boundary resulted in expected small simulation errors. Monitor D1 was located on the boundary of two pressure zones, with flow governed by a regulating valve. The significant delay in the simulated median CAAB time compared to observed is a good indication that flow through the regulating valve is significantly greater than simulated, and thus that the downstream valve setting in the network model was in error. Locations D2, D3, and D4 were in close proximity to each other, with D3 located at a storage tank. Flow to this tank was actively controlled via a solenoid operated valve on the inlet/outlet line; the status of this valve was modeled explicitly by the real-time hydraulic model. When filling or draining, this tank level changed by 5 to 10 feet over short intervals of approximately 2 hours. Errors between simulated and observed conductivity at this tank were due to errors in pulse arrival times relative to when the tank was filling. Observations indicate that each conductivity pulse was transported past the tank while it was draining, whereas the simulation suggested that pulse arrival coincided with a tank fill period about 15 hours after the first injection. Location D4 was on a 16-in. main leading from the injection site to the tank, and was a good indication of the simulation error along one of the largest mains in the study area. Location D2, however, was on an 8-in. distribution main off of the 16-in. Locations D6 and D7 were both on 6-in. dead-end mains; D6 was on a pipe that branches off of a 12-in. main, and D7 was on a pipe that branches off of a 16-in. main — both leading from the injection site. The errors at both these locations indicated the simulation was too slow by about 5 hours, but those delays could be due to greater velocities in transmission mains, or within the dead-end pipes than predicted by the model. It is interesting that the observed conductivity signal at D7 did not exhibit the pulse attenuation and dispersion observed at D6 and at other dead-end locations.

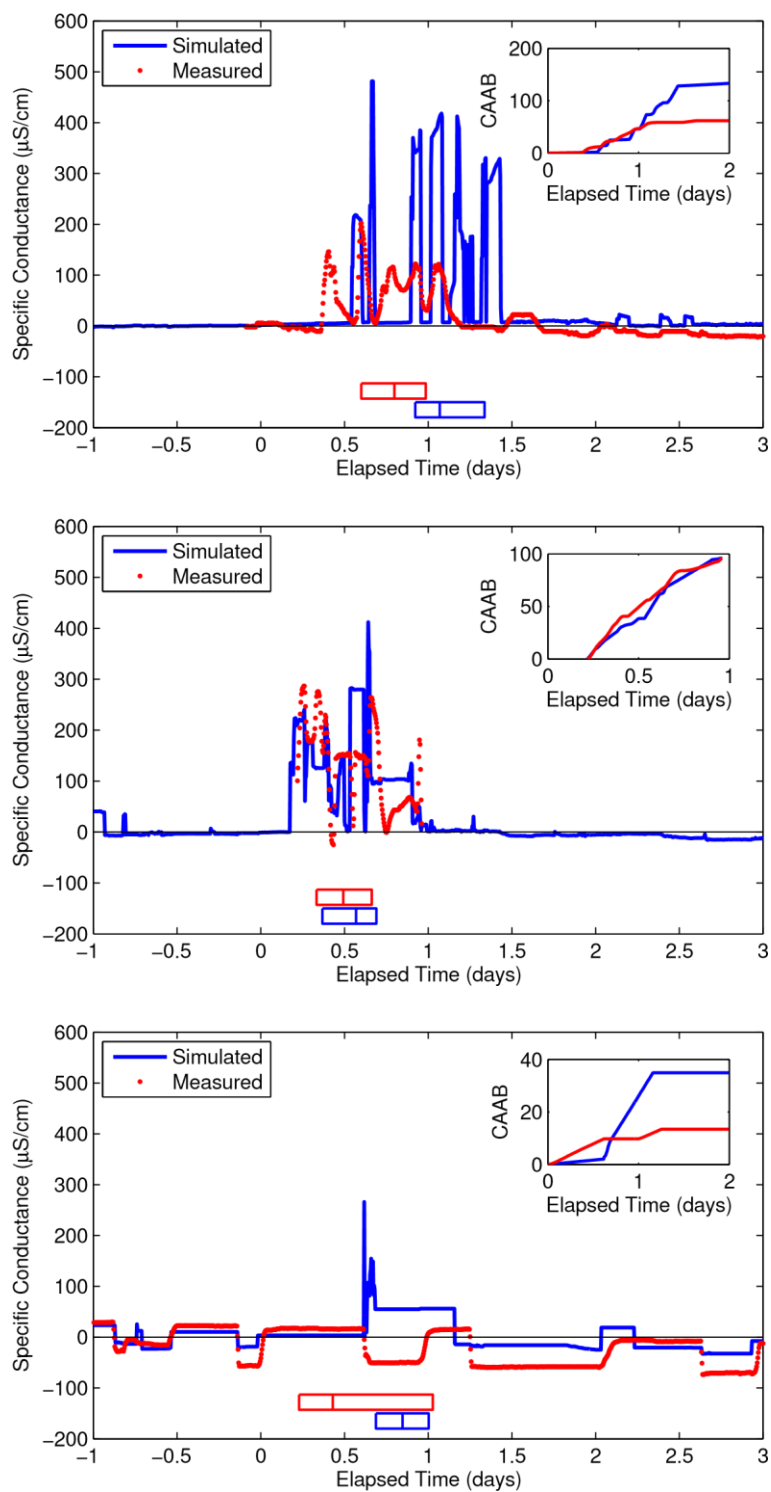


Figure 4.3-13. Observed and simulated tracer movement, Locations D1 (top), D2 (middle), and D3 (bottom). Inset figure shows the CAAB over a 2-day period commencing with the start of tracer injection, for both simulated and observed time series. The symbols along the x-axis show the IQRs (outer box boundaries) and medians (lines within boxes) for each time series.

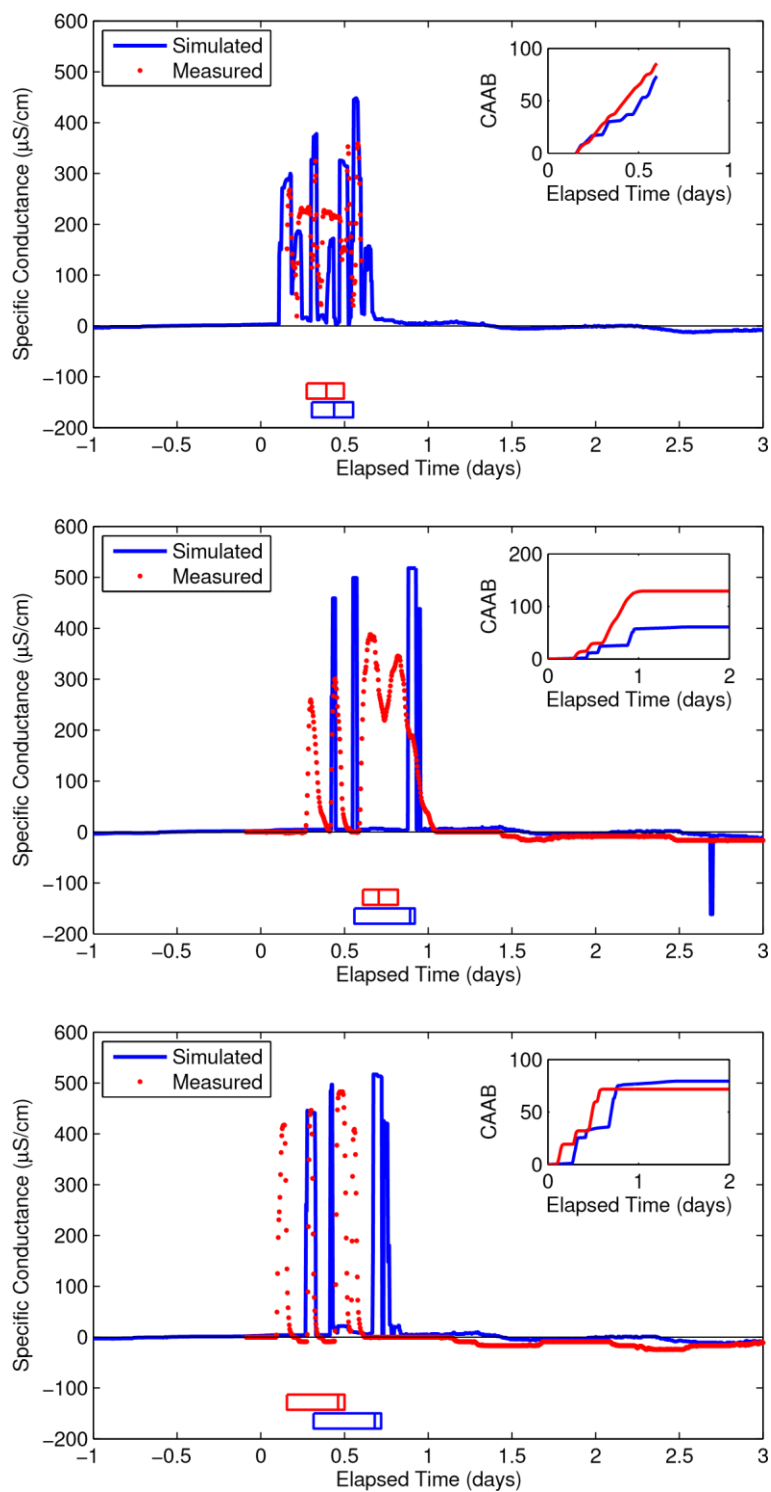


Figure 4.3-14. Observed and simulated tracer movement, Locations D4 (top), D6 (middle), and D7 (bottom). Inset figure shows the CAAB over a 2-day period commencing with the start of tracer injection, for both simulated and observed time series. The symbols along the x-axis show the IQRs (outer box boundaries) and medians (lines within boxes) for each time series.

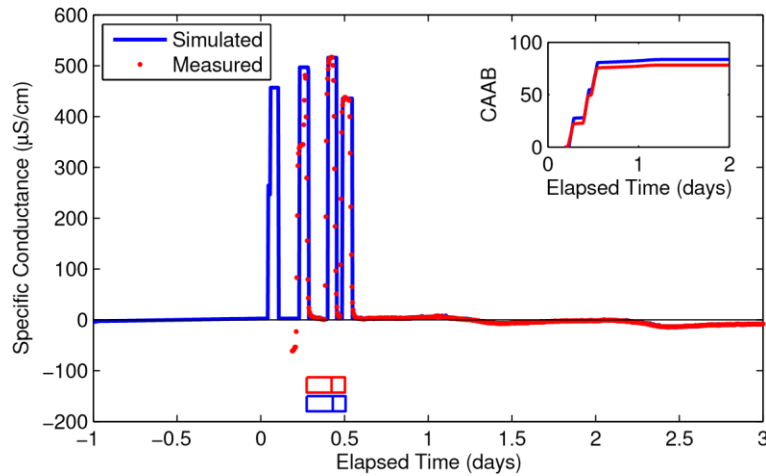


Figure 4.3-15. Observed and simulated tracer movement, Location D8. Inset figure shows the CAAB over a 2-day period commencing with the start of tracer injection, for both simulated and observed time series. The symbols along the x-axis show the *IQRs* (outer box boundaries) and medians (lines within boxes) for each time series.

Region E

Figures 4.3-16 through 4.3-17 show the simulated and observed conductivity signal results from the five monitors located in Region E. Location E5 was omitted due to monitor malfunction. Monitors E2 and E6 were located at storage tanks. Location E1 was on a pressure zone boundary and under the influence of both the tank at E2 and a downstream regulating valve. Locations E3 and E4 were both on dead-end mains in between the two storage tanks at E2 and E6. The simulation errors at E3 and E4 suggest again how simulation results may be strongly affected by highly localized demand characteristics within small diameter pipes and especially within dead-end segments.

