Real-Time Network Hydraulic Modeling: Data Transformation, Model Calibration, and Simulation Accuracy

Technical Report

Prepared for

National Institute of Hometown Security
368 N. Hwy 27
Somerset, KY 42503

By

James G. Uber   Dominic Boccelli   Hyoungmin Woo   Yuan Su
The University of Cincinnati

Sam Hatchett
CitiLogics, LLC

October 2013

Contents

Abstract

We report on the process of creating a real-time network hydraulic model for the Northern Kentucky Water District (NKWD), including the underlying data transformation steps, and real-time model calibration. This is the first known study to document these steps in detail as well as the resulting real-time simulation accuracy. The accuracy of the real-time model was assessed using the complete operational record, for a one week period in November 2012. This week was selected to coincide with the time period of a full scale tracer study conducted in a portion of the NKWD service area, and which will be reported on separately. The real-time hydraulic model faithfully represents hydraulic SCADA data within the field study area (tank levels, pump station flows, and pressures), with average correlation coefficients of 0.79, 0.81, and 0.83 for all available flows, tank levels, and pressures, respectively.

1 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. Yet these investments should be further leveraged to support a wider scope of utility decision making. Indeed, 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 should yield numerous practical benefits, each enabled by the ability to simply and accurately forecast distribution system hydraulics and water quality, in real-time. Operators could 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 could 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 could 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. Such benefits are not unrealistic, and in fact are already supported by the existing investments in SCADA and 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 should be connected with network models, and the resulting accuracy of network simulation models that are driven by operational data. Absent that 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). And there will continue to be skepticism about the ability of network models, which were developed to support master planning, to provide meaningful predictions that reflect particular system operational decisions. This paper is an attempt to fill that gap in understanding, by describing the real-time network modeling process, and resulting hydraulic model accuracy, in the context of a case study conducted with the Northern Kentucky Water District.

1.1 Previous work

The use of SCADA data in water distribution system model development and calibration is not novel. A SCADA database may contain years of data for hundreds to thousands of relevant instruments, at sub-minute resolution – and the availability of these data is widely known. Yet published model calibration studies, and the standard practice in the field (**?**), remains focused on sparse, manually collected data sets for a small observation time window, perhaps 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 systems abundant data resources. In case a continuous stream of SCADA data is essential or advantageous for research purposes, (e.g., testing a state estimation methodology, as in **???**), it is usually based on synthetic data, perhaps with random noise superimposed to simulate real SCADA measurements.

Examples of SCADA-model fusion do exist, but are usually burdened by many intermediate steps between data access and synthesis, or by the closed philosophy of proprietary software systems (**??**). Software titles claim the ability to use "real time" data, but may in reality support only a batch-oriented "import" of SCADA information, requiring the export of a dataset as a text file, with offline processing. Such data connectivity can clearly be useful, but falls far short of the goal for real-time data fusion, which promises a persistent connection between model and SCADA database with automated data transformation and synthesis. Proprietary systems 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 USEPA National Homeland Security Research Center 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 (**??**). Epanet-RTX (RTX) was released as an open-source software project on September 24, 2012, to support commercialization opportunities[1]. It is intended that Epanet-RTX be 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.

[1]see http://openwateranalytics.github.com/epanet-rtx/

The real-time modeling results presented here use the Epanet-RTX library, and constitute the first full-scale case study of this technology.

1.2 Study goals

As mentioned, the fidelity that should be expected of real-time network hydraulic predictions is unknown - and therein lies a critical gap between research and development, and practical data fusion applications with measurable benefits. 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 can not 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 are likely. 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 a lower bar more than anything – representative of a significant initial effort.

1.3 Organization

Subsequent sections will describe the field study site, including the important SCADA data streams that were used, followed by a deeper discussion of Epanet-RTX, and the configuration of the real-time model. The field site and data are described first in order to provide context and motivation for the object design and functionality of Epanet-RTX. The configuration of the real-time model is focussed on the data processing aspects as well as how the data streams are connected to model elements. This is distinct, then, from model calibration, which is described next, as the various modifications that were made to the network model, in pursuit of (informal) maximizing of correlation between simulated values and SCADA data. It is stressed that there was no formal process of algorithmic calibration employed here; adjustments to the model were limited to those which made sense given the SCADA and infrastructure data. It is not claimed or assumed that these calibration activities have achieved the best relationship between data and model simulated values. We then present the detailed results that compare all available SCADA data streams with the real-time simulations, and draw conclusions. An appendix provides more detail about the decisions made in the model calibration process.

2 Field Study Description

2.1 Study Area and Distribution System Infrastructure

The Northern Kentucky Water District (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 Greater Cincinnati Northern Kentucky International Airport. It covers over 300 mi^2 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 daily 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 1 shows the study area, a subregion of the NKWD service area east of the Licking River with a total demand of 7.48 MGD. The northern portion of the study area (within the bounding box in Figure 1 is shown in greater detail in Figure 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 5) have been limited to the pictured subregion, and thus only the subregion results are discussed here. This particular study area was chosen because a tracer study was conducted in the same region during November, 2012. The real-time hydraulic results discussed here will be used to drive water quality predictions, providing a further rigorous evaluation of real-time model accuracy that complements the present evaluation. The study area consists of three hydraulically distinct regions, referred to as District Metered Areas (DMAs), and numbered 1-3 in Figure 1. The hydraulic characteristics of each study area DMA will be discussed and described further in section 5; they are introduced here for convenience, and they also serve to define clearly 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 1017 ft. From a broad topographic perspective, the study region – bordered by the Ohio River to the north and east, and by the Licking River to the west, which drains into the Ohio – is a ridge with watersheds that drain into either the Ohio or the Licking. The north-west 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 shows the locations of the two treatment plants (TP) within the study area, represented in the network model by reservoirs (head boundaries). Production from the northern TP supplies DMA 2 by gravity from the clearwell, and then DMA 3 through booster pumping. 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. The 1017 zone can be either "split" or

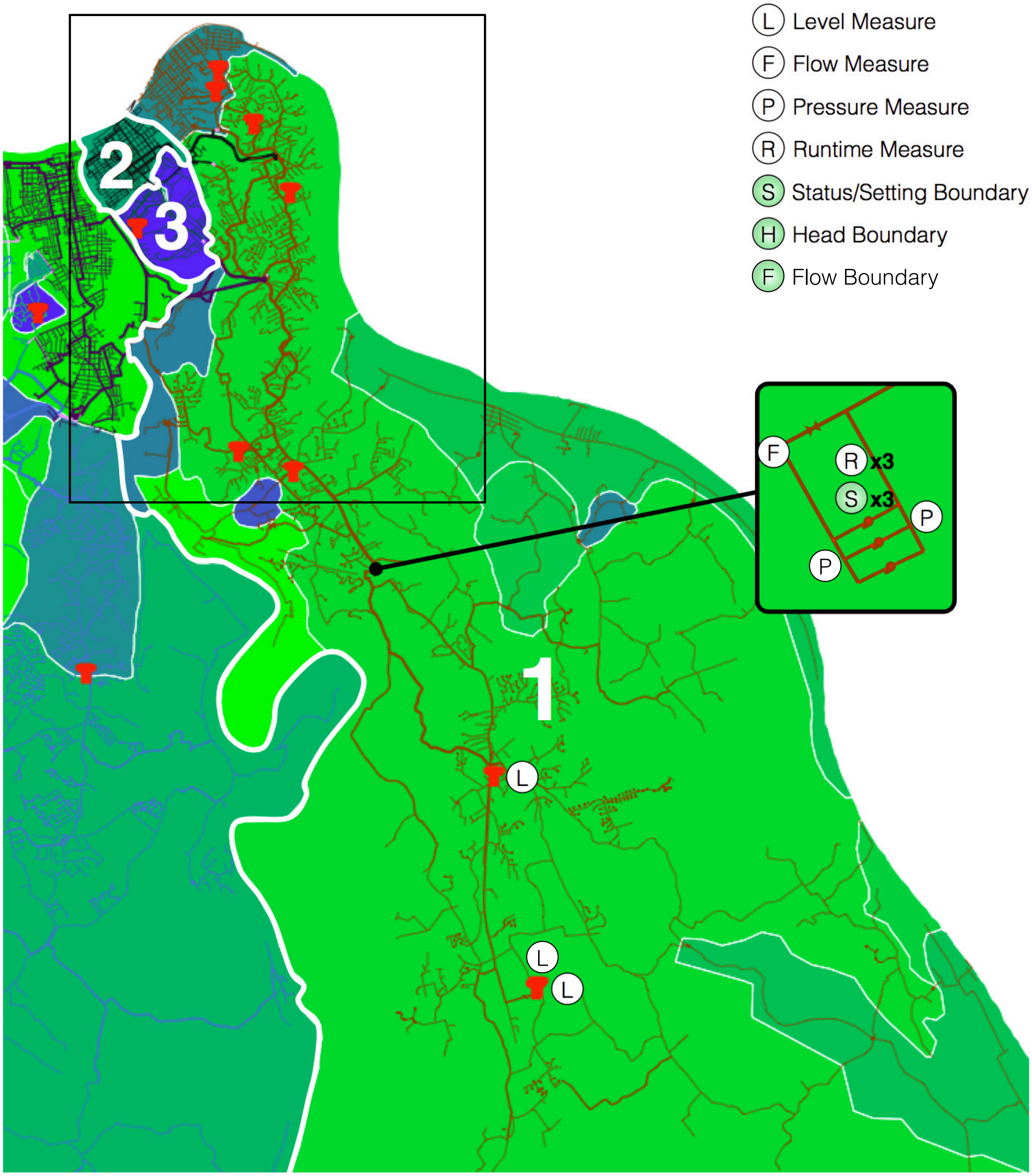

Figure 1: Distribution system study area map showing supply infrastructure, pressure zones, district metered areas, and categorized real-time data streams. Data streams within the bounding box are shown in Figure 2.