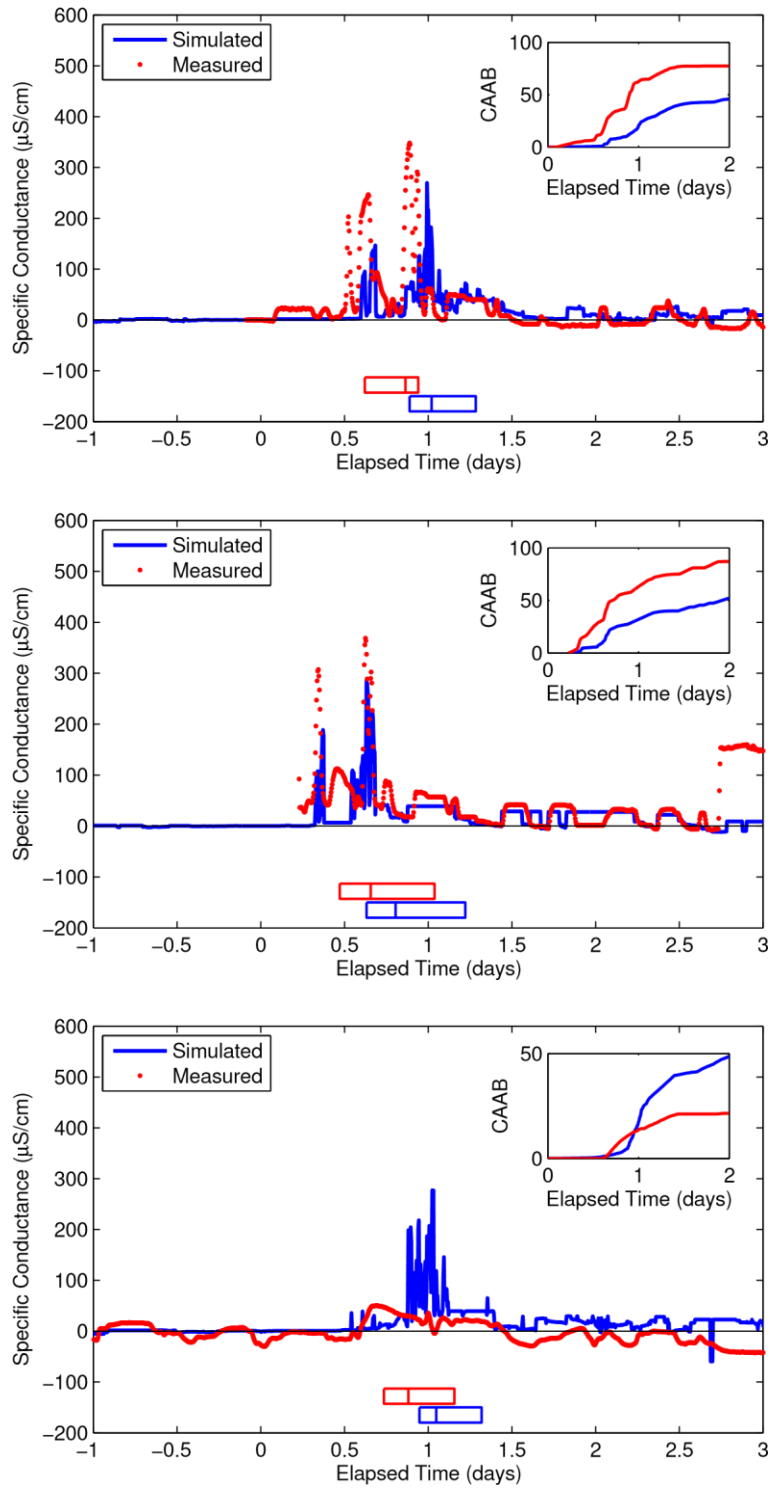


Figure 4.3-16. Observed and simulated tracer movement, Locations E1 (top), E2 (middle), and E3 (bottom). Inset figure shows the CAAB over a 2-day period commencing with the start of tracer injection, for both simulated and observed time series. The symbols along the x-axis show the IQRs (outer box boundaries) and medians (lines within boxes) for each time series.

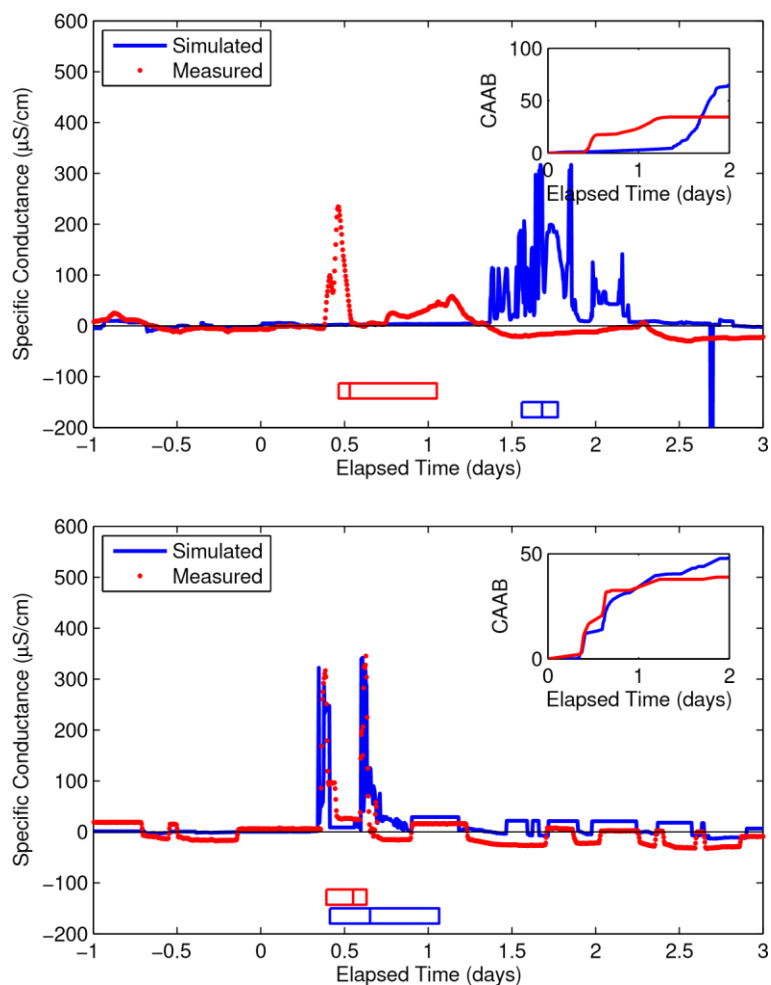


Figure 4.3-17. Observed and simulated tracer movement, Locations E4 (top) and E6 (bottom). Inset figure shows the CAAB over a 2-day period commencing with the start of tracer injection, for both simulated and observed time series. The symbols along the x-axis show the *IQRs* (outer box boundaries) and medians (lines within boxes) for each time series.

Region F

Figures 4.3-18 through 4.3-19 show the observed and simulated conductivity signal results for the six monitors in Region F. Locations F1, F2, and F3 were on distribution mains of 8, 6, and 12 in., respectively, that branch off of a 20-in. transmission main leading from the injection site to the south. None of these locations were on what could be called dead-end mains; they each had significant demand downstream, even if the network structure is mostly branched. Locations F1 and F2 observed three or four distinct pulses, and the simulations were slow compared to the observations. Location F3, however, observed a single pulse, even though it was downstream of F1 and F2 and off the same main. Obviously the flow was being affected downstream of F1 and F2, and in a manner that was contrary to the simulation. A booster station is located nearby and to the south of F2, off of the same 20-in. transmission main, that serves the area to the south (this area was outside of the study region). It is interesting that the real-time hydraulic simulation to the south of F3 was noticeably less accurate than other regions, including the operation of the booster station and the cycling behavior of the three tanks that it serves (see Section 4.2). Briefly, these three tanks are used more in terms of the depth of the fill/drain cycles, as compared to how they are used in the real-time hydraulic simulation. Thus, one explanation for the tracer simulation errors at F3 was that the increased draw on the south tanks when the booster pump station was off prevented some of the tracer pulse mass from being transported to the south, between Locations F1, F2, and F3. These same hydraulic errors likely affected the other locations in Region F, especially F4, although the errors in simulated signals at F5 and F7 were reasonable. The decreased amplitude of the observed pulses at Location F7 may have been due, in part, to the impacts on flow from booster pumping and demands to the south.

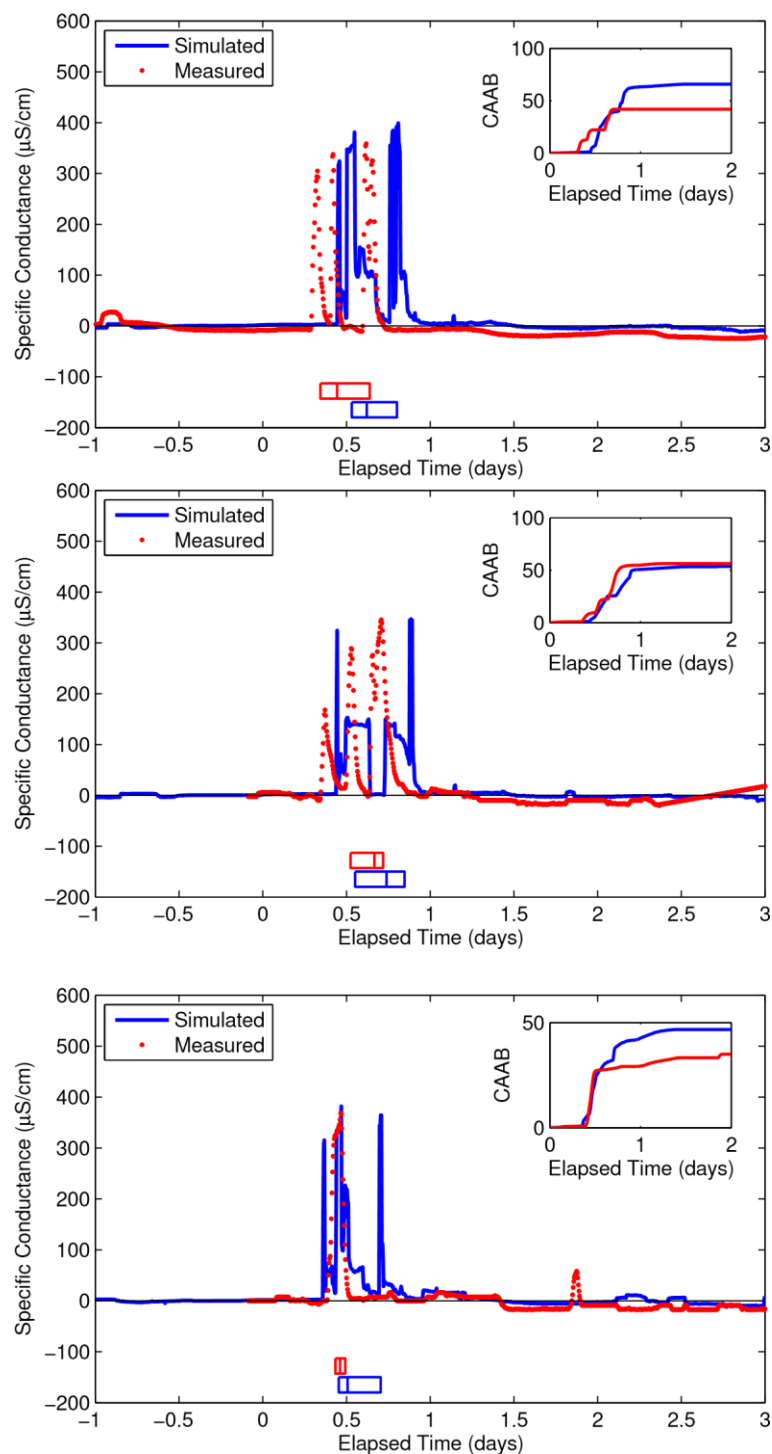


Figure 4.3-18. Observed and simulated tracer movement, Locations F1 (top), F2 (middle), and F3 (bottom). Inset figure shows the CAAB over a 2-day period commencing with the start of tracer injection, for both simulated and observed time series. The symbols along the x-axis show the *IQRs* (outer box boundaries) and medians (lines within boxes) for each time series.

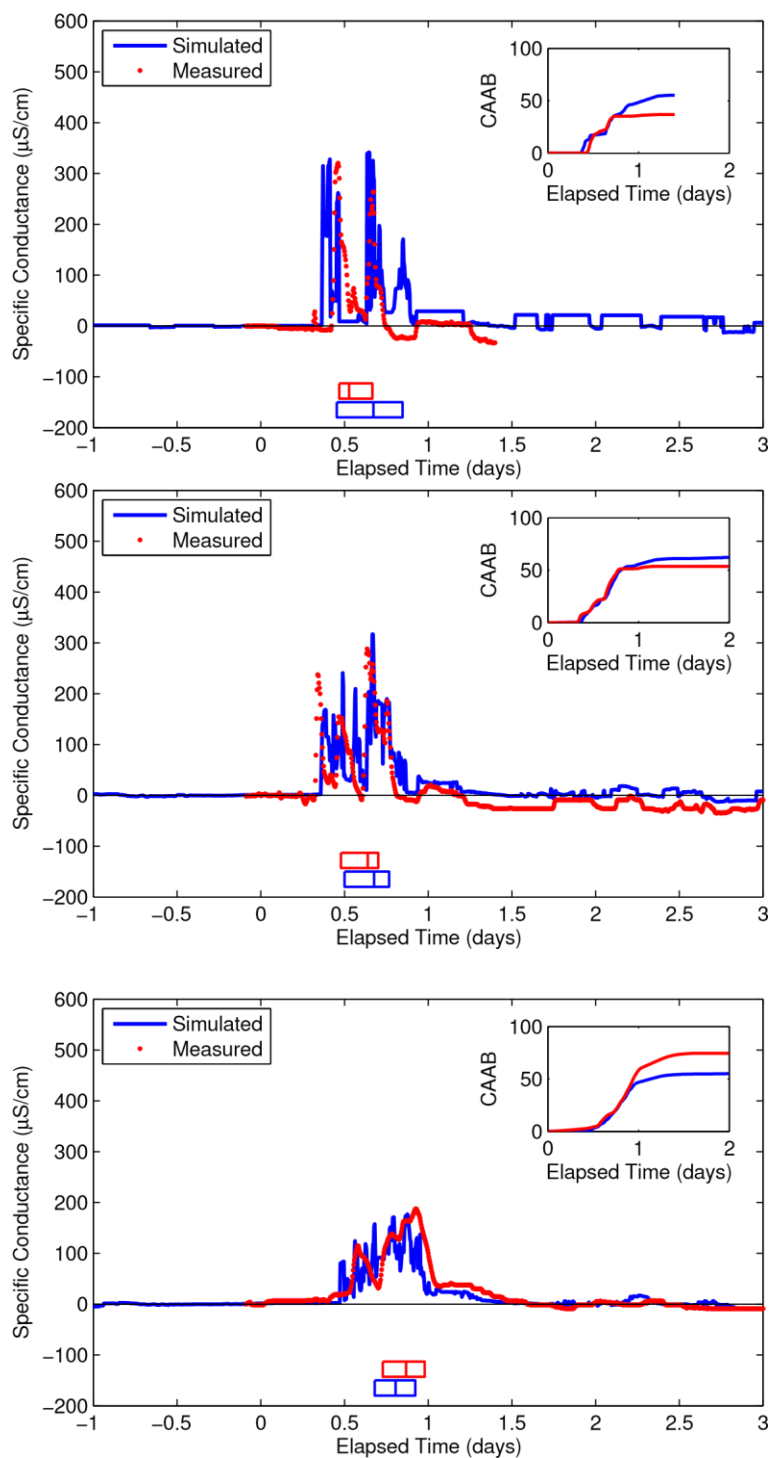


Figure 4.3-19. Observed and simulated tracer movement, Locations F4 (top), F5 (middle), and F7 (bottom). Inset figure shows the CAAB over a 2-day period commencing with the start of tracer injection, for both simulated and observed time series. The symbols along the x-axis show the IQRs (outer box boundaries) and medians (lines within boxes) for each time series.

4.3.5 Summary Results

The key characteristics of simulated and observed tracer time series are summarized in Table 4.3-1. These include the quantiles and total CAAB, as discussed in Section 4.3.4. The absolute errors in median and *IQR* of CAAB, and the percent error in total CAAB, are summarized in Table 4.3-1. Table 4.3-2 gives the median and mean of each error statistic over all measurement locations; the median (mean) Q_2 error is 3.88 (4.47) hours, the median (mean) *IQR* error is 1.46 (2.57) hours, and the median (mean) C_{TOT} error is 37.11% (50.33%). These results suggest that errors affecting the average speed of the tracer pulses, as characterized by the Q_2 errors, dominate those that affect the dispersion of the pulses, as characterized by the *IQR* errors. The C_{TOT} errors usually reflect an attenuation of the observed time series relative to the simulated (if these are computed using signed instead of absolute errors, the median C_{TOT} error is +13.83%).

Table 4.3-1. Characteristics of simulated and observed tracer time series data. All quantiles in days and C_{TOT} in (day $\times \mu\text{cm}$).¹⁶

Location	Simulated				Observed			
	Q_1^{sim}	Q_2^{sim}	Q_3^{sim}	C_{TOT}^{sim}	Q_1^{obs}	Q_2^{obs}	Q_3^{obs}	C_{TOT}^{obs}
A2	0.615	0.753	0.833	67.946	0.632	0.788	1.122	64.096
A3	0.892	1.069	1.25	71.135	0.736	0.899	1.215	74.454
A4	0.569	0.767	1.073	71.277	0.66	0.823	1.118	72.018
A5	1.247	1.483	1.642	139.783	1.347	1.622	1.878	73.739
A6	0.667	0.757	1.167	75.481	0.646	0.785	1.115	54.429
A7	0.524	0.604	1.038	87.37	0.545	0.788	1.042	78.179
A8	0.913	1.045	1.476	78.415	0.66	0.833	1.135	70
B1	0.698	0.896	1.101	93.229	0.753	0.958	1.194	30.422
B2	0.49	0.597	1.035	66.101	0.726	0.847	1.156	65.67
B3	0.622	0.785	1.163	62.745	0.868	1.201	1.375	33.67
B5	0.486	0.618	1.021	65.624	0.618	0.983	1.306	122.286
B7	0.351	0.476	0.941	68.869	0.448	0.663	0.962	68.212
B9	0.389	0.503	0.941	78.368	0.521	0.788	1.01	83.297
C1	0.604	0.851	0.979	128.116	0.861	1.118	1.274	52.043
C2	1.097	1.177	1.243	83.718	1.399	1.538	1.771	40.529
C3	1.41	1.486	1.677	76.637	1.149	1.247	1.476	33.626
C4	0.674	0.837	1.108	77.546	0.642	0.792	0.979	48.172
C6	0.642	0.927	1.052	123.268	0.969	1.146	1.243	111.945
C7	0.542	0.826	0.889	116.072	0.663	0.872	1.052	78.424
C8	0.413	0.615	0.722	122.127	0.392	0.587	0.67	90.107
D1	0.924	1.069	1.337	133.188	0.601	0.799	0.986	62.177
D2	0.368	0.569	0.691	95.86	0.333	0.493	0.663	95.631
D3	0.688	0.847	1.003	34.971	0.229	0.431	1.028	13.403
D4	0.306	0.438	0.552	73.513	0.274	0.392	0.497	85.615
D6	0.559	0.892	0.92	60.895	0.611	0.705	0.819	128.969
D7	0.316	0.681	0.719	79.373	0.156	0.462	0.5	71.73

¹⁶ C_{TOT} is the total CAAB or conductivity area above background after 2 days of elapsed time.

Water Utility Case Study Using EPANET-RTX Libraries

Location	Simulated				Observed			
	Q_1^{sim}	Q_2^{sim}	Q_3^{sim}	C_{TOT}^{sim}	Q_1^{obs}	Q_2^{obs}	Q_3^{obs}	C_{TOT}^{obs}
D8	0.274	0.431	0.507	83.598	0.274	0.424	0.503	78.251
E1	0.889	1.021	1.285	45.998	0.622	0.865	0.941	77.388
E2	0.632	0.806	1.222	52.17	0.472	0.656	1.038	87.035
E3	0.948	1.049	1.319	48.707	0.736	0.882	1.156	21.49
E4	1.559	1.681	1.774	65.256	0.465	0.531	1.052	34.392
E6	0.413	0.653	1.066	48.179	0.392	0.552	0.632	38.88
F1	0.531	0.622	0.802	65.859	0.344	0.444	0.639	41.987
F2	0.552	0.74	0.847	53.608	0.524	0.667	0.719	56.286
F3	0.455	0.507	0.705	46.768	0.434	0.462	0.493	35.008
F4	0.455	0.674	0.847	55.345	0.469	0.528	0.667	36.993
F5	0.5	0.677	0.767	62.123	0.479	0.639	0.701	53.725
F7	0.681	0.806	0.924	55.101	0.729	0.868	0.979	74.474

Table 4.3-2. Differences between simulated and observed tracer time series data.¹⁷

Location	$ Q_2^{sim} - Q_2^{obs} $ (hr)	$ QR^{sim} - QR^{obs} $ (hr)	$\frac{100 \times C_{TOT}^{sim} - C_{TOT}^{obs} }{C_{TOT}^{obs}}$ (%)
A2	0.83	6.5	6.01
A3	4.08	2.92	4.46
A4	1.33	1.08	1.03
A5	3.33	3.25	89.56
A6	0.67	0.75	38.68
A7	4.42	0.42	11.76
A8	5.08	2.08	12.02
B1	1.5	0.92	206.45
B2	6	2.75	0.66
B3	10	0.83	86.35
B5	8.75	3.67	46.34
B7	4.5	1.83	0.96
B9	6.83	1.5	5.92
C1	6.42	0.92	146.17
C2	8.67	5.42	106.56
C3	5.75	1.42	127.91
C4	1.08	2.33	60.98
C6	5.25	3.25	10.12
C7	1.08	1	48.01
C8	0.67	0.75	35.54
D1	6.5	0.67	114.21
D2	1.83	0.17	0.24
D3	10	11.58	160.92
D4	1.08	0.58	14.14
D6	4.5	3.67	52.78
D7	5.25	1.42	10.65
D8	0.17	0.08	6.83
E1	3.75	1.83	40.56
E2	3.58	0.58	40.06
E3	4	1.17	126.65
E4	27.58	8.92	89.74
E6	2.42	9.92	23.92
F1	4.25	0.58	56.85
F2	1.75	2.42	4.76
F3	1.08	4.58	33.59

¹⁷ C_{TOT} is the total CAAB or conductivity area above background after 2 days of elapsed time.

Location	$ Q_2^{sim} - Q_2^{obs} $ (hr)	$ IQR^{sim} - IQR^{obs} $ (hr)	$\frac{100 \times C_{TOT}^{sim} - C_{TOT}^{obs} }{C_{TOT}^{obs}}$ (%)
F4	3.5	4.67	49.61
F5	0.92	1.08	15.63
F7	1.5	0.17	26.01
Median	3.88	1.46	37.11
Mean	4.47	2.57	50.33

In an attempt to identify trends within the location error statistics, the three error statistics are plotted in Figure 4.3-20 as a function of pipe diameter at the measurement hydrant, after isolating those measurement locations at storage tanks and on dead-end mains. The horizontal and vertical lines on these plots indicate the median errors. The same error statistics are plotted in Figure 4.3-21 as a function of location region, by the same location labeling convention used in Figure 2.1-1. Neither of these analyses present visually compelling arguments for clear trends in the error statistics, of the sort that could help explain the sources of the simulation errors. It might be said that dead-end locations, for example, are poorer overall in simulation accuracy, as 5 of 7 have C_{TOT} errors that exceed the median, yet the same statement does not hold up for the quantile error statistics. In any case, the expectation that dead-end mains would have strikingly different, and poorer, overall simulation accuracy compared to looped mains does not hold up to scrutiny in this case. Similarly, the regional categorization does not identify any one region with strikingly different overall simulation accuracy. While it could be argued, for example, that Region A presents overall improved accuracy compared to B, especially in terms of the CAAB median, the results do not indicate a clear and consistent trend. Perhaps the lack of such trends reflects on the overall difficult test posed by the tracer study design, with 24 of 37 monitoring locations on dead-ends or mains with diameters less than or equal to 8 in. Just as likely, it may reflect the fact that the complexity of transport dynamics in looped networks cannot be captured using simple concepts related to local pipe characteristics or geographic proximity.

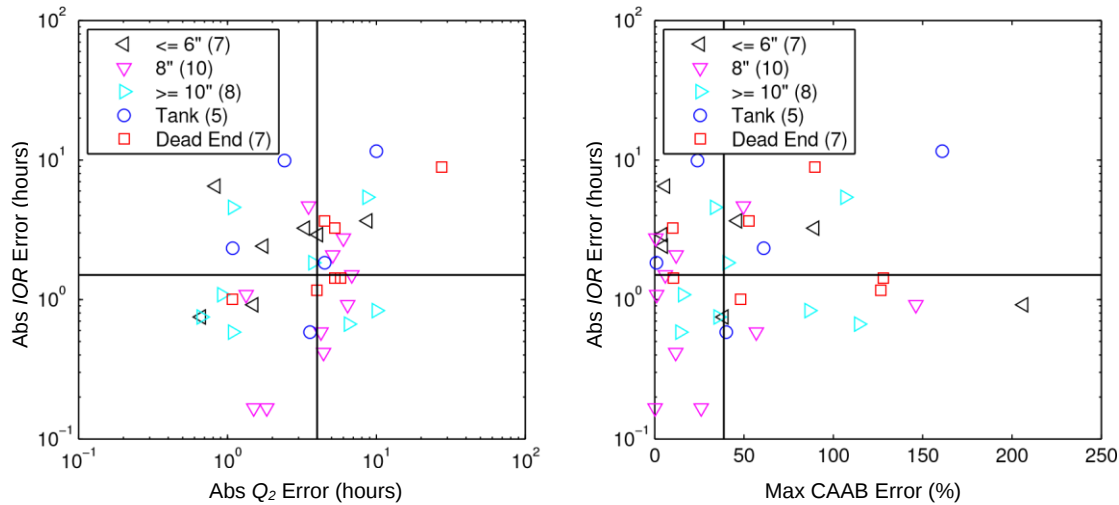


Figure 4.3-20. Comparison of tracer data and real-time simulations at 37 monitoring locations, as a function of pipe diameter or location characteristic. Conductivity results observed at storage tanks and dead-end mains are isolated for comparison and are not included in the pipe diameter categories. Statistical comparison of observed and simulated conductivity pulses is focused on three pulse movement characteristics: absolute error in *IQR*, absolute error in median (Q_2), and percent error in the total CAAB, measured 2 days after tracer injection.