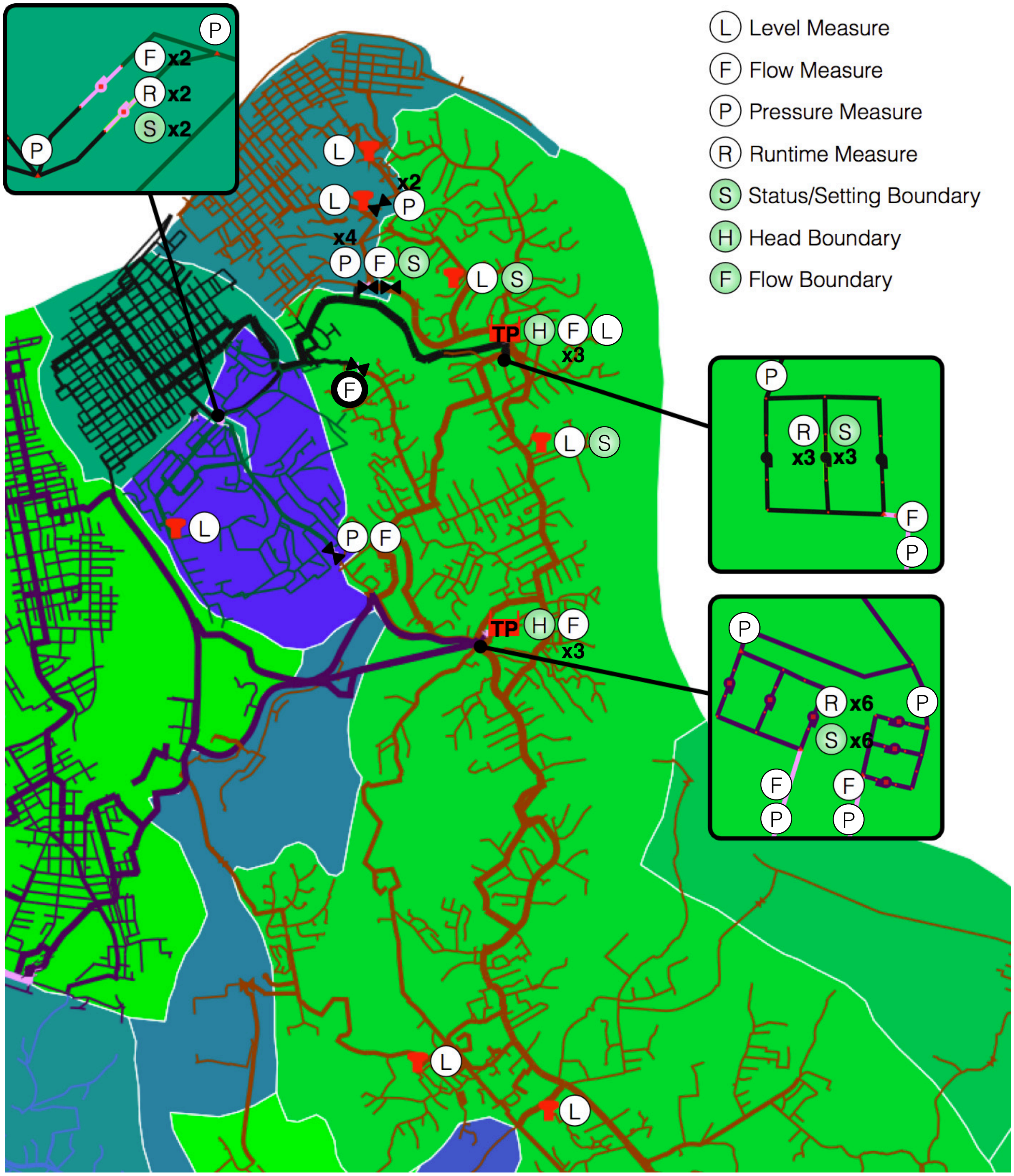

Figure 2: Northern portion of distribution system study area showing supply infrastructure, pressure zones, and categorized real-time data streams.

“un-split” depending on the status of three valves near 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 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, however, getting its clues directly from the real-time SCADA information (as discussed later in section 4.

2.2 Instrumentation: 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 that 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 1 and 2 show the approximate locations of measurement and boundary data streams within the study area. Measures are shown using open circles with a single letter indicating the type: water level (L), flow rate (F), pressure (P), and pump runtime (R). Boundaries are similarly shown using filled circles: pipe, valve, or pump on/off status, and pump or valve setting (S), reservoir head (H), and junction flow (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 (pump/pipe status, valve setting, head boundary, flow boundary, or demand). Even SCADA data used purely as measurements may sometimes be resampled and filtered, to reduce noise and 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 later, in section 4.

In general, each storage tank has a level measurement; each pump has a runtime measurement and status boundary; each pump station has suction and discharge pressure measurements and a station discharge flow measurement; and each treatment plant 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 the setting 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, 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 is discussed in section 4). 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 closed 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 its 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, dominate 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 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 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 1: Summary of measurement and boundary data streams within the NKWD study area (omitting nodal demand flow boundaries.

Category	Model Element	Number
Level Measure	Tank	10
Flow Measure	Pump Station	6
	PRV	3
	Source (TP)	6
Pressure Measure	Pump Station	10
	PRV	7
Runtime Measure	Pump	14
Status Boundary	Pump	14
	Alt. Valve	2
Setting Boundary	PRV	1
Head Boundary	Reservoir	2
Flow Boundary	–	0
Total	–	75

2.3 SCADA Data Quality

All SCADA data streams were inspected for visually 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 data density. 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.

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 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 surely eliminate many points that otherwise would be labelled "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 timeseries 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 three month period from Oct 1, 2012 through Jan 1, 2013, as a colormapped 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, and the maximum data gap (both computed on a daily basis) – as well as the mean value and inter-quartile range (also computed on a daily basis). The visual data analysis is described in more detail for each of the data categories below. Here 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 3, for the 27 SCADA flow measures listed in Table 2. The integer index in Table 2 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 three month span for one data stream is represented by one row of the image. Thus the visual matrix in Figure 3 has dimension 27×92; one matrix element for each data stream and each day. Moving from top to bottom changes the data streams from index 1 through 27, while moving from left to right 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 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.

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 4-7 as well as 11 are essentially absent from the record and thus were discarded. Data stream 11 is the flow through the regulating valve singled out above that prompted the assumption of zero flow, in order to establish the boundary for DMA 2. There are also significant data gaps for data stream 21, which is the discharge from the high service pumps for the northern TP; this gap, however, is coincident with the 1017 pressure zone being "un-split," when these high service pumps are not expected to be in service. Thus the data exhibit

Table 2: 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 Reg Flow	10
CHES_FI200	Chesapeake Regulator Pit Flow	11
NEW1_FI301	St. Therese Reg 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	TaylorMill 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 Flow	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

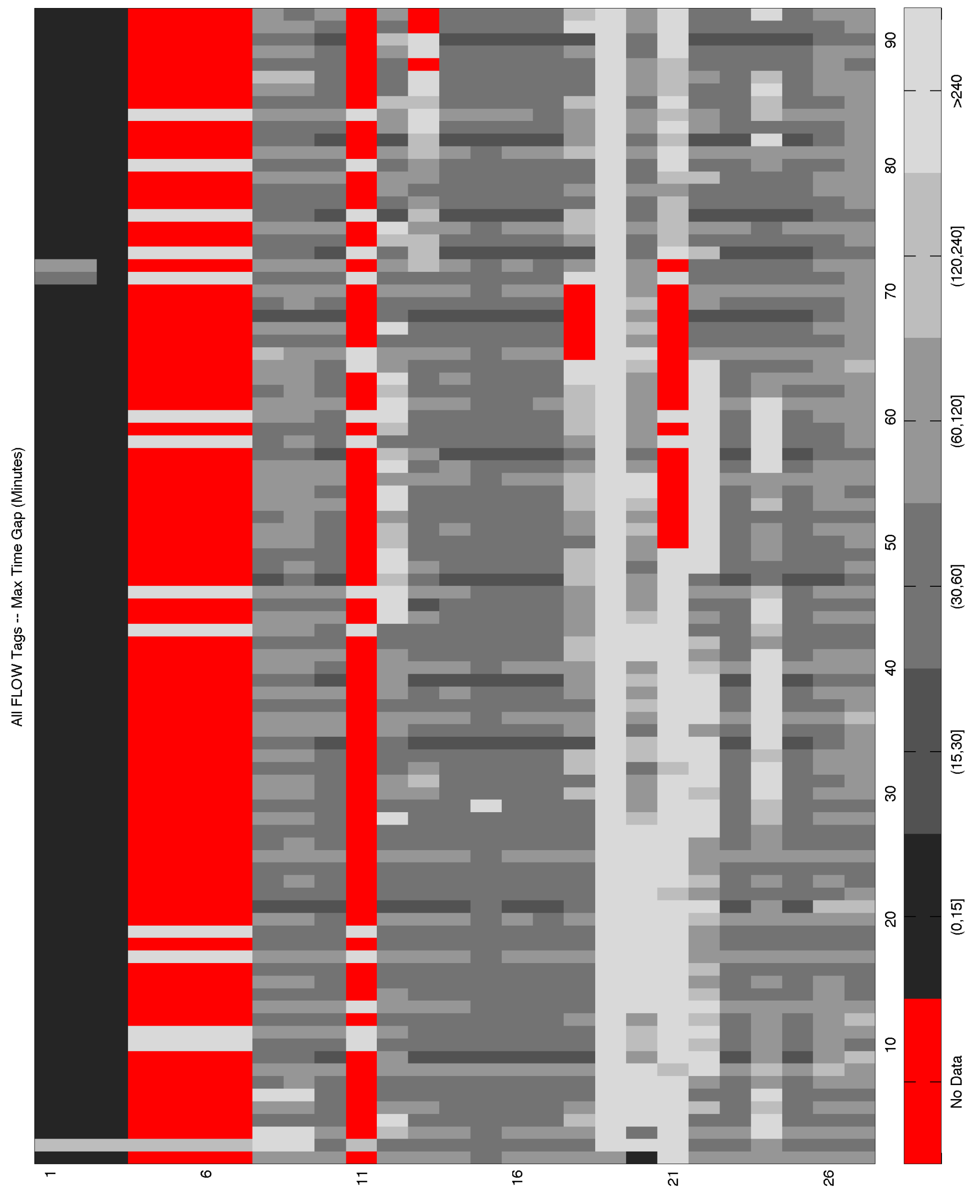

Figure 3: Maximum time gap (minutes) between valid measurements for all flow measure data streams, for each day from Oct 1, 2012 through Jan 1, 2013. Indices refer to the SCADA tags in Table 2.

the characteristics of "delta mode" storage, where new data points are stored, only when a significant change in value occurs. Presumably the pump flow 21 is zero for the entire duration when no data exists. It would be preferable for data quality assurance, however, if the delta-mode SCADA configuration allowed for a minimum data density (say, 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 scheme (e.g., minimum value change that triggers storage of a new point), and how they 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 4 and 5, respectively, for the data stream indices in Tables 3 and 4. 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 5, but 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.

The maximum gap data for all three categories show strong relationships among the individual data streams. For example in Figure 5 there are days when data are written for every pump runtime, presumably independent of pump status or runtime change, while in Figures 3 and 4 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.

3 Real-Time Modeling Using Epanet-RTX

RTX is an object library for building real-time hydraulic modeling environments. It is a set of building blocks, which can be used and extended to create real-time data fusion applications. In essence, RTX provides interoperable access to several different technologies which are foundational to realtime modeling. These technologies involve accessing a SCADA historian database, using filtering, smoothing, and other data transformation methods, and running hydraulic and water quality simulations. 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.