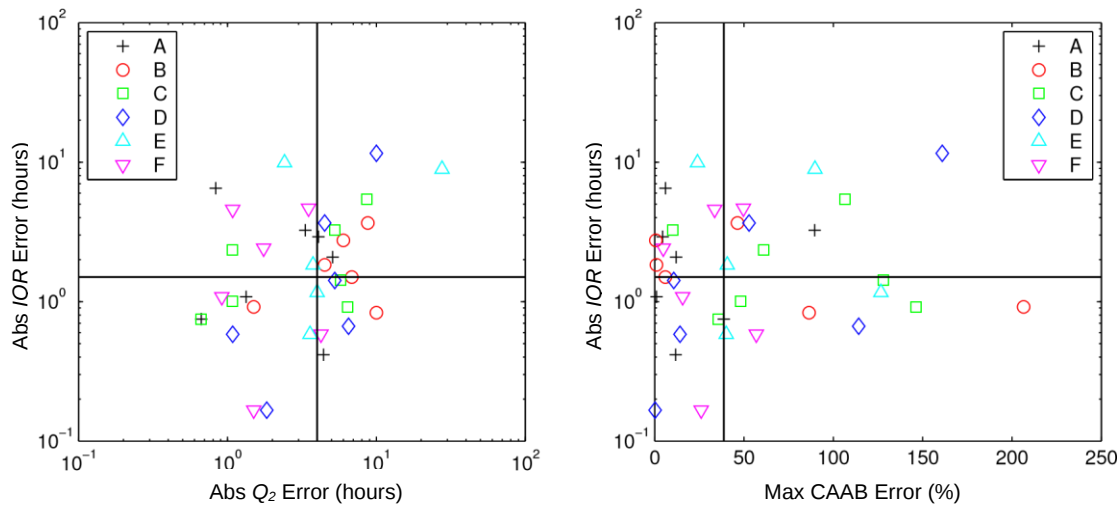


Figure 4.3-21. Comparison of tracer data and real-time simulations at 37 monitoring locations, as a function of location region (A-F). Statistical comparison of observed and simulated conductivity pulses is focused on three pulse movement characteristics: absolute error in *IQR*, absolute error in median (Q_2), and percent error in the total CAAB, measured 2 days after tracer injection.

5.0 Water Security Application: Demonstration of Using Real-Time Water Quality Simulation Results for Contamination Detection

If real-time hydraulics and water quality can be represented with sufficient accuracy, then they could be used toward real-time detection of water quality anomalies (sometimes called “event” detection) and fault identification/isolation, with wide-ranging benefits for asset management and public health protection. We provide a framework for model-based detection of water quality anomalies. We believe

this analysis is a first. While preliminary, we provide a quantitative evaluation of a real-time model-based contamination event detection approach versus the prototypical methods being used for contamination warning. The model-based algorithm relies on a real-time hydraulic and water quality model to simulate water quality results for comparison with water quality conductivity measurements. The difference between the modeled and the measured water quality values is processed (e.g., filtered) to detect changes that are inconsistent with noise, and thus likely to be associated with a contamination incident. Preliminary results using the NKWD tracer data show that most conductivity monitoring locations could benefit from a real-time model-based event detection method, in terms of enhanced sensitivity to real water quality anomalies. Specifically, in Section 5.1, we demonstrate and find that the median event detection threshold when processing conductivity signals using a model-based approach can be as much as about 30% less than when processing the raw signals against an accurate baseline. This decrease in detection threshold implies a more sensitive detector, yet one with a similar false positive rate.

Current contamination warning systems (CWSs) use a network of multi-probe water quality sensors to measure water quality and detect anomalous behavior (i.e., “event”); these anomalies are assumed to be characteristic of upstream contaminant introduction. For example, various microbial or chemical contaminants can be expected to produce deviations in disinfectant residual, total organic carbon (TOC), pH, specific conductance, and turbidity, compared to “normal” background water quality.

The current signal processing approaches used by CWSs consist of simple filters applied individually to each set of water quality signals at each monitoring location. There is typically no understanding of the dynamic operational decisions that affect water quality within the distribution system, nor any attempt to link the signals at different monitoring locations to understand system-level behavior. A practical problem with this simple approach rests in the determination of “normal” water quality for comparison against the real-time signals. This process is often referred to as “background estimation.” This is probably inappropriate because the term “background” implies a steadiness that should not be assumed characteristic of water quality in the distribution system (e.g., Uber et al., 2004). Water quality metrics vary over time — and perhaps rapidly — due to variability in source water quality, treatment efficiency, water usage, and system operation. Thus the determination of an “abnormal” water quality deviation is a difficult problem that should be solved in an adaptive manner, in real time. Standard approaches like control charts and thresholds can be used, based on historical water quality observations, but such approaches may not be sufficiently sensitive to water quality fluctuations if the thresholds are set to reduce the false positive rate to acceptable levels. Essentially, a problem with the typical CWS signal processing approach is that the natural variability in the “background” signal could be on par with the variability expected from real contamination events. In such a case, one approach is to attempt to estimate a more accurate background signal, one that reflects actual system boundary conditions and operational decisions — in short, a background obtained using a real-time predictive model that incorporates such factors. An alternative approach is simply to reduce the sensitivity of the detector by raising the threshold required for an alarm. While simple and effective for reducing false positives, the resulting decrease in sensitivity will harm the ability of the detector to identify real contamination events. For this reason, raising the threshold to reduce false positives should be considered a questionable practice.

Real-time model-based event detection is a novel approach that attempts to reduce the false positive rate without damaging sensitivity (or, equivalently, increase sensitivity without damaging the false positive rate), by filtering sensor water quality measurements prior to their analysis by the standard event detection algorithms. This filter is based on a real-time prediction of network hydraulics and water quality. Essentially, the real-time prediction of what the water quality *should be* is subtracted from the

sensor measurements, leaving a signal that reflects only variability that is unexplained; if such variability triggers an alarm, then it is more likely that the alarm is real.

The heart of a model-based event detection scheme is the real-time hydraulic and water quality model used to generate the prediction error signal between the measured water quality and corresponding model predictions. The advantage of processing such time series signals for event detection, rather than the raw local water quality signals, is improved sensitivity and specificity. This improved performance derives from the ability of the water quality model predictions to incorporate network-scale transport, known operational changes, and measured source water quality. Nevertheless there is understandable skepticism about the damaging effects of various modeling errors, and currently it is unknown whether such errors can be controlled sufficiently to yield the demonstrated increased sensitivity and specificity discussed next.

5.1 Evaluation of Real-Time Model-Based Event Detection Using the NKWD Tracer Study Data

Here we report preliminary results from the first evaluation of a model-based approach for detecting anomalous water quality events using real-time model predictions and field scale measurements. The test site and data set is provided by the NKWD tracer study previously described (Section 4.3), and the real-time hydraulic and water quality model is the EPANET-RTX water quality predictions, also evaluated previously (Section 4.2). The tracer data do not, of course, constitute a contamination event, so one must first put into context how such data could be used to evaluate the promise of model-based event detection.

The tracer conductivity data were regarded as representing normal system operational signals, and the challenge (to the real-time model-based approach) is to represent the causes of the signal variability so that it is discovered as normal, and not anomalous, behavior. This approach is defensible, because the variability in tracer signals is not unlike the variability that one can expect to see with water quality signals collected in distribution systems. The rapid and significant changes over time are not unlike the changes commonly observed in free residual chlorine due to changes in system operation and specifically tank operation (Uber et al., 2004). This “background” signal variability due to boundary condition and operational changes is likely responsible for the high false positive rates that have plagued current CWSs. These signals also, however, represent a challenge to model-based event detection methods, to track them closely enough so that subtraction of the model prediction allows the resulting signal to be classified as “normal.” Finally, the tracer study data set is unusually rich, with 38 monitoring locations available for evaluation within the study area.

To evaluate and compare both simple and model-based event detectors, we use the simplest prototype event detection algorithm, described below. We wish to emphasize that while more sophisticated approaches can be used, both the simple and model-based approaches use the same algorithm, just applied to different signals.

1. Create a filtered signal $\hat{S}$ from the original signal to be processed, S , using a standard moving average filter with averaging time window W_s . As with all filters, such an algorithm would be easily deployed as a real-time process.
2. Create a binary status signal $B = (S > T)$, where T is the detection threshold.
3. Raise an alarm if status B consistently equals 1 within a detection time window W_d ; otherwise no alarm is raised.

This simple event detection algorithm was applied at each monitoring location to both the absolute conductivity above background signal $S = |C - C_b|$, and the real-time model absolute prediction error signal $S = |C - C_m|$, where C is the measured specific conductance ($\mu\text{S}/\text{cm}$), C_b is the mean “background” concentration prior to tracer injection (a constant), and C_m is the specific conductance predicted by the real-time model.

For each of the recorded conductivity observations and predicted signals at each of the 38 conductivity monitoring locations, we calculated the *minimum detection threshold* T , such that B does not raise any alarm during the time period of the tracer monitoring. If T^* denotes a minimum threshold applied to the signal $S = |C - C_b|$, and T_m^* is the minimum threshold applied to the signal $S = |C - C_m|$, then the (location specific) detection threshold ratio T_m^*/T^* is a metric related to the potential improvement in detection sensitivity when using a real-time model-based event detector. As both T^* and T_m^* are calculated to result in zero (false) positive alarms, both are equivalent in terms of the false positive rate, for the evaluation data set. If, say, $T_m^*/T^* = 0.5$, then the 50% decrease in the model-based event detection threshold is directly available to decrease the concentration at which the water quality contamination event could be detected. A global measure of the potential power of using a model-based prediction error signal is the median value of the detection threshold ratio T_m^*/T^* . If the median value of the detection threshold is significantly less than 1, then it would indicate a real-time model could add value, in terms of increased sensitivity, for event detection at many monitoring locations. If the median value of the detection threshold is significantly greater than 1, then that would indicate that the real-time model predictions are not sufficiently accurate to use for real-time event detection at most monitoring locations.

Table 5.1-1 shows the minimum detection thresholds and the threshold ratio for all 38 conductivity monitoring locations, using the simple event detection algorithm with $W_s = 12$ hours and $W_d = 4$ hours. (Various event detection parameters were experimented with and results, in terms of the ratios T_m^*/T^* , were not observed to be overly sensitive to reasonable ranges of these values.) The median threshold ratio is 0.68, indicating that the detection threshold using model-based real-time signals could be set at least 30% lower (1 minus 0.68) compared to using an accurate background conductivity, for half (given the median threshold ratio) of the 38 monitoring locations. This is a positive and encouraging result for using real-time water quality model predictions, and indicates that further study should consider methods for improving the prediction accuracy for specific conductance and other common water quality signals.

Figure 5.1-1 shows the histogram of threshold ratios for all 38 monitoring locations. As shown, a significant number of locations would not benefit, in this data set, from a model-based event detection approach. On the other hand, the real-time model prediction accuracy at other locations is sufficient to allow for dramatic reductions in the detection threshold. It would seem logical that any implementation of real-time event detection should consider the estimated real-time model accuracy, as a function of location, when making decisions about monitor locations. Alternatively, additional model calibration work should be anticipated and planned for if monitors need to be placed in areas of low expected real-time model accuracy, to achieve water security objectives.

Finally, Figure 5.1-2 illustrates the event detection signals at three different monitoring locations, selected to represent the 10th, 50th, and 90th percentile of the detection ratios T_m^*/T^* (corresponding roughly to excellent, median, and poor behavior of the model-based event detection approach). These graphs make apparent the expected relationship between real-time water quality prediction accuracy and the opportunity for enhanced detection algorithm sensitivity.

Table 5.1-1. Minimum event detection thresholds for filtered measurement signals (T^*), filtered prediction error signals (T_m^*), and detection threshold ratio (T_m^*/T^*).