Table 3: 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

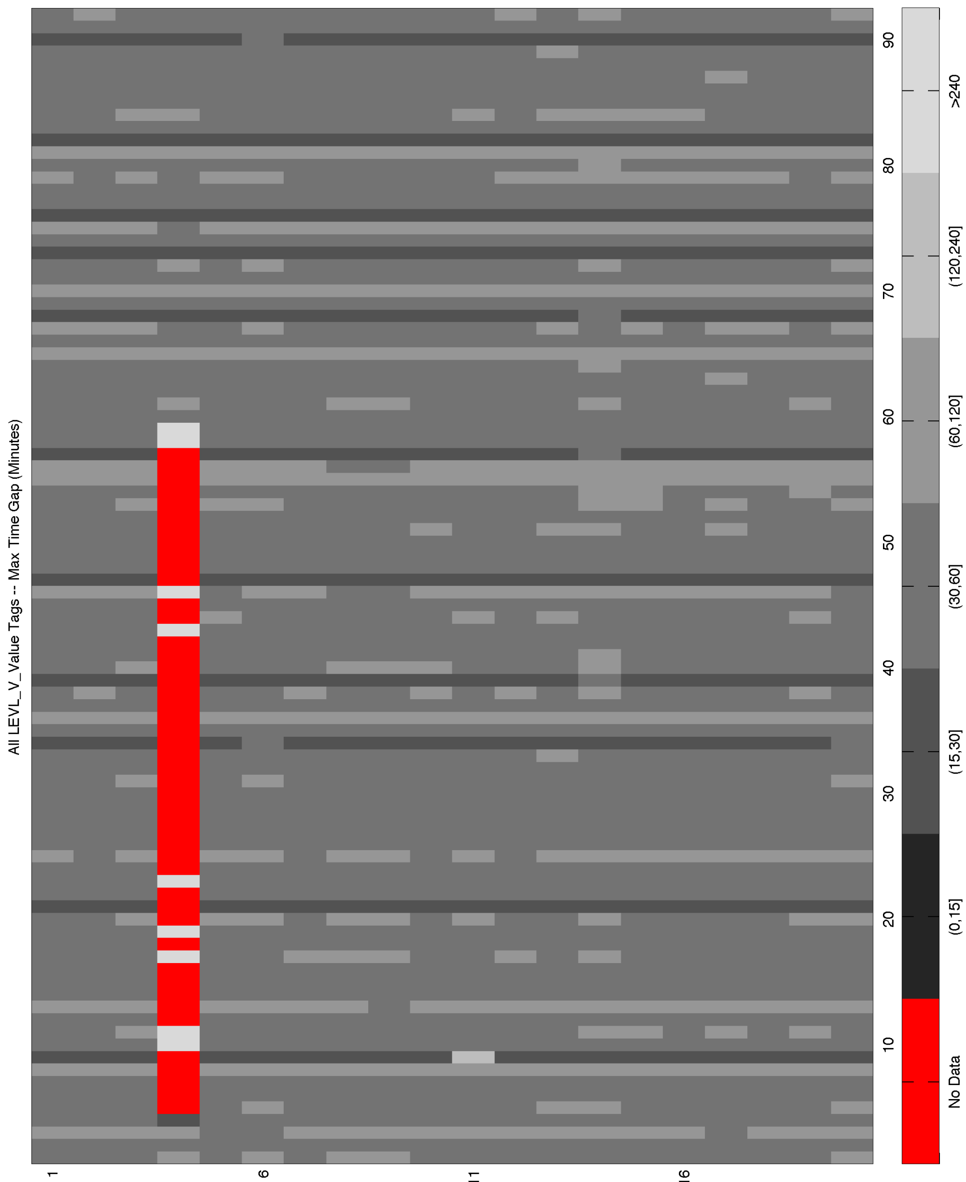

Figure 4: Maximum time gap (minutes) between valid measurements for all tank level measure data streams, for each day from Oct 1, 2012 through Jan 1, 2013. Indices refer to the SCADA tags in Table 3.

Table 4: 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	TaylorMill 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	TaylorMill HS Pump 3 Status	23
DUD1_KQP532NR	Dudley 1040 Pump 2 Status	24
HAND_KQP532NR	Hands Pike Pump 2 Status	25
TMHS_KQP535NR	TaylorMill 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	TaylorMill HS Pump 4 Status	31
BROM_KQP531NR	Bromley Pump 1 Status	32
TMHS_KQP532NR	TaylorMill 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
DUD1_KQP533NR	Dudley 1040 Pump 3 Status	41
TMHS_KQP531NR	TaylorMill HS Pump 1 Status	42
BROM_KQP532NR	Bromley Pump 2 Status	43

Figure 5: Maximum time gap (minutes) between valid measurements for all pump runtime measure data streams, for each day from Oct 1, 2012 through Jan 1, 2013. Indices refer to the SCADA tags in Table 4.

3.1 Real-time simulation process

The real-time modeling results presented here were obtained using a simple RTX client application, similar to that shown in Figure 6. The application uses built-in RTX objects that read the real-time model specification using the libconfig configuration file library (**?**). This one configuration file specifies the SCADA databases and how to access their data records; the timeseries to query in the databases (i.e. the SCADA "tags"), and their properties (e.g., units); the transformations to be applied to each timeseries; and the connections between the transformed timeseries, and the 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 6: Prototype Epanet-RTX client code (C++) for executing real-time simulation using an RTX configuration file.

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

0. Ignore 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 timeseries 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 timeseries pipelines specified in the RTX configuration file (these are described in section 4). The design

of RTX elegantly handles the execution of such transformation pipelines, as each RTX 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. 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.

Here, the above steps were executed for a particular historical time frame (i.e. corresponding to `start` and `dur` in Figure 6). In a true real-time simulation, the RTX client software would periodically wake up from an idle state, perform the above steps 1-4, and then go to sleep for a specified interval. Such a persistent real-time simulation would provides a constantly updated view of system status and model performance.

Also, as a practical matter, the results presented below were not obtained through a live connection to the SCADA historian database (Wonderware SCADA historian based on Microsoft SQL Server). To avoid the need to be on-site (the NKWD SCADA historian database was not connected to the internet), a disk image was created of the SCADA historian server, so that a virtual SCADA historian could be run off-site. The RTX software and client code used was 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 virtual machine was created.

4 Epanet-RTX Real-Time Model Configuration

Real-time model configuration specifies how SCADA data are accessed and transformed into the real-time data streams associated with the measurements and boundaries shown in Figures 1 and 2. Such data transformation sequences are called *timeseries pipelines*. Configuration also specifies the associations between the terminus of each timeseries pipeline, and the model elements that they represent. Finally, the estimation of real-time water usage, and its distribution to the network nodes, is an important and potentially complex configuration task that would be required of any real-time modeling application. RTX includes specialized objects that support the configuration of real-time demand calculations, simplifying the configuration requirements from the application perspective.

Here we describe the RTX data access and transformation processes and decisions that were implemented for the NKWD case study. For completeness we also explain the RTX demand estimation process, although responsibility for those computations is assumed by the RTX DMA objects.

4.1 Epanet-RTX Data transformations

The timeseries pipelines represented below 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 RTX object model was designed with such experimentation in mind, acknowledging its important role 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 timeseries pipeline comprised of RTX objects, and simply request the data points. Such requests are propagated backward through the timeseries pipeline (as needed - some points may be already available from prior requests), and results may be automatically persisted in a database.

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

4.1.1 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 7 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, is 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 prior to using tank level data for DMA demand estimation, which requires converting tank level 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, e.g., **?**, or a practical online introduction by **?**). 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 nth derivative; thus for a first derivative is it often sufficient to use two passes of a moving average (equivalent to a single pass of a triangular weighted moving average - see **?**).

Figure **??** represents the RTX timeseries pipeline implemented for smoothing all tank level data. The pipeline begins with a TimeSeries 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 TimeSeries object are input to the Resampler object (more accurately, the Resampler fetches its points from the TimeSeries 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. The

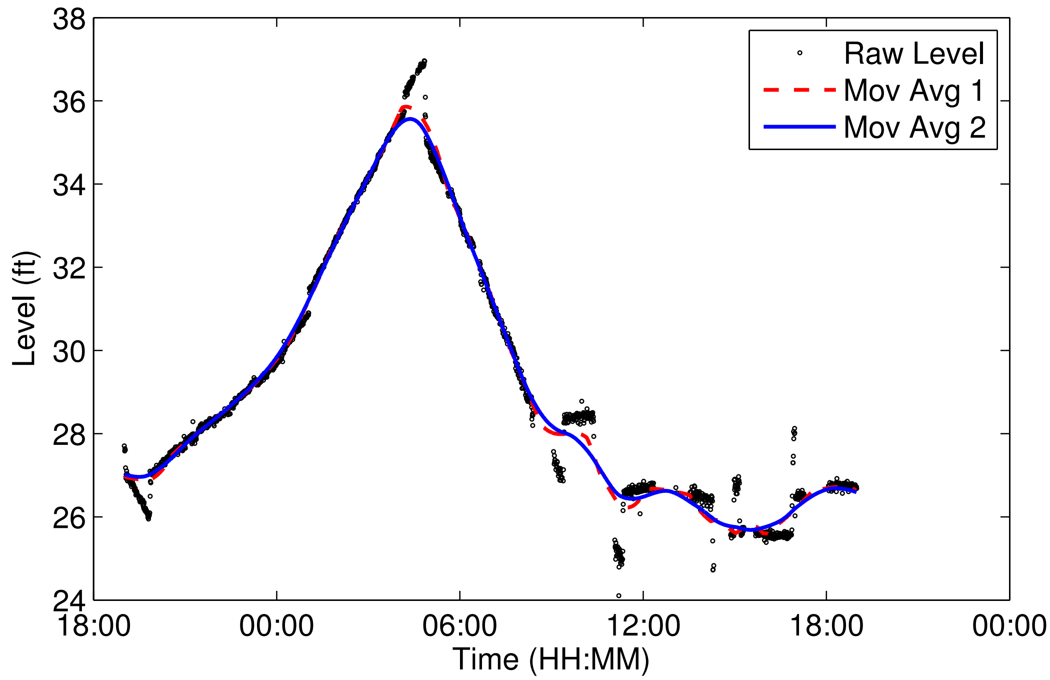

Figure 7: Representative raw storage tank water level data with 1st and 2nd moving average filters. Note signal noise and spikes separating fill and drain cycles.

tank level resampler uses a one minute clock, so interpolated points are produced at one 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 above, is equivalent to a single pass of a 90 minute window triangular weighted moving average filter. The last object in the timeseries pipeline represents the association with a model element – in this case, to the element associated with the SCADA "Tank Level."