Location	T^*	T_m^*	T_m^*/T^*
A2	72.570801	31.311035	0.431455
A3	83.435059	28.747559	0.34455
A4	83.679199	22.888184	0.273523
A5	84.899902	103.45459	1.218548
A6	73.242188	37.658691	0.514167
A7	107.299805	73.425293	0.6843
A8	75.805664	52.246094	0.689211
B1	63.354492	84.899902	1.340077
B2	84.838867	49.865723	0.58777
B3	71.350098	55.908203	0.783576
B5	86.05957	59.326172	0.689362
B7	63.537598	24.353027	0.383285
B9	92.163086	36.804199	0.399338
C1	58.59375	132.019043	2.253125
C2	66.28418	112.426758	1.696133
C3	65.917969	87.890625	1.333333
C4	68.969727	37.658691	0.546018
C6	156.37207	69.641113	0.445355
C7	104.797363	66.40625	0.633663
C8	132.751465	46.325684	0.348966
D1	87.890625	127.075195	1.445833
D2	133.850098	15.075684	0.112631
D3	82.397461	80.01709	0.971111
D4	105.529785	27.160645	0.257374
D6	197.631836	174.01123	0.880482
D7	103.515625	61.462402	0.59375
D8	148.193359	10.131836	0.068369
E1	100.280762	107.60498	1.073037
E2	183.47168	200.805664	1.094478
E3	52.79541	46.630859	0.883237
E4	76.599121	120.178223	1.568924
E6	58.349609	27.160645	0.465481
F1	68.908691	62.866211	0.912312
F2	336.669922	354.797363	1.053843
F3	77.880859	97.229004	1.248433
F4	64.880371	39.489746	0.608655
F5	88.867188	70.922852	0.798077
F7	125.12207	141.296387	1.129268

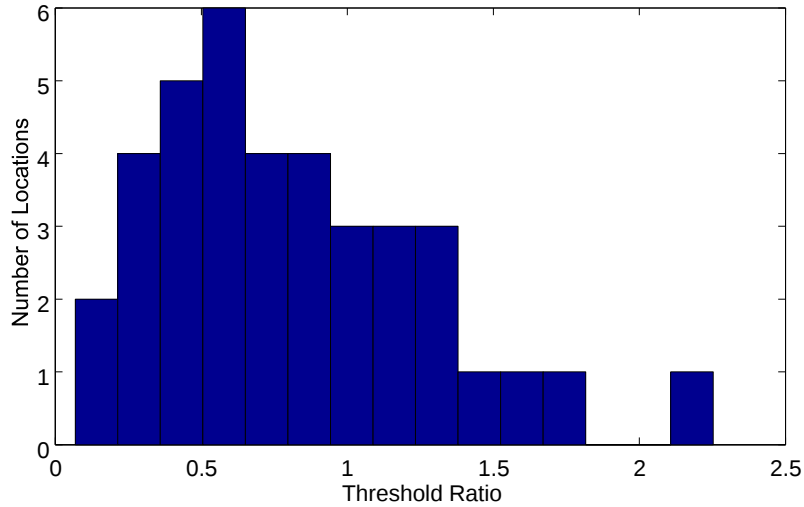


Figure 5.1-1. Histogram of detection threshold ratio, T_m^*/T^* , for all 38 tracer monitoring locations. The median threshold ratio is 0.69, indicating that half the monitoring locations would allow a decrease in the detection threshold (a tightening of the threshold used to trigger an alarm) of 30% or greater. This decrease should lead to increased sensitivity to actual contamination events, while leaving the false positive rate unaffected.

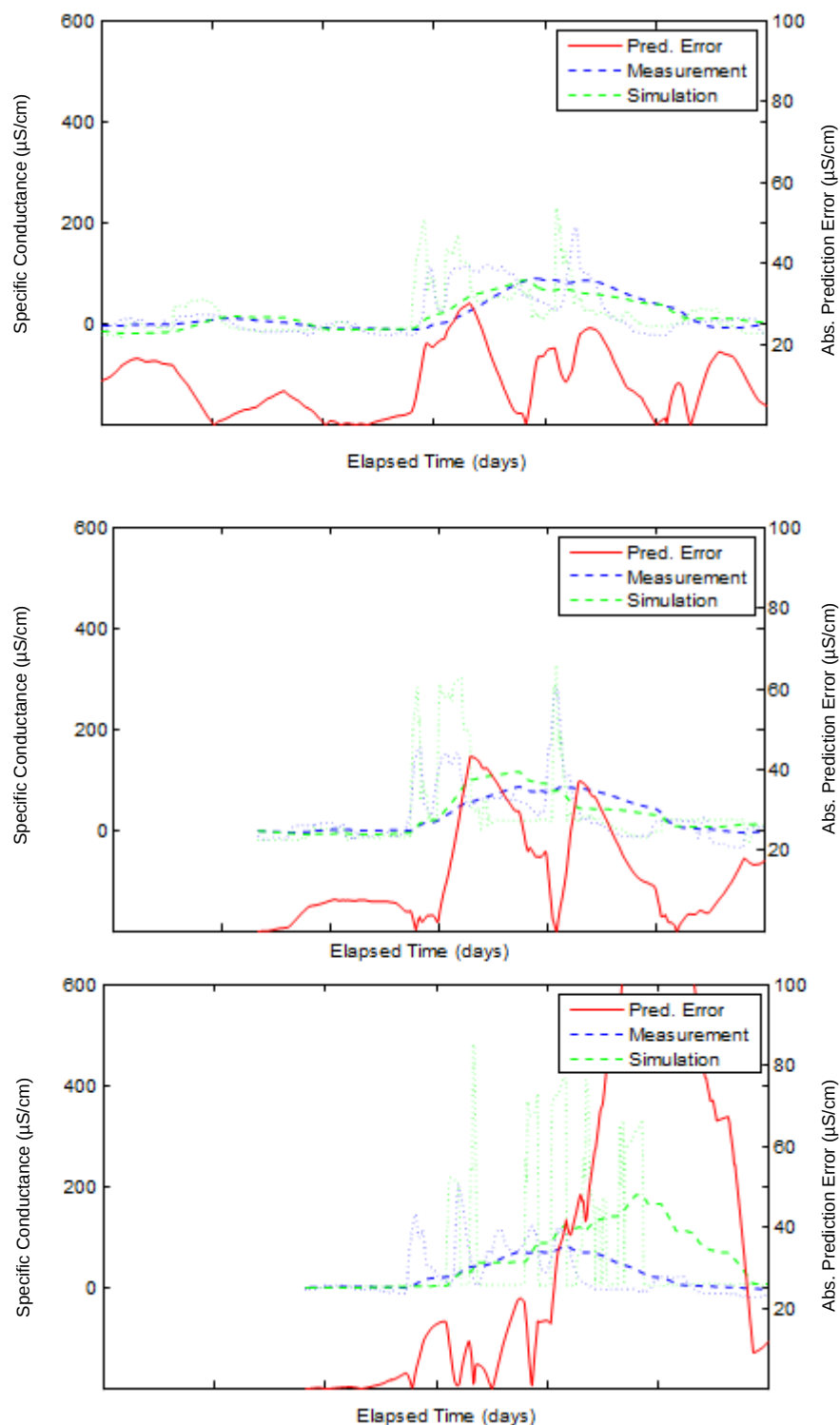


Figure 5.1-2. Illustrated event detection results for three different locations. Location A4 (top) represents the 10th percentile of threshold ratio T_m^*/T^* (0.27); Location A7 (middle) represents the median T_m^*/T^* (0.68); Location D1 (bottom) represents the 90th percentile T_m^*/T^* (1.45). The heavy dashed lines are the filtered measurement (blue) and simulation (green) signals, with the lighter dotted lines corresponding to the unfiltered signals. The solid red line is the prediction error — the difference between

filtered measurement and simulation signals — and is the signal used for real-time model-based event detection.

6.0 Outcomes

In this section, we describe a set of outcomes that resulted from the development and application of the EPANET-RTX libraries to our partnering water utility's (NKWD) network model and SCADA data assets. We define an *outcome* as either a specific deliverable provided to our partnering utility (e.g., improved network model) or a finding, strategy, or product provided to the wider water community to address a need or to demonstrate a useful result that could be obtained with real-time modeling.

On May 4 and 5, 2011, the American Water Works Association (AWWA), in collaboration with the USEPA NHSRC, convened the eighth meeting of the AWWA Water Utility Users Group in Cincinnati, Ohio (Janke et al., 2011). The AWWA established the Water Utility Users Group in 2003 to form a partnership between the USEPA NHSRC's research program and interested water utilities from across the country to investigate research questions associated with the design, implementation, and evaluation of contamination warning systems (Morley et. al., 2007). Partnering water utilities from the Users Group, such as NKWD, have collaborated with our research program by sharing information, data, and operational experiences. The partnership has helped to ensure that research is focused on the development of useful tools and methodologies that can be directly used by the water community (Morley et. al., 2007). Since the inception of the Users Group, utilities have emphasized the importance of field verification of models and tools and the development of improved capabilities for data analysis and data integration (Janke et al., 2011).

The focus of the May 2011 meeting was on developing and applying real-time modeling tools for water systems. The primary goal for the meeting was to receive water utility feedback on research priorities for advancing real-time modeling. One of the top research priorities identified was the development of methodologies and tools to support real-time hydraulic and water quality modeling (Janke et al., 2011). In this report, we demonstrated the application of the EPANET-RTX technology to a large and complicated water distribution system to support the refinement and calibration of the utility's hydraulic and water quality model. In Sections 6.1 and 6.2 we provide as a product to NKWD the identification of infrastructure model errors and an improved water distribution system network model, respectively. While the investigation is ongoing, in Section 6.3 we provide a discussion of how the application of the EPANET-RTX technology worked to identify a potential critical valve failure or sensor failure in the NKWD distribution system. The feasibility to easily and efficiently leverage existing model and SCADA data assets successfully for real-time modeling has been questioned by the commercial industry for more than two decades. In Section 6.4, we summarize three critical findings from this research with respect to using SCADA data for real-time modeling and simulation. In Section 6.5, using the case study demonstration results, we provide a detailed analysis of a water security application for real-time modeling. The water security application also demonstrates the use of real-time water quality simulation results for contamination detection.

In Section 6.6, as a result of this case study demonstration, we outline the five major steps for developing a real-time model and implementing real-time analytics using SCADA data assets. And, finally, in Section 6.7 we provide a discussion of some potential barriers to real-time model development and implementation for the water community.

6.1 Identification of Infrastructure Model Errors

We identify the major model modifications that were made to the NKWD infrastructure model in Appendix B.

6.2 Improved NKWD Model

Through this EPANET-RTX case study, we developed an improved NKWD model as demonstrated in this report and further detailed in Appendices B and C. With the implemented and recommended network model changes, NKWD is well positioned to implement a real-time model and begin investigating opportunities for improved operations and management practices.

6.3 Identification of Potential Valve Failure or SCADA Sensor Problem

As part of this case study, negative demands in a particular DMA were identified to have begun during July 2012 and continued through the present day (July 2014). This DMA is shown in green in Figure 6.3-1 and is associated with a pressure zone at service level 876 ft. This DMA is served by one set of high service pumps, and three sets of booster pumps lift water 200 ft., from the 876 zone into the adjacent 1080 zone. Further investigation of the causes of the negative demand in 876 showed a sudden drop in DMA demand of approximately 2000 GPM occurring over several hours, and driving that demand to become negative with a corresponding sudden increase in the minimum nightly flow (demand) in the adjacent 1080 DMA. The two time series showing this behavior are also shown in Figure 6.3-1, along with long-term average trends. The events indicated in the figure are generated by the application of simple statistical control chart concepts to the data series, and would serve to notify utility personnel of a potential issue affecting the demands in these two adjacent DMAs.

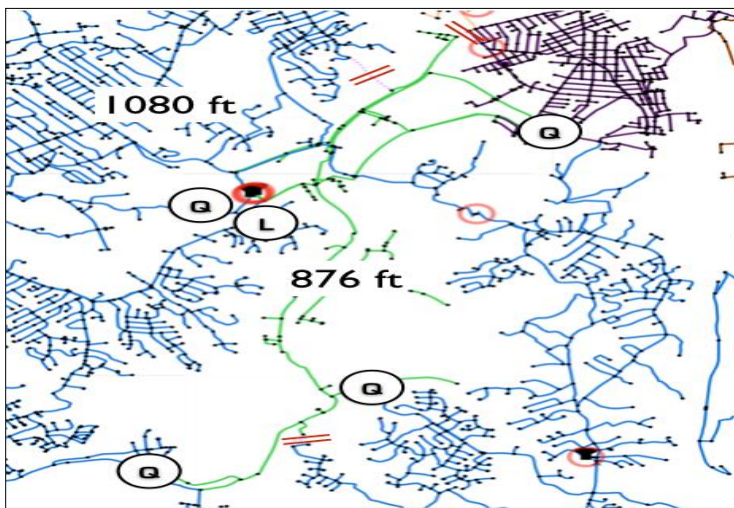


Figure 6.3-1. Identification of potentially excessive water pumping (over a long period of time) at partnering utility. Possible excessive pumping from a lower elevation district metered area (DMA) (green pipes) to a higher elevation DMA (blue pipes) due to a possible broken control or isolation valve at a higher DMA allowed pumped water to fall back each day to a lower DMA.