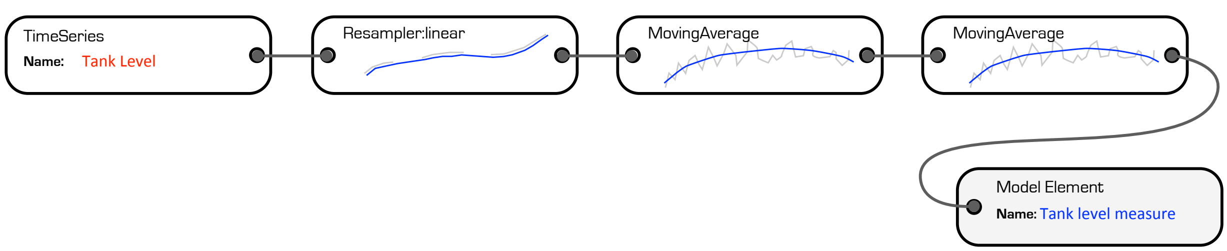

Figure 8: Timeseries pipeline for resampling and smoothing raw SCADA tank level data. Resampler uses 1 minute clock with linear interpolation. Sequential moving average filters each use a 91 point, or 90 minute, window.

Figure 7 shows representative results from the tank level pipeline, including 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.

4.1.2 Pressure data streams

Figure 9 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 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 previously in Figures 3 and 4 there are unexplained relationships between the maximum daily time gaps across different data streams.

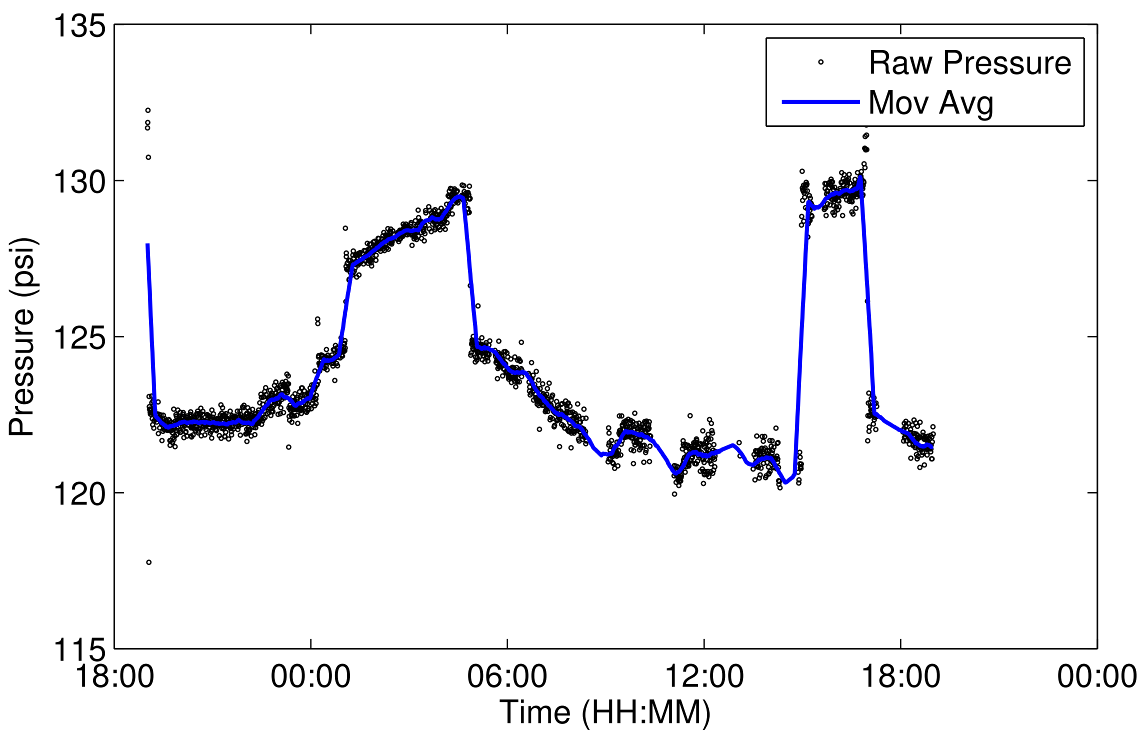

Figure 9: Representative raw pressure data with moving average filter. Note the presence of cycles of intense data polling activity with interspersed data gaps. This behavior is observed in many raw SCADA time series.

The timeseries pipeline used for all pressure data is shown in Figure 10. 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 points, or 24 minutes, wide. Results from applying this timeseries pipeline to the representative pressure data are shown in Figure 9.

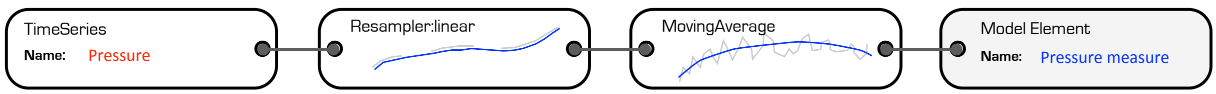

Figure 10: Timeseries pipeline for resampling and smoothing raw SCADA pressure data. Resampler uses 1 minute clock with linear interpolation. Sequential moving average filter uses a 25 point, or 24 minute, window.

4.1.3 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 state[2]. 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 standard RTX objects is represented by the timeseries pipeline in Figure 11. The runtime data stream is first resampled using a 1 minute clock and assigned as the source to a FirstDerivative 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 source to a Threshold object, which compares its source 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 timeseries pipeline are shown in Figure 12. The left figure shows four 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 timeseries pipeline in Figure 12, 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 timeseries pipeline in Figure 12 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 12 show large errors in cumulative runtime when

[2]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.

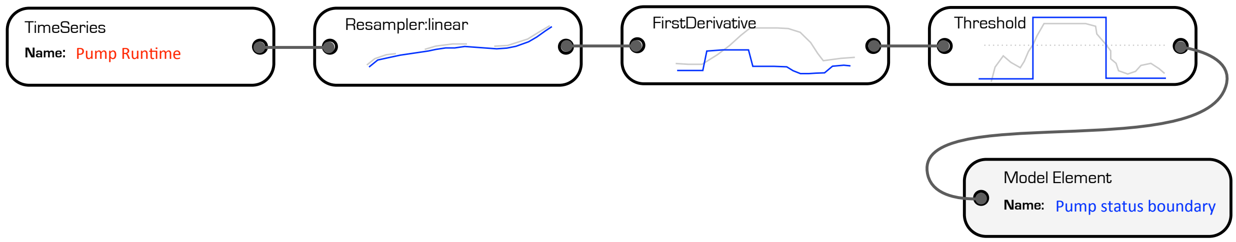

Figure 11: Timeseries pipeline using derivative and threshold to derive binary pump status from raw SCADA pump runtime data.

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).

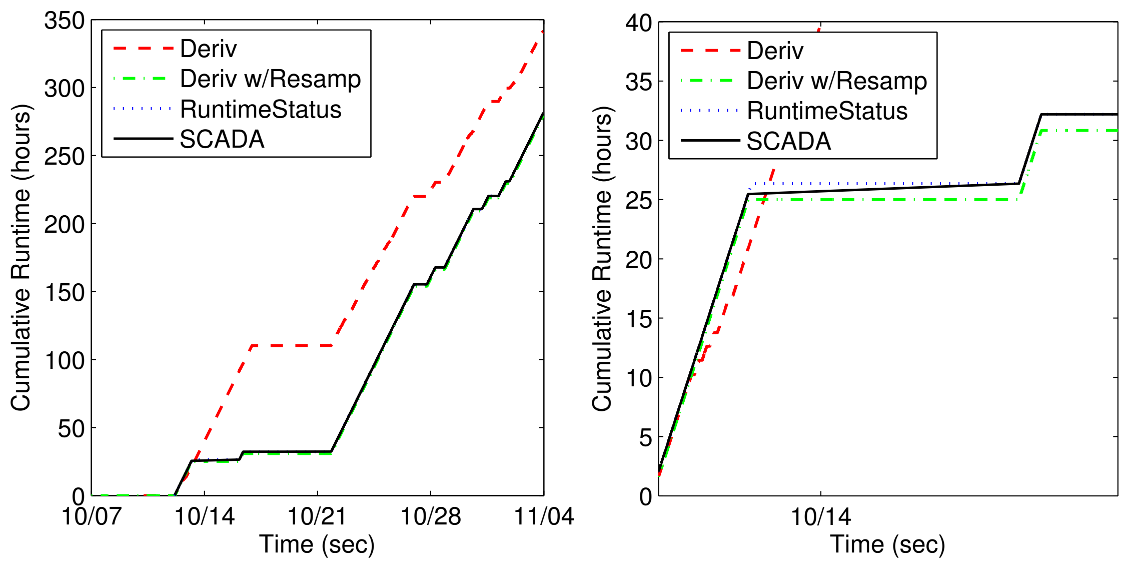

Figure 12: Representative cumulative pump runtime calculated from raw SCADA pump runtime data, as well as three pump status data streams derived from runtime data. Right figure shows detail around two separate pump status changes, illustrating pump status errors introduced when differentiating runtime data, which are resolved by the 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 12 illuminates detail over several days surrounding two pump cycles. While the SCADA data *should* yield a cumulative runtime with a slope of either 0 or 1[3], the results show a time span exceeding two 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 one 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 two day time gap – a much greater error than the one hour of lost runtime. Such errors originating in large time gaps are common enough in this SCADA system, that 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 RTX class named RuntimeStatus that processes runtime clock data and accurately detects the status changes; the new timeseries pipeline for pump status, which was used for all the real-time modeling results, is shown in Figure 13.

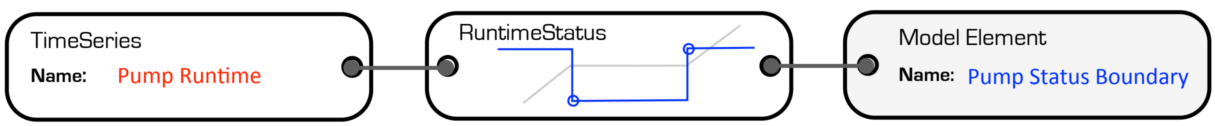

Figure 13: Timeseries pipeline using special-purpose RuntimeStatus class to derive binary pump status directly from raw SCADA pump runtime data.

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 RTX derivative object that is limited by its local perspective. The data in Figure 12 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[4] This behavior is illustrated in Figure 12, 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]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.

[4]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.

4.1.4 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 passed to two sequential MovingAverage objects, each with a 91 point averaging window. The two moving averages used for tank level data was 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 effect the mean values.

For most pump station flows, additional processing of the data streams was performed prior to moving average smoothing, to force the flow to zero when all pumps were off[5]. The motivation for this additional processing was the presence of significant and regular time gaps between points in the flow data streams. Figure **??** 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 timeseries 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, but in real time the data processing must be automatic and robust. The core idea is to 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 transformation pipeline that accomplishes this is shown in Figure 15. While this pipeline appears significantly more complex than those examined previously, each transformation component is represented by an existing RTX object, which shoulder all of the data processing work. The upper portion of the timeseries 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 by the usual objects, producing

[5] A zero flow assumption is valid when the flow sensor does not measure station bypass flow – true for all by two NKWD pump stations

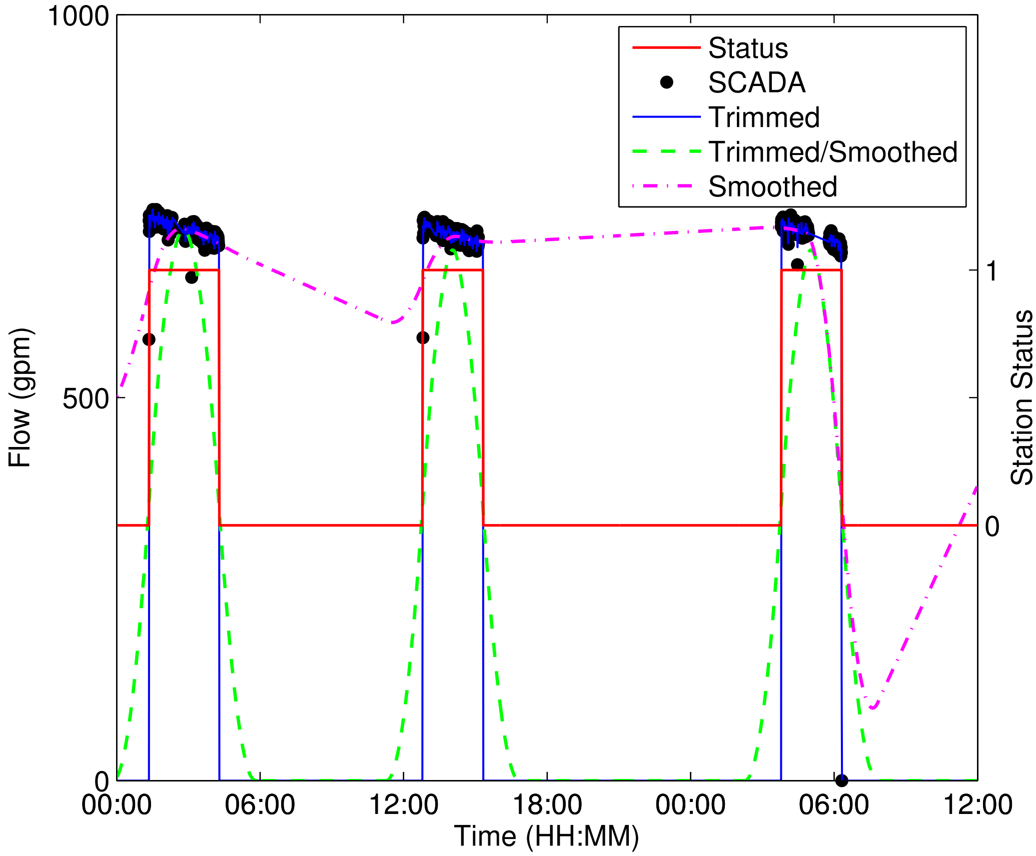

Figure 14: Raw pump station flow SCADA data and trimmed/smoothed station flow (left axis), along with station status (right axis) used to produce the trimmed data stream. Between pump on cycles, flow data gaps make it difficult to use simpler interpolation methods, which could lead to significant non-zero flows when all pumps are off.

the trimmed and smoothed flow measure. Figure 14 shows the Trimmed/Smoothed flow measure, which better represents the true station flow.

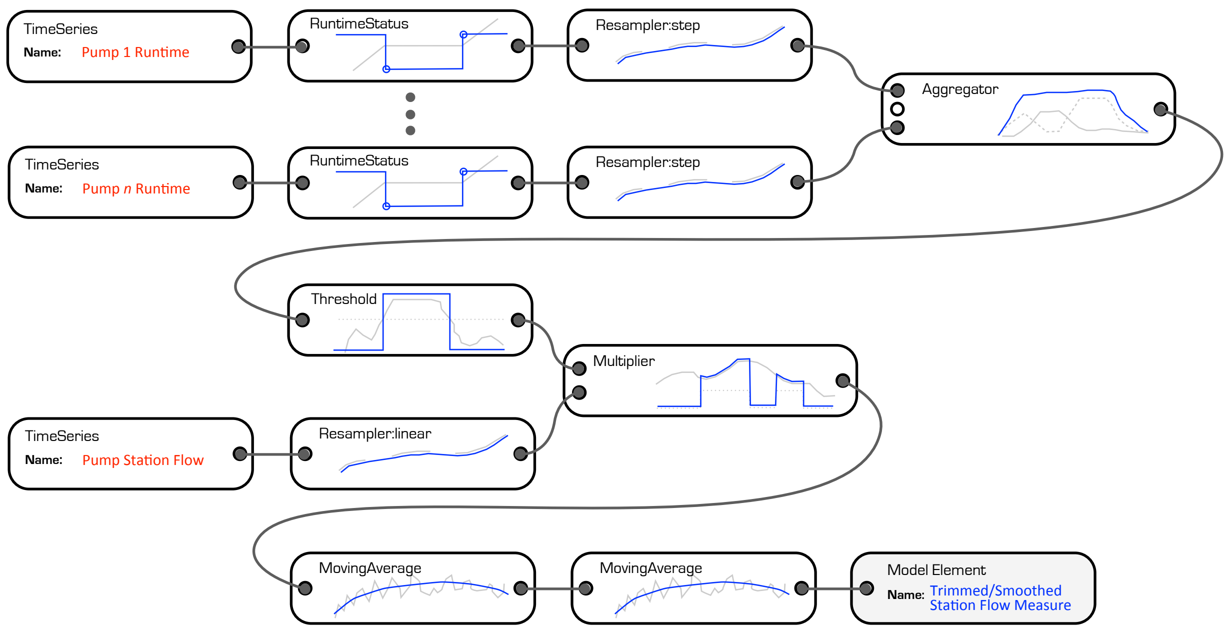

Figure 15: Timeseries pipeline for trimmed and smoothed pump station flow measures. Trimming eliminates flow out of the pump station when all pumps are off; the resampled flow data stream is multiplied by the pump station status (the output data stream from the threshold transformation).

4.1.5 Altitude valve status 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)[6]. Consider, for example, the SCADA tank level data in Figure 16. These data show extended periods during which the tank level is either 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 timeseries pipeline is shown in Figure 17; this pipeline is similar to that used to calculate pump status with a FirstDerivative object, but this time we filter the data series first, as was done for the tank levels. The resampler has a clock of 1

[6]Currently, the hydraulic model does not include such bypass piping; it could be argued that it should.

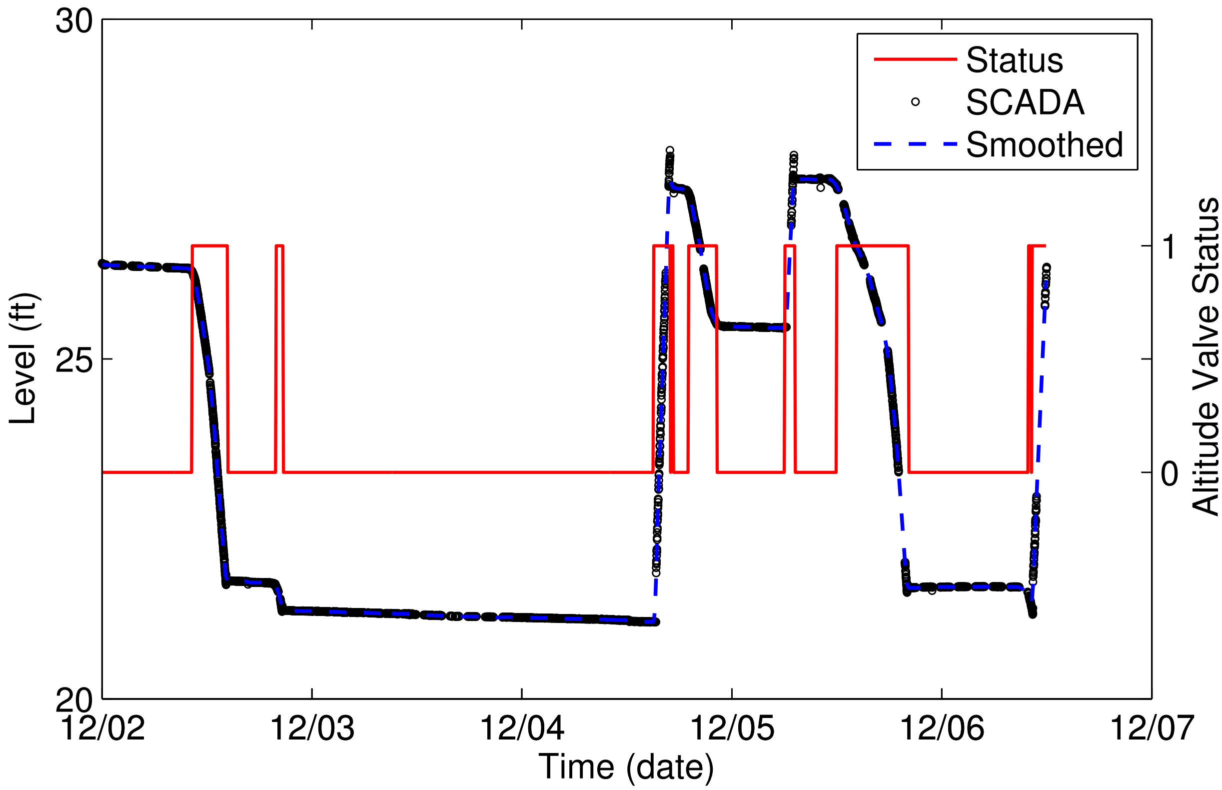

Figure 16: Raw and smoothed tank level data (left axis), along with altitude valve status (right axis) used to model control action of altitude valves on specific tank inlet/outlet lines.

minute, the MovingAverage objects each use a window size of 19, and the Threshold object sets the altitude valve status to closed if the rate of change in tank level drops below 0.2 ft/hr. Representative results from this timeseries pipeline are shown in Figure 16 (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.

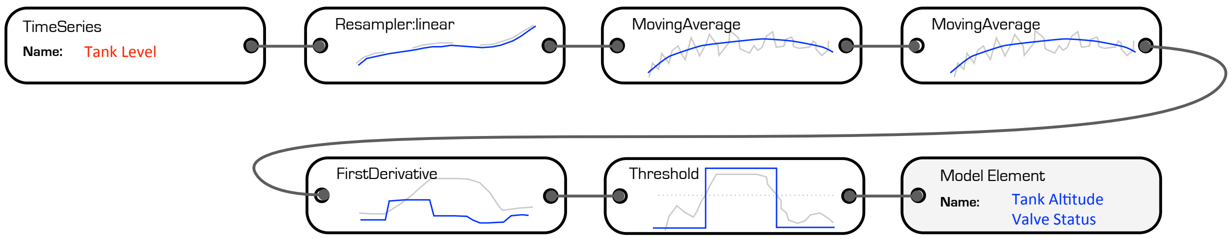

Figure 17: Timeseries pipeline to determine status of tank inlet/outlet pipe with altitude control valve.