One possible cause of these statistical changes in real-time demands is a broken control valve or open isolation valve, allowing for water lifted up to the 1080 zone to fall back into the 876. Such a situation would appear as an unrepresented demand in 1080 (i.e., a “leak”) and an unrepresented supply in 876 (causing the apparent drop in demand). While this analysis was performed retrospectively as part of the case study, a set of real-time demand analytics would have been able to catch this problem when it

occurred, possibly with significant cost savings from reduced pumping during the 2 years that it went undiscovered. Indeed, a visual exploration shows that one of the pump station flows increases by 2000 GPM suddenly, coincident with the time when the 876 zone demand drops. An estimate of the costs of excessive pumping alone suggest an \$80,000/year electricity charge, and these costs could well be outweighed by infrastructure costs associated with excessive pump and impeller wear.

Another possible cause of the DMA demand statistical anomalies is flow meter calibration errors. Regardless of which cause is determined to be true, the power of real-time analytics lies in bringing such issues to the forefront so that utility personnel can diagnose them rapidly and put them to rest, before the anomalous behavior becomes the “new normal” — and hence forgotten. Utility personnel are currently investigating the piping detail of the particular pump station, including valve position testing and flow meter calibration, in order to accurately and efficiently identify and fix the root cause.

6.4 Demonstration of SCADA Data Evaluation, Analysis, and Use in Real-Time Modeling

This report represents a field-scale, quantitative demonstration of the fidelity of a real-time simulation model for hydraulic and water quality predictions. Documentation of accuracy for tank levels, flows, pressures, and pulsed tracer signals has been provided. While questions about the practicality of leveraging existing model and SCADA data assets have been present in the industry for more than two decades, the analyses and discussion provided here provide solid, evidence-based demonstration of the practical potential of real-time simulation technology. This outcome provides a clear and compelling example of how real-time data processing and modeling can be performed.

- Proof of ability to process ordinary/raw SCADA operational data streams to determine accurate pump status, tank inflow, pump station flow rate and head gain, aggregate district metered area demand, and altitude valve status, as well as ability to automatically drive network hydraulic and water quality simulations using such data streams, as enabled through the EPANET-RTX software objects and software derived from them. (See Section 3.2)
- Succinct documentation of time series data transformation pipelines (see Section 3.2.1 through 3.2.6) necessary for processing raw SCADA data streams into usable model inputs. This outcome provides important documentation for the successful data processing schemes developed as part of the EPANET-RTX research. These time series data transformation pipelines or schemes are fundamental to the EPANET-RTX software libraries (USEPA 2013).
- Proof of scalability of real-time (hydraulic and water quality) modeling tools to the industry, by showing that real-time modeling results obtained using an EPANET-RTX based application can be obtained without custom programming. Custom programming would likely limit applicability to only the largest utilities. This outcome demonstrates that the EPANET-RTX technologies for data acquisition, transformation, and analysis can be used to provide real-time predictions and forecasts for any water system that has a suitable infrastructure network model and sufficient SCADA data assets.

6.5 Demonstration of Using Real-Time Water Quality Simulation Results for Contamination Detection

We reported preliminary utility results demonstrating the first evaluation of using a model-based approach for detecting anomalous water quality events using real-time model predictions and field scale measurements. We demonstrated that using a real-time model-based event detection system could lead to a potential 50% decrease in the event detection threshold or concentration at which the water quality parameter could be observed to indicate that a contamination event occurred.

6.6 Steps for Developing a Real-Time Model and Implementing Real-Time Analytics

This case study has helped to refine the process for developing a real-time model and implementing real-time analytics using SCADA data. The five major steps for developing a real-time model and implementing real-time analytics are described as follows:

1. Gathering and analyzing network and SCADA data

- Collect network and mapping data. Identify SCADA system software characteristics and develop preliminary plan for data extraction.
- Harvest 1 year of SCADA data, at a minimum, from the SCADA historian database, through the creation of a virtual version¹⁸ of the SCADA historian machine. This is performed using appropriate virtualization software, and the resulting virtual machine is then able to be run off-site and support historian database connections and queries that are identical to those that would be used in a live, on-site, implementation.
- Based on interviews of operations and engineering personnel, internal documentation about the network model, SCADA historian database, and standard protocols followed by operators, identify the linkages between model elements and SCADA data tags. Assess adequacy of SCADA data streams for real-time application.
- Develop strategies to overcome any significant obstacles related to linkage between SCADA data and model elements.

2. Data diagnosis

- Explore SCADA data to identify important gaps related to data integrity/reliability.
- Make recommendations for necessary improvements to instrumentation, data cleaning/filtering, and data access interfaces.

3. Building and testing the real-time hydraulic model

- Reconfigure the hydraulic network model for real-time application, including any necessary changes to model topology and detail in order to accommodate accurate data connections.
- Implement connections of real-time data streams to model infrastructure elements, including appropriate and necessary data transformations.
- Assess adequacy of district metered area real-time demand computations.
- Perform continuous simulation of the distribution system hydraulics that reflect the actual changes in system operation as a function of system demand. Examine various time frames within that represented by the harvested SCADA data, to explore real-time model validity.
- Summarize the performance of the real-time hydraulic model as compared to SCADA data.
- Identify potential model issues that may be responsible for perceived performance gaps.
- Propose calibration activities as needed for further diagnosing and correcting any such deficiencies.

4. Perform model calibration activities

- Develop calibration strategy (including detailed work tasks and budget).
- Perform calibration activities.
- Reassess the performance of the real-time hydraulic model as compared to SCADA data.

5. Product delivery and demonstration

- Demonstrate end-user real-time software. Operational planning exercises should be developed to highlight the key decision support role of the real-time model. Demonstration

¹⁸ Essentially a virtual version is a copy, but with the added advantage of being able to access the SCADA database through the potentially, proprietary historian software.

should involve utility operations and engineering staff using the software to develop and evaluate alternative operational schemes and their benefits.

6.7 Potential Barriers to Real-Time Implementation

The future of drinking water distribution system network modeling will undoubtedly rely on an increased level of model-data integration. The drivers toward that end are strong: less expensive sensors and data communications technologies are creating a richer data environment in terms of number, type, and density of information; less expensive and more powerful ways to store and access data with mobile platforms are challenging traditional solutions that tie analysis to the desktop; and increasing interest in the water sector from traditional information technology companies will drive up the technology expectations of water utilities. There will, however, likely be a period of transition, and given the heterogeneity of the water industry sector in terms of size/resources, this period may be prolonged, especially for the (significant) number of smaller utilities or those resource limited. In this section, we discuss some significant barriers to implementation of real-time model-based solutions in the short run.

One significant barrier to real-time data integration technologies at many water utilities is the development and active promotion of a vision for the continuous use of data to diagnose and solve problems within the utility organization. This barrier may be difficult to overcome. Currently, data is highly fragmented in terms of its availability outside the group of primary owners. Movement of data within the organization is not automated, and usually includes a human “in-the-loop.” This fragmentation and lack of automation limits the expectations of data use and quality throughout the organization. Individuals expect that data is hard to obtain or, once it is obtained, data quality is to be suspect and challenged, and difficult and time consuming data transformations are likely to be required prior to putting data into practical and beneficial use. Water utility managers need to set expectations for data access and quality within their organizations, and put systems into place that make such visions real. Yet those same leaders currently lack access to information about what expectations and visions are realistic, and what the benefits to their organizations would be from implementing a utility-wide data management plan. While such plans should be regarded as foundational — just like the development and deployment of SCADA systems for operation and control — low expectations for efficient resolution of complex problems are continuing to restrict the sense of urgency that is necessary to drive investment in data technologies.

Another barrier to real-time data integration — very much related to the lack of vision for continuous use of data — is the particular problem involving the collection, sharing, and quality assurance of water pipe network infrastructure data. These data have largely been collected in isolation and under the purview of a small number of engineers, and it is not surprising that they are often not trusted outside of that small cadre of individuals. Specifically, operations staff are often distrustful that the infrastructure data, as encoded in a hydraulic network model, represent the true state of the system. All areas are open to suspicion — in particular the status of isolation valves and the accurate representation of pipeline, pump, and control valve characteristics. As part of the development of an organization-wide vision of data sharing and use, these suspicions must be allowed to surface, be discussed, and find resolution — in ways that are helpful and non-confrontational. It must be emphasized that such problems of trust in infrastructure data are, often frustratingly, the same problems that would be the target of a comprehensive vision of data sharing and use. That is why the development of such a vision is paramount in importance — without it as a framework to set expectations, the status quo of data access restriction and mistrust is allowed to continue, often unspoken except within trusted groups.

Finally, the simplest barrier to overcome is the one of insufficient data that must be used to drive real-time model predictions. This is the simplest because of the availability of straightforward requirements

and software/hardware technology for implementation. The first requirement is that SCADA data access must be available either through industry-standard OPC data connections or through SCADA historian databases. The choice of data access method will depend on the downstream software requirements, but SCADA historians are implemented using standard database technologies and allow for flexibility in integrating with various applications. Thus one very straightforward question is whether or not such a historian was implemented with the SCADA system delivery, as it is normally an optional product. Connection through OPC data access layers can create tension with SCADA managers, who are always concerned about outside software connections to SCADA and may want to scrutinize downstream software design and requirements.

Once data access is assured, minimum data requirements can be presented and assessed. Here are probably the top 5 highest priority data requirements:

1. First, a system flow balance must be possible using only real-time flow and level sensor data (i.e., not relying on assumptions about average production rates).
2. Item 1 essentially means that all floating storage tanks must include a level measure.
3. All supply sources must include flow measures allowing the calculation of cumulative flow rates at all points of entry to the distribution system.
4. It is highly recommended that all significant points of supply have actual flow measures — in particular, significant unregulated wholesale supplies should be metered, and those values should be available through real-time data acquisition, rather than being replaced by an average rate or historical diurnal curve (while the latter approach is feasible for real-time model implementation, it ignores demand variability and as such is not recommended).
5. All significant control elements must have their status either monitored directly and available via SCADA, or have adequate data streams that allow their status to be inferred from real-time data analytics.

For Item 5, the control elements that must be included are any pump or control valve that is under direct operator or automatic control. Sometimes the necessary data streams may be available, but they may not have been configured for storage in the SCADA historian database — a problem that can be corrected via SCADA configuration. Some relevant examples of SCADA data requirements follow.

- Any variable speed pump must have the pump speed measured and available, as it determines the pump operating characteristics (essentially replacing the “pump on/off” statuses typical of fixed speed pumps).
- Any fixed speed pump must have its status stored directly from contact sensors or inferred from other data streams (other possible data streams include suction and discharge pressures, individual pump flow rate measurements, and pump run-time or kilowatt meters).
- Remote-control PRVs or other control valves must have their statuses monitored and available, such as the upstream and downstream pressures of a remote-controlled PRV, the sensed valve stem position of a remote controlled PRV (better), or the open/closed position of a remote-controlled valve (expressed as a percent or fraction open).

Finally, beyond these minimum data requirements, it is suggested that system sensing include:

- Suction and discharge pressures for all pumps at all pump stations.
- Total flow rates, for all pump stations.

- Multiple pressure sensors within each pressure zone, to allow for useful mapping of the system-wide hydraulic grade line.
- Any additional flow and pressure monitoring, to be selected with an eye toward the use of that data for improving real-time model calibration and accuracy. (For example, additional flow monitors can be chosen at locations that optimally subdivide the system into DMAs, within which separate demand balances can be constructed in real time.)

The average size of those DMAs (in flow units) is a measure of the real-time demand disaggregation, and logically the more disaggregated the real-time demands, the more powerful the real-time model predictions and associated data analytics (e.g., for non-revenue water or leak detection within DMAs).

7.0 Conclusions

We provided a comprehensive description of the development and performance of a real-time hydraulic network model, including a description of the data processing steps and an evaluation of model accuracy using all available operational (SCADA) data streams in a complex real distribution system. We also provide a comprehensive analysis and discussion of a large-scale calcium chloride tracer study. We compare EPANET-RTX-based simulation results to conductivity measurements obtained during the field study to evaluate the accuracy of the real-time water quality model. The work described here, however, is not meant to be complete, but rather only illustrative of the insight and value that can be obtained from the fusion of a network model with SCADA data assets.

We described and demonstrated EPANET-RTX technologies through a detailed case study analysis of the NKWD. The case study results presented here were obtained using PolarisTM, a commercially available product developed using the EPANET-RTX open source library of software objects. We provided EPANET-RTX-based hydraulic model performance results for a 1-week evaluation period and proved the feasibility of calculating accurate real-time simulations for complex distribution systems. We found correlation coefficients averaging approximately 0.80 for flows, pressures, and tank levels for the NKWD study area. Our demonstration was without the use of complex micro-calibration of system parameters. That is, real-time hydraulic simulation results were demonstrated and shown to be sufficiently accurate that water utilities can now investigate improving their existing work flows and designing new ones to achieve desired endpoints, such as improved operations and water quality management, emergency preparedness, or water loss determination.