4.1.6 Reconstruction of missing plant production data stream

The north treatment plant 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 recreate that flow from other available data sources. Figure 18 is a schematic of the essential north treatment plant infrastructure[7]. The missing flow 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).

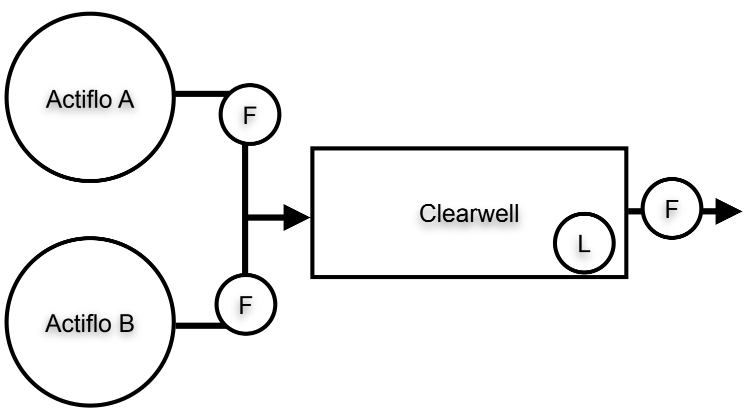

Figure 18: Treatment plant clearwell schematic showing flow and level measures used for construction of boundary flow data stream.

[7]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.

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,

$$\frac{dV}{dt} = F_a + F_b - F, \tag{1}$$

or,

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

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 RTX timeseries pipeline that implements this strategy is shown in Figure 19. 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 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 Eq. 2 to produce the estimate of supply flow leaving the clearwell.

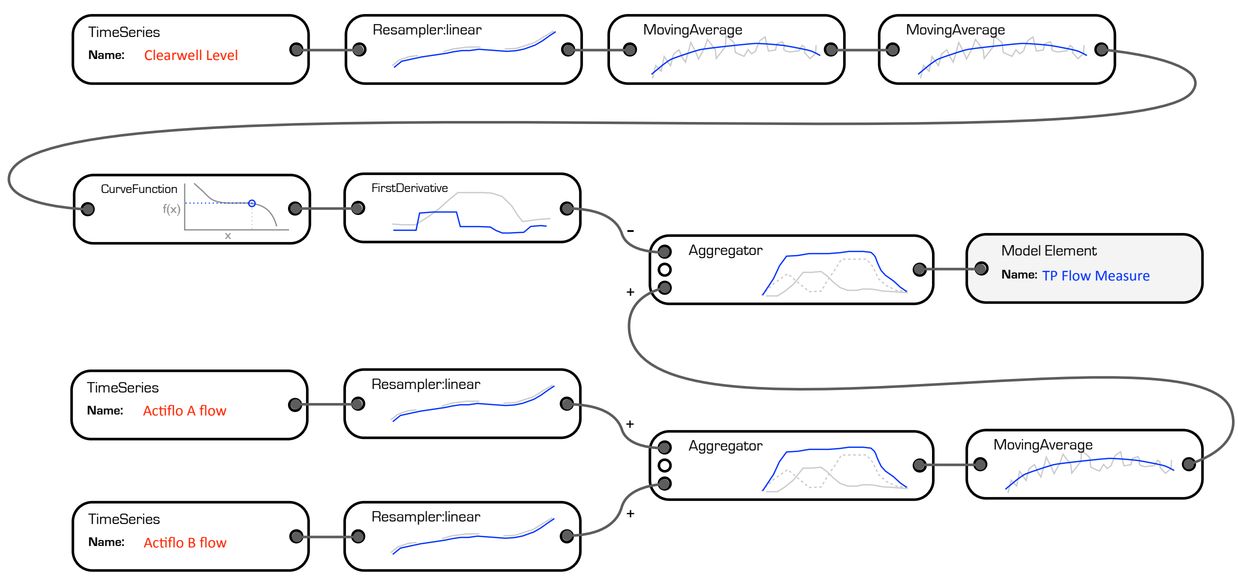

Figure 19: Timeseries pipeline to determine flow measure at north treatment plant, from conservation of fluid volume within clearwell.

Figure 20 shows representative results from the above timeseries 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 axes.

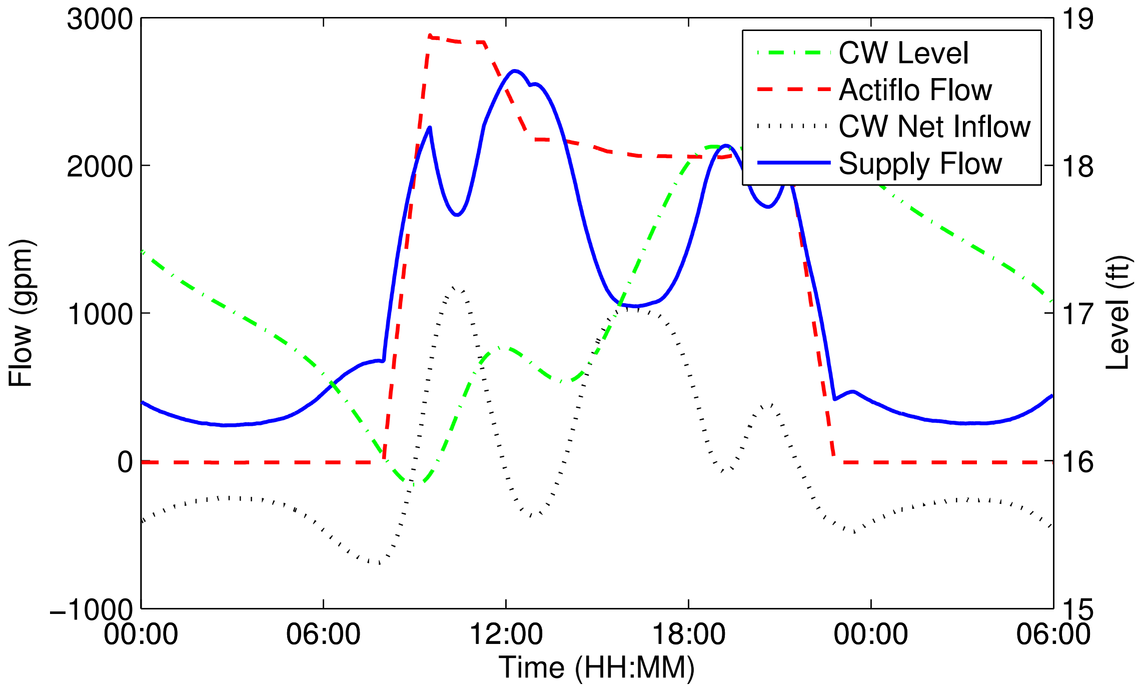

Figure 20: Components of clearwell flow balance and clearwell level at north treatment plant, from timeseries pipeline in Figure 19.

4.2 District metered area real-time demands

The district metered area, or DMA, is a demand management concept introduced in the UK in the early 1980s. UK Report 26 (**?**) defined a DMA as an area of a distribution system which 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 elsewhere, 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 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 comprised 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. Cities like Dublin, serving 1.5 million customers with a supply rate of 143 MGD, have developed an infrastructure monitoring strategy that relies on DMAs consisting of between 1000 and 2000 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 presumably have a positive impact on the accuracy of real-time demands and and model predictions.

It is possible to confuse the DMA with a pressure zone; their boundaries may often be similar, simply because pressure zones 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 1, a single DMA can contain multiple pressure zones (and vice-versa).

4.2.1 DMA demand timeseries 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 its closed and measured boundary pipes.

As an illustration of DMA construction, the boundary elements describing DMA 3 are shown in Table 5. 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[8], the associated flow measure data stream, a multiplier for the DMA demand aggregation, and a brief description. 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 5: Summary of boundary elements for DMA 3 demand timeseries aggregation.

Model Element	Flow Measure Time Series	Mult.	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

The data for DMA 3 in Table 5 can be expressed more usefully as a DMA demand timeseries pipeline, through a flow balance on the DMA. While the detailed boundary elements

[8]Included only to emphasize the information that is collected to describe the DMA, through the process of graph search.

will vary from one DMA to the next, the template for the demand timeseries pipeline does not, and so can be automated for any network. The pipeline for DMA 3 is shown in Figure 21. 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 timeseries pipeline that is responsible for generating it – but for clarity most of these details are omitted from Figure 21. What is shown, however, is the piece of the timeseries pipeline that converts tank level to net inflow, using the CurveFunction and FirstDerivative objects, prior to aggregating the boundary flows.

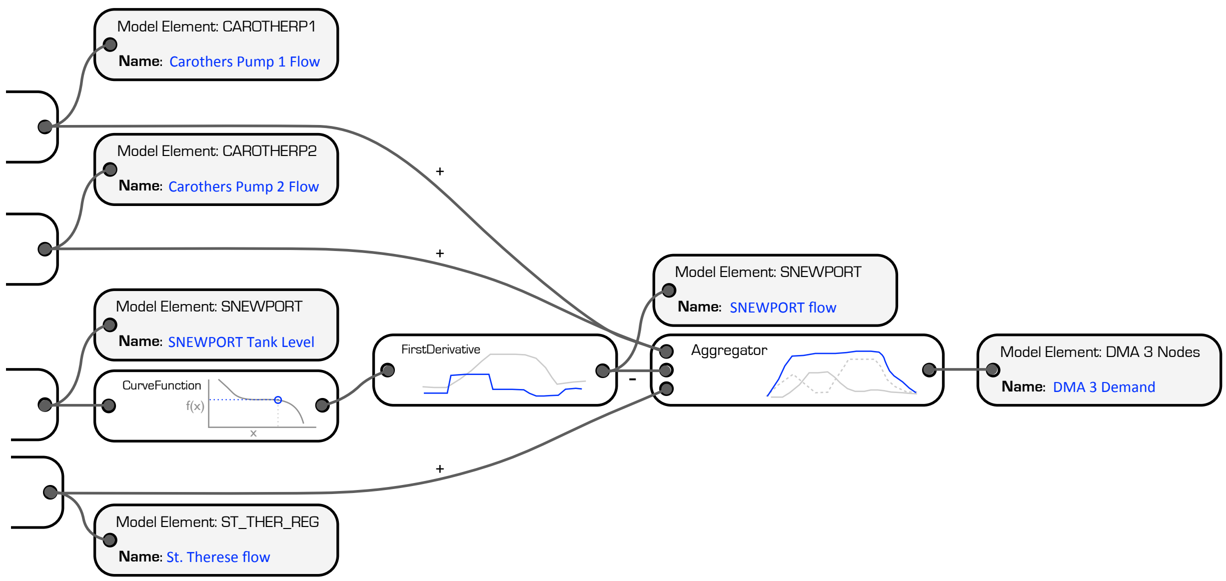

Figure 21: Timeseries pipeline constructed automatically by Epanet-RTX to aggregate boundary flows (DMA 3 demand). Timeseries pipelines for boundary model elements (not shown) are specified as part of real-time model configuration.

The general process of producing real-time DMA demands, including the identification of DMAs and their boundary elements, and the construction of the demand timeseries pipelines for each DMA (e.g., the objects in Figure 21 for DMA 3), are automated by Epanet-RTX algorithms. Figure 22 shows representative calculated demands for all three DMAs; these demands drive the real-time simulation results for the one week period examined in section 6.

4.2.2 DMA demand disaggregation

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

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

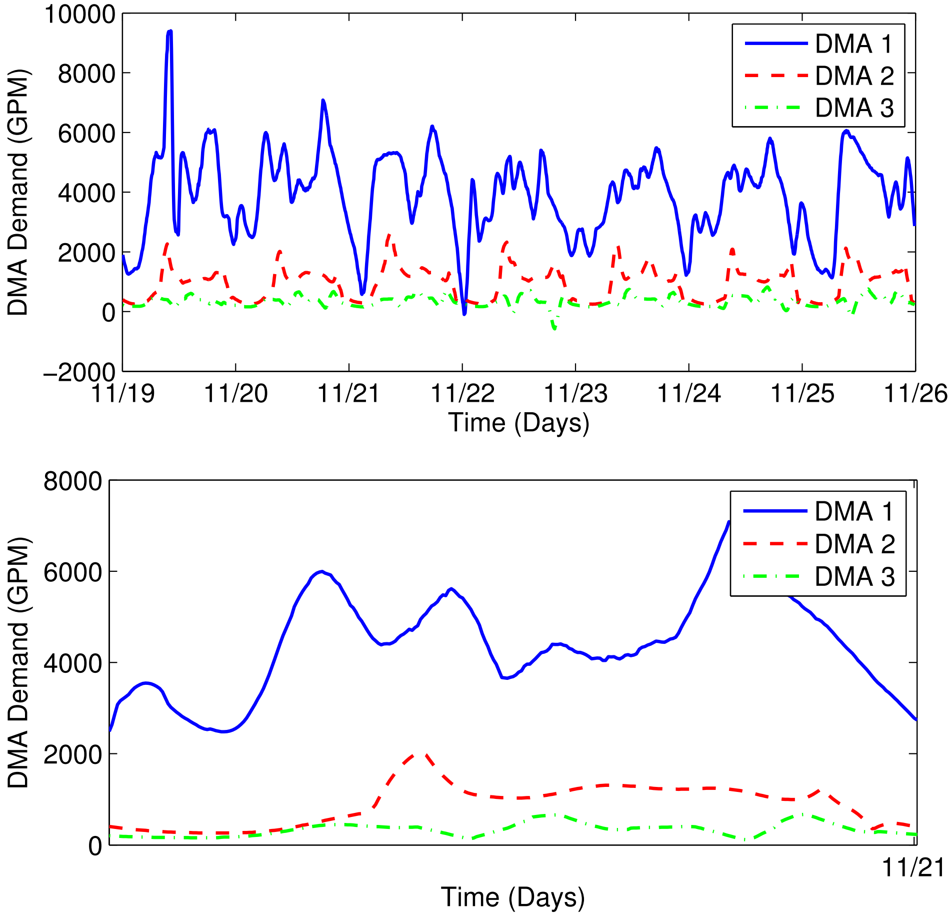

Figure 22: Real-time aggregate demand for DMAs 1-3 for November 19-26, 2012 (top), and expanded for a one day period within the same time frame (bottom).

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 kth 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,

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

$$d_j(t) = b_j(t), \; j \in B_i, \tag{5}$$

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 timeseries pipeline, and $b_j(t)$ is the measured boundary flow from the timeseries 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.

5 Real-Time Model Calibration

This section provides a broad summary of modifications that were made to the NKWD 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, as well as a summary of recommendations and open issues, is provided in Appendix A.

5.1 Calibration process

Many processes that would typically be considered part of network model calibration are implemented automatically by the 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 section 5.

The calibration process described here is consistent with macro-calibration, as it is known in the field (**?**). 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 – i.e. we did not seek to optimize, say, pipe roughnesses, 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 are enumerated below.

1. Updates to PRV settings and elevations based on field measurements.
2. Updates to tank geometry and elevations, based on utility 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 A catalogs the actual 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 source 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 at this stage. 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 RTX). The two computations were expected to be identical – and were observed to be – as RTX is setting the real-time demands based on the SCADA data, and the hydraulic simulations must balance the flows within each DMA.

6 Real-Time Simulation Results and Discussion

6.1 Results

A real-time extended period simulation model was run in a continuous retrospective mode for a one week evaluation period, from midnight, Nov 19, 2012 through midnight Nov 26, 2012[9]. 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 one-week period in 2012. No special data processing was performed, beyond the data transformations described above 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 extended period hydraulic solution. All results were obtained

[9]This one week period includes the Thanksgiving holiday in the United States, Thursday, November 22.

using an Epanet-RTX client application, and a real-time configuration as specified in an RTX configuration file; after that, the simulation was driven automatically by Epanet-RTX.

The data used for evaluation are those available in the SCADA record for the study area over the one week evaluation period, as presented in section 2. 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 23 summarizes the quality of real-time simulation results, for all individual measurements, using the 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 5 do not have zero correlation coefficients, as would seem to be indicated by the Figure; 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.

Perhaps the most 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 24-28; for the 8 flow data streams, in Figures 29-31; and for the 10 tank levels, in Figures 32-35. In each Figure the data points are the red circles, the RTX simulated values 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 one week period; thus it is important to note the scale when comparing results across the different measurements.

6.2 Discussion

In general, it can be said that the real-time simulation results faithfully reproduce the hydraulic behavior of the distribution system, as described by this set of SCADA measurements. That is not to say 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 model at different times of the year, or in different operational modes. Nevertheless, the real-time simulation results are encouraging. Given that the data processing and hydraulic simulation is automated, and no special data processing was done for this particular time period, the results anticipate similar levels of accuracy being available for other time frames – whenever they might be needed for productive purposes. Such is the promise of a real-time model.

The results clearly show some areas where improvements are needed, and we have outlined some fruitful macro-calibration issues to be looked at further in Appendix A. Some of the tanks have significant errors in the mean values, and two tanks – the Campbell County and

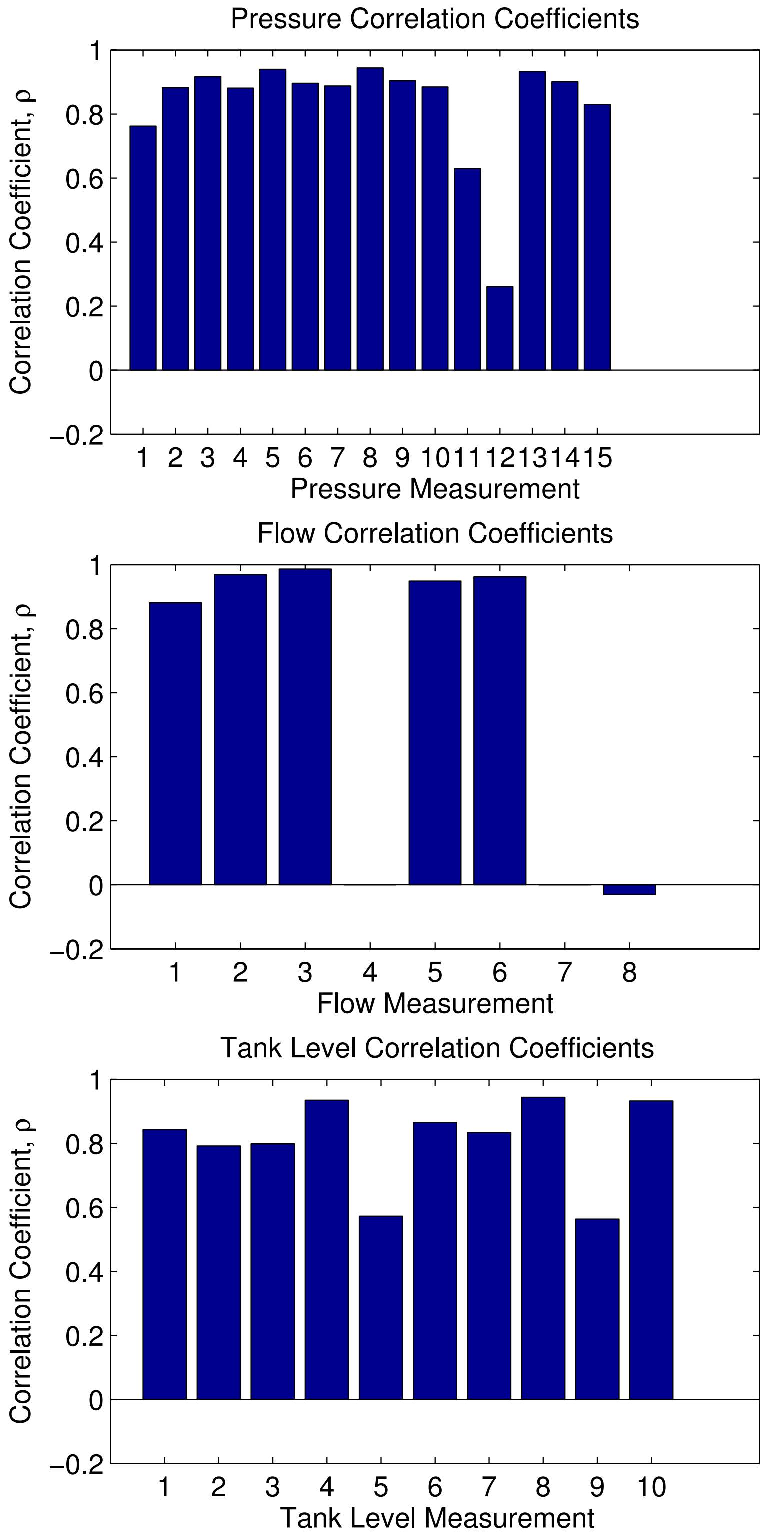

Figure 23: Pearson's correlation coefficients between measured and real-time simulated heads, flows, and tank levels.