We generally showed that real-time simulation results accurately reproduced the hydraulic behavior of the distribution system, as described by this set of SCADA measurements. We do not, however, conclude that the real-time model is validated. A denser grid of data points, as well as performing similar evaluations using the same network model at different times of the year, or in different operational modes would still be needed to demonstrate calibration. We believe these real-time simulation results are, however, encouraging. We believe that the results presented are especially encouraging given that the data processing and hydraulic simulation were automated and no special data processing was performed for the study time period analyzed. Our results indicate similar levels of accuracy achievable for other time frames, which is the promise of a real-time model.

We also fully described a water quality model tracer field experiment. Our field tracer study was unique in that it was likely one of a few distribution system water quality studies to follow a large volume of finished water through an extensive portion of the distribution system. Our study is the first study to specifically use real-time modeling to drive the tracer simulations, and thus evaluate the fidelity of real-

time simulation data processing techniques. Our study design represented a challenging test of water quality model accuracy, as 24 of 38 monitors were located on small diameter distribution mains (17) or dead-end mains (7); thus our test not only evaluated the ability of a real-time model to predict movement through transmission mains, but also evaluated the accuracy at a neighborhood scale.

We provided and described a set of outcomes that resulted from the development and application of the EPANET-RTX libraries to our partnering water utility's (NKWD) network model and SCADA data assets. We defined an *outcome* as either a specific deliverable provided to our partnering utility (e.g., improved network model) or a finding, strategy, or product now available to the wider water community to help address an identified need or demonstrate a useful result that could be obtained with real-time modeling. Here are the outcomes that we found:

- We provided to NKWD a list of infrastructure model errors and an improved water distribution system network model. (Sections 6.1 and 6.2., Appendices B and C)
- We provided an EPANET-RTX-based analysis finding a potential critical valve failure or sensor failure in the NKWD distribution system. (Section 6.3)
- We demonstrated the feasibility of easily and efficiently leveraging existing model and SCADA data assets for real-time modeling and we summarized three critical findings from this research with respect to using SCADA data for real-time modeling and simulation. (Section 6.4)
- We demonstrated a water security application using the EPANET-RTX technology showing how real-time water quality simulation results could be used for improved contamination detection. (Section 6.5.)
- We provided a concise list of five major steps for developing a real-time model and implementing real-time analytics using SCADA data assets. (Section 6.6)
- We provided a discussion of some potential barriers to real-time model development and implementation for the water community. (Section 6.7)

References

- Association, U. W. A. (1980). Report 26 leakage control policy and practice. Technical report. London: UK Water Authorities Association.
- Bartolin, H., Martinez, F., and Cortes, J. A. (2006). Bringing up to date WDS models by querying an EPANET-based GIS geodatabase. In *Proceedings of the 8th Annual Water Distribution Systems Analysis Symposium*, Cincinnati, OH, pp. 1-17. Baltimore, MD: American Society of Civil Engineers, A.S.C.E.
- Davidson, J. and Bouchart, F. J.-C. (2006). Adjusting nodal demands in SCADA constrained real-time water distribution network models. *J. Hydraulic Engineering*, 132(1):102–110.
- Hatchett, S., Uber, J. G., Boccelli, D., Haxton, T., Janke, R., Kramer, A., Matracia, A., and Panguluri, S. (2011). Real-time distribution system modeling: Development, application, and insights. In *Proceedings of the Eleventh International Conference on Computing and Control for the Water Industry, Exeter, Devon, UK*.
- Janke R., Morley K., Uber J. G., Haxton T. (2011). Real-time modeling for water distribution system operation: integrating security developed technologies with normal operations. In: *Proceedings, American Water Works Association, Water Security Conference and Distribution Systems Symposium*, Nashville, TN, 11–14 Sept. 2011. Denver, CO: American Water Works Association.
- Johnson, J. G., Allen, R., Green, A., and Molia, S. (2007). Using sensitivity analysis and SCADA/model comparisons to diagnose model and system issues. In *Proceedings of the World Environmental and Water Resources Congress*, volume 243: 492.
- Kang, D. and Lansey, K. (2009). Real-time demand estimation and confidence limit analysis for water distribution systems. *J. Hydraulic Engineering*, 135(10):825–837.
- Lindner, M. A. (2012). Libconfig: A library for processing structured configuration files. Technical report. Accessed October 1, 2014 at <http://www.hyperrealm.com/libconfig/libconfig_manual.html>
- Morley, K., Janke, R., Murray, R., Fox, K. (2007). Water utilities driving water security research on drinking water contamination warning systems. *Journal AWWA*, 99(6): 40–46, June 2007.
- O’Haver, T. (2013). A pragmatic introduction to signal processing. Technical report. Accessed on October 1, 2014 at <<http://terpconnect.umd.edu/~toh/spectrum/IntroToSignalProcessing.pdf>>
- Ormsbee, L. E. and Lingireddy, S. (1997). Calibration of hydraulic network models. *Journal AWWA*, 89(2):42–50.
- Rossman, L. A. (1999). The EPANET programmer’s toolkit for analysis of water distribution systems. In *Proceedings of the Annual Water Resources Planning and Management Conference*, Tempe, AZ. Volume 102. American Society of Civil Engineers, A.S.C.E.
- Rossman, L. A. (2000). EPANET 2 User’s Manual. Technical report. Cincinnati, OH: U.S. Environmental Protection Agency. EPA/600/R-00/057

- Shang, F., Uber, J. G., van Bloemen Waanders, B. G., Boccelli, D., and Janke, R. (2006). Real time water demand estimation in water distribution system. In *Proceedings of the 8th Annual Water Distribution Systems Analysis Symposium*, Cincinnati, OH. pp. 95.
- Uber, J. G., Murray, R., and Janke, R. (2004). Use of systems analysis to assess and minimize water security risks. *Journal of Contemporary Water Research and Education*, Issue 129: 34–40.
- USEPA (2013). EPANET-RTX Real-Time Extension for the EPANET Toolkit. Open source software object library, U.S. Environmental Protection Agency. Accessed on October 1, 2014 at <<http://openwateranalytics.github.io/epanet-rtx/>>
- Walski, T. (1983). Technique for calibrating network models. *J. Water Resour. Plann. Manage.*, 109(4):360–372.
- Wood, G. A. (1982). Data smoothing and differentiating procedures in biomechanics. *Exerc. Sport Sci. Rev.*, 10:308–362.

Appendix A: Prediction Accuracy of Chloride Levels Based on Measured Specific Conductance

The purpose of this analysis was to estimate the accuracy of chloride-level estimates based on specific conductance. Paired chloride and specific conductance measurements collected from Northern Kentucky Water District (NKWD) finished water (Fort Thomas Treatment Plant) between 9/7/11 and 8/22/12 were used in this analysis. A simple linear regression model was applied to the data set, as shown in Figure A-1. The prediction interval at a 99% confidence level was calculated. This provides the ability to determine a range for chloride concentration based on a given specific conductance, and based on this data set.

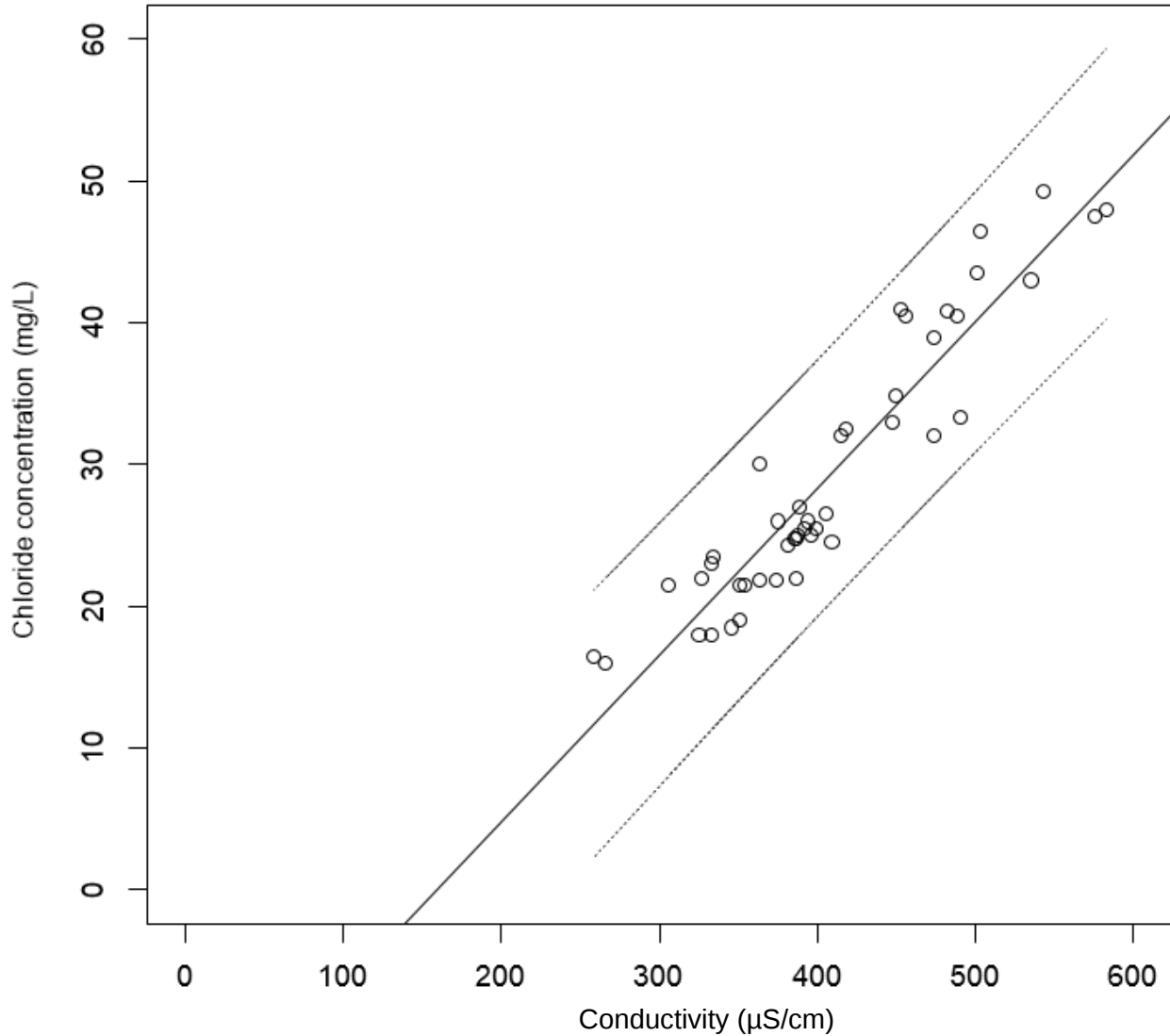


Figure A-1. Relationship between specific conductance and chloride measurements.¹⁹

¹⁹ 99% prediction interval shown. NKWD data collected between 9/7/11 and 8/22/12.

The following equation represents the lower bound ($[Cl^-]_{lb}$) on the prediction interval:

$$[Cl^-]_{lb} = 0.117EC - 27.5, \quad (6)$$

The upper bound ($[Cl^-]_{ub}$) is found by:

$$[Cl^-]_{ub} = 0.118EC - 9.68. \quad (7)$$

Using these two equations for a measured specific conductance value provides the 99% prediction interval on chloride concentration in that sample. For example, given a measured specific conductance of approximately 400 $\mu S/cm$, we can state that the measured chloride concentration will fall between 19 and 37 mg/L, with 99% confidence.

Appendix B: A Catalog of Operational Notes and Network Model Updates for the Northern Kentucky Water District (NKWD) System

B.1 Operational Notes

Important known operational issues that affected the drinking water distribution system during the November 2012 study period include the following:

1. Taylor Mill Treatment Plant (TMTP) was off-line 9/7/12-2/5/13. TMHS lift pumps 3, 4, 5 ran less than 1 hour combined during the study time range (to clear water diverted to clearwell by Fort Thomas Treatment Plant (FTTP) diversion valve, at a calculated average rate of 18 feet per minute [FPM]).
2. The 1017 pressure zone was “un-split.” Pipes 7555, P7454, and P7754 are closed to split the 1017 zone, and opened to un-split the zone. From the operational record, the 1017 zone split was modified: January 1, 2012 (un-split); July 23, 2012 (split); November 16, 2012 (un-split); December 12, 2012 (split); April 9, 2013 (un-split, confirmed through April 26, 2013).
3. Bromley Tank was off line for painting. The operational record for the Bromley tank level can be used to confirm when the tank was taken out of service and put back into service.

B.2 Model Updates

The following tables summarize structural modifications (Table B-1) and parametric modifications (Table B-2) that were implemented for the real-time hydraulic simulations described in Section 4.2. These modifications were made to the NKWD network model.

Table B-1. Summary of structural model changes.

Model Elements	Description
New Flow Control Valve TMTP DIVERSION VALVE	The TMTP diversion was modeled as a flow control valve at 18 GPM during the period the TMTP was off-line. Runtime data for TMTP lift pumps, plus assumed 8000 GPM for each pump, gives an average of 18 GPM diverted to the TMTP clearwell from the FTTP during the period when TMTP was off-line.
W. Covington PS	Modifications to include bypass (as per drawings)
FTTP Clearwell and Gravity Mains	Opened pipe P117 so that both clearwells feed the 763 zone. Also linked the US 27 PS so that flow can come from any of the three gravity feed mains from the two FTTP clearwells. This modification requires field confirmation, but it is consistent with SCADA data for the three FTTP clearwell pipes and the US 27 pump station flows, which show clear signature connections between each of the clearwell discharge pipes and the US 27 pump activity. There is essentially equal hydraulic pull on both clearwells, indicating they are not serving separate zones. Collapsed both clearwells to a single reservoir, to avoid recirculation between the two modeled clearwells due to the difference in head.
St. Therese Interconnect	Added St. Therese PRV.
Pipe 17000	Added pipe near US 27 per utility personnel.
Multiple Nodes	Added hydrant nodes for tracer study.
Multiple Pipes	Added pipes for tracer study, adjacent to added hydrant nodes.
Pipe Statuses	Updated pipe statuses: 9778 (OPEN).
Various Pipe Diameters	Updated various pipe diameters near several tanks and elsewhere with confirmation of utility personnel. IDs: 1031, 2911, 2912, 2913, 4751, 4783, 5131, 5132, 5133, 5138, 5179, 5180, 5181, 7394, 7395, 7398, 7604, 7605, 7643, 8100, 8209, 8228, 8229, 8230, 8650, 10214, 10215, 10220, 10222, 10228, 10229, 10230, 10231, CLARYVILLETANKPIPE, INDEPENDENCETANKPIPE, JOHNSHILLTANKPIPE, KENTONLANDSTANKPIPE.

FTTP, Fort Thomas Treatment Plant; GPM, gallons per minute; PRV, pressure reducing valve; TMTP, Taylor Mill Treatment Plant

Table B-2. Summary of parametric model changes or confirmation.

Model Elements	Description
PRV Settings	Site surveys collected upstream and downstream pressures collected using redundant high-precision analog dial gauges, and it was noted whether there was flow through the valve. Model PRV settings were updated to field downstream pressures for all flowing valves with active pressure control. For valves observed to be closed (BENTONRD, CENTERST, COVERTRUN, MOOCKRD, CARLISLERT10, WINTERSLANE, WOODLAWN), the measurement was considered an upper bound and settings were further adjusted downward so that modeled flow was zero. Note MEMORIALPRV is controlled in the real-time model using its downstream pressure as a setting boundary.
PRV Elevations	Pipe centerline elevation (above mean sea level) values were computed and updated from two sources of data. Centerline-to-landmark vertical measurements were made by tape measure, and mean sea level elevations for each landmark — usually the valve pit hatch cover — were retrieved from aerial survey data.
PRV Diameters	Updated from site survey which noted the true core valve diameter.
PRV Statuses	Changed the fixed status of the following PRVs from closed to none: NewportLow, Chesapeake2, LincolnPRV, Woodlawn, St. Therese Regulator.
PRV Minor Losses	Added minor loss coefficients for all model regulators from Cla-Val documentation as per the model number of the valve (provided by the utility) and the valve diameter (noted in the field). There are two main types, one with a reduced internal port size and one with a full-size internal port. The valve coefficients are given separately for these two types (and differ significantly). The reduced port is for the 600 series valves and loss coefficients are taken from Cla-Val data for the basic valve model 100-20, while the full size port coefficients are taken from data for basic valve model 10001. (Note: The K coefficients given by Cla-Val are the same dimensionless ones to be used for the model.)
Tank Elevations	Aerial survey data yielded the elevation at a certain landmark — usually a large poured concrete slab at or near the tank. Then the vertical distance between the landmark and the tank's pressure transducer was measured by tape. This measurement, combined with overflow pressure readings from the transducer (converted to feet of water), gave the overflow elevation for that specific tank. Design information on the maximum tank height then gave the corresponding bottom elevation.
Tank Minimum Levels	Modified all tanks so that minimum level = 0. Even if not physically realistic, it will increase operational flexibility for the real-time model. Unrealistically low values will be highlighted by mismatch with SCADA levels.
Tank Diameters	Reviewed and updated all tank diameters to be consistent with spreadsheet "tank summary for UC 2010 updated sept 8.xls," spreadsheet provided to water utility personnel. Note: Corrected 50 ft. discrepancy in the Dudley 1080 diameter.
Tank Altitude Valves	Updated tank maximum levels for Lumley, Main Street, and Rossford to reflect altitude valve maximums, as determined from SCADA.
Tank Volume Curves	Added/updated tank volume-depth curves for the following tanks, based on drawings and data provided by utility personnel: Barrington, Campbell County, Devon, Independence, Industrial, Kenton Lands, Lumley, Main Street, Rossford, South Newport.
Pump Station Elevations	Updated all pump station elevations, including assumed discharge/suction pressure node locations, to reflect updated information provided by utility personnel.
Pump Characteristics	Updated following pump head-discharge curves based on analysis of SCADA data: Bristow, Bromley, Carothers, West Covington, Dudley 1040, Dudley 1080, Hands Pike, Latonia, Richardson, Ripple Creek, US 27 1–6, Taylor Mill.
TMTTP Valving	The entire flow coming from TMTTP was discovered to be accounted for by the venture meter tracked in SCADA as TMTTP FI500, or model link 4602. The valve configuration in this area was inferred from the above information; manually checking all of the valves was impossible since some valve stems were not accessible by valve key. It was then inferred that all of the plant flow must be directed through just one of the two pipes exiting TMTTP; model link 4606 carries flow, and link 13326 does not.

PRV, pressure reducing valve; TMTTP, Taylor Mill Treatment Plant

Appendix C: Recommendations and Open Questions

The following are significant recommendations, and existing open questions, that were generated through the development of the real-time hydraulic model. While this list may not be complete, it is illustrative of the insight and value that can be obtained from the fusion of a network model with supervisory control and data acquisition (SCADA) data assets. This list includes issues related to SCADA, model infrastructure and operations data, and real-time model configuration.

1. Review the normal procedure for testing pressure reducing valve (PRV) settings to ensure reliable data are being collected. Equipment used for field pressure measurements should be upgraded. Procedures should stipulate how to reliably determine if the valve is active or closed, and how to identify settings accurately when there is no flow through the valve. This would presumably involve inducing flow through the valve by identifying a downstream hydrant for each valve that should be flowed when necessary.
2. Review and verify key PRVs that are SCADA controlled or used actively for pressure management have both stem position and upstream/downstream pressures transmitted via SCADA. This would allow the valve status to be determined, and thus enable accurate interpretation of downstream pressure values with respect to the valve setting. Such data would be important for consistent and reliable real-time model predictions.
3. Investigate and solve SCADA issues creating data gaps across wide spectrum of measurements. It does not appear that Delta storage mode is reliably configured in historian.
4. Investigate and fix SCADA measurements for: Bullock Pen meter pit flows 1 through 3 and pressure; Chesapeake 1 regulator pit flow (critical for district metered area [DMA] demand aggregation); Walton meter pit flow and pressures; Devou park pressures; St. Therese pressures; waterworks suction pressure; Taylor Mill clearwell (TM LI502); Latonia pump 1 non-reset runtime (should not reset)
5. Identify valid SCADA data stream (SCADA tags) that record speeds of waterworks variable speed pumps. The SCADA historian includes SCADA tags that should contain those speeds, but there are no data. This would be critical for real-time model predictions whenever waterworks pumps are running, and should also serve to identify typical operational modes for off-line model simulations. There is evidence in the total dynamic head and flow data from SCADA that speed is being varied, or is actively controlled to regulate discharge pressure.
6. Review and verify SCADA measurements. For instance, SCADA data indicate a 10-ft. head loss through the check valve at Ripple CreekPS at a flow of 800 GPM [gallons per minute] (when the pumps are off and flow is through the bypass). This is equivalent to a minor loss coefficient of 965, which is extremely high. The elevations of one or both of the pressure transducers may be in error, one or both pressure transducers may have a bias, or the check valve may be stuck. This should be inspected.
7. Review pump station flow signals. For instance, the Dudley 1080 pump station flow signal is very noisy; it is understood that this sensor was been replaced in 2013.
8. Investigate and confirm piping details surrounding the three gravity feed mains from the two Fort Thomas Treatment Plant (FTTP) clearwells. Confirm the SCADA data for the three FTTP clearwell pipes, and the US 27 pump station flows. These data show clear signature connections between each

of the clearwell discharge pipes and the US 27 pump activity. Particularly when one of the US 27 4–6 pipes is turned on, we saw about 2000 GPM increase from FTTP 3, 1500 increase from FTTP 2, and 1000 increase from FTTP 1, which is not entirely out of line with the approximately 4000 GPM increase out of the US 27 pump station. Also, we saw essentially equal hydraulic pull on both clearwells, indicating they are not serving separate zones.

9. Review tank curves. For instance, update the Ida Spence tank curve, as it is not a cylindrical cross section. Also, the maximum diameter appears to be larger than specified for the assumed cylinder, as per Google Earth™ calculations. It is understood that NKWD has no drawings; is the shape the same as the Kenton Lands tank, which was built at the same time? *Resolution:* Confirmed with utility that this is a good assumption — implemented. Also confirmed that Google Earth-measured diameters are approximately the same for both Kenton Lands and Ida Spence (approx. 50 ft.).
10. Confirm Memorial Parkway Treatment Plant (MPTP) clearwell bottom elevation.
11. Confirm diameters of pipes 8348 and 8444 (Rossford tank), 8993, 9539, 9540, 8993, and 15093.
12. Confirm status of pipes 9778, 10157. *Partial Resolution:* Confirmed status of 9778 should be changed to open — implemented.
13. Update the regulator valve diameters listed in “pressure regulator settings” to accurately reflect the valve internal diameter and not the pipe diameter. (Valve diameters in the model reflect the 2010 field observations.) *Resolution:* Conveyed list of all discrepancies to utility and referenced 2010 field observations for correct values.
14. Review hydraulic heads for certain tanks. For instance, the calculated hydraulic heads for Dayton and Bellevue tanks, using the light detection and ranging (LIDAR) ground elevations, puts Dayton at –3.82 ft. compared to Bellevue. This does not seem physically realistic, and SCADA shows that both tanks float together. Investigate LIDAR and SCADA elevation data to determine the reason for the calculated hydraulic head difference.
15. Update demands to latest billing data.
16. Real-time EPANET-RTX modeling team - experiment using step interpolation for all resampling of pump station flows that will ultimately be trimmed by the status time series. This allows gaps to propagate the last flow value when the pump is on, as opposed to interpolating to a flow when the pump might be off. Carothers pump 1 — 11/19/2012 — is illustrative.
17. Investigate altitude valve control for three tanks in 1017 (there may be others in zones that are outside the current study area): Lumley, Rossford, and Main Street. Lumley and Rossford may have more complex controls, compared to a max level cutoff. They may have a non-modulating level control valve or may be SCADA controlled. The piping and control systems should be modeled adequately in the real-time model. *Resolution:* Confirmed with utility personnel that both Lumley and Rossford have solenoid-controlled altitude valves that are set to close at one level and then open again when the level has dropped to a low level. Also these solenoid positions are under operator control. From SCADA level data, it is apparent that operators are using these solenoids, possibly in order to force turnover of these tanks. The existing EPANET-RTX pipe status time series seems an adequate method to mimic these valve positions, in lieu of actual data on solenoid position.
18. Add bypass PRVs to model.
19. Include measured pressures from field study in real-time model assessment.

20. Examine behavior of Memorial-Newport regulator; flow is zero in SCADA but simulation includes sporadic, yet significant, flow. Modeled setting may be too high.
21. Review treatment plant supply flows. For instance, the MPTP calculated supply flow should probably be changed to include a second moving average on the Actiflo[®] flow rate. *Resolution:* A second moving average operation was added to the EPANET-RTX configuration.
22. Investigate MPTP clearwell diameter. Current value is 150 ft; Google Earth puts the value closer to 170 ft. *Resolution:* MPTP clearwell diameter changed to 168 ft. in EPANET-RTX configuration. This value is not in the model since the clearwell is represented as a reservoir; no model modifications are required.
23. Assess modeled pump station infrastructure to reliably locate pressure transducers and their elevations, and represent minor loss components within each station (aimed at being able to use SCADA data more reliably to determine operating points).

SCIENCE



PRESORTED STANDARD
POSTAGE & FEES PAID
EPA
PERMIT NO. G-35

Office of Research and Development (8101R)
Washington, DC 20460

Official Business
Penalty for Private Use
\$300