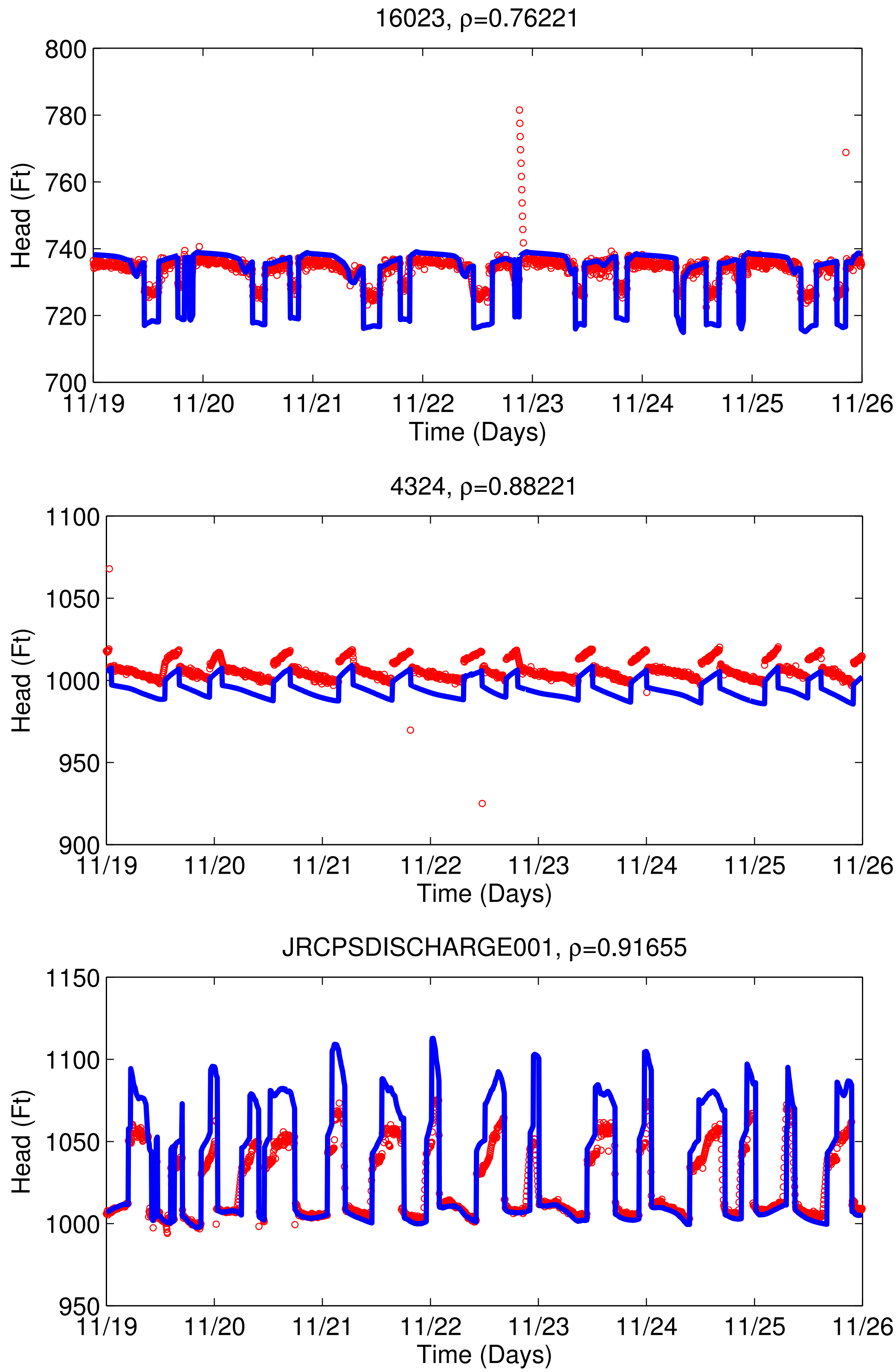

Figure 24: Measured and real-time model simulated heads.

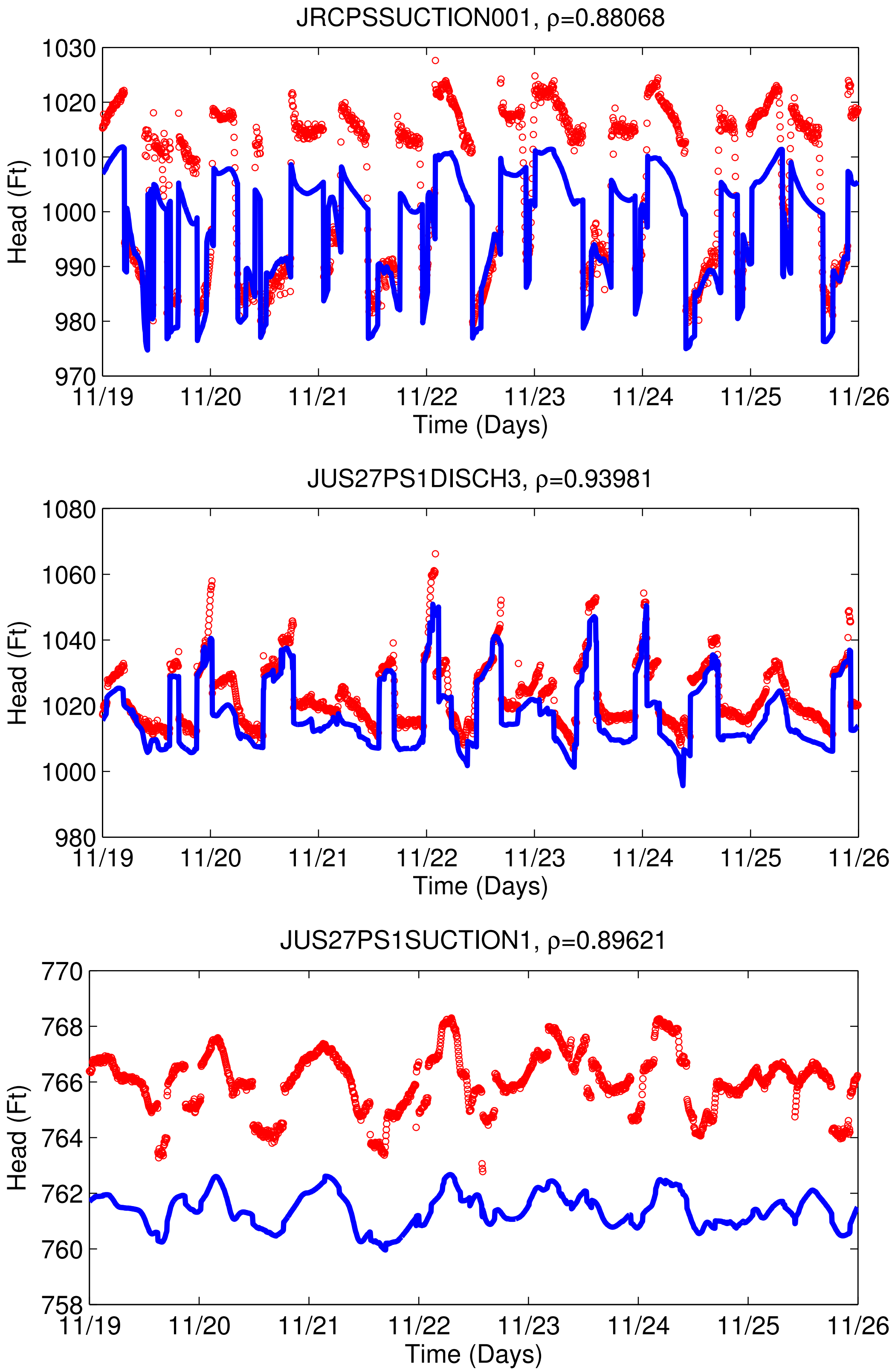

Figure 25: Measured and real-time model simulated heads.

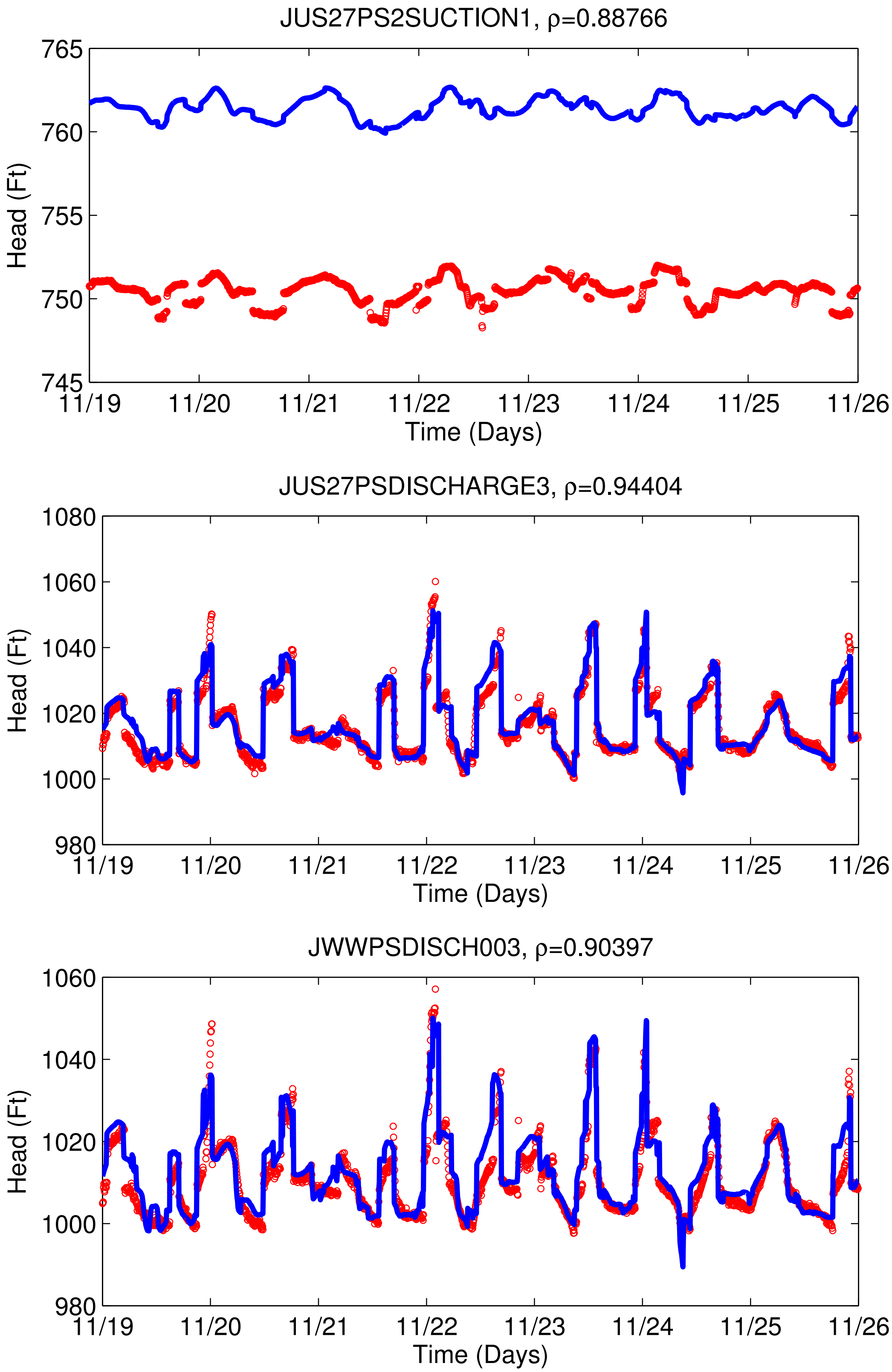

Figure 26: Measured and real-time model simulated heads.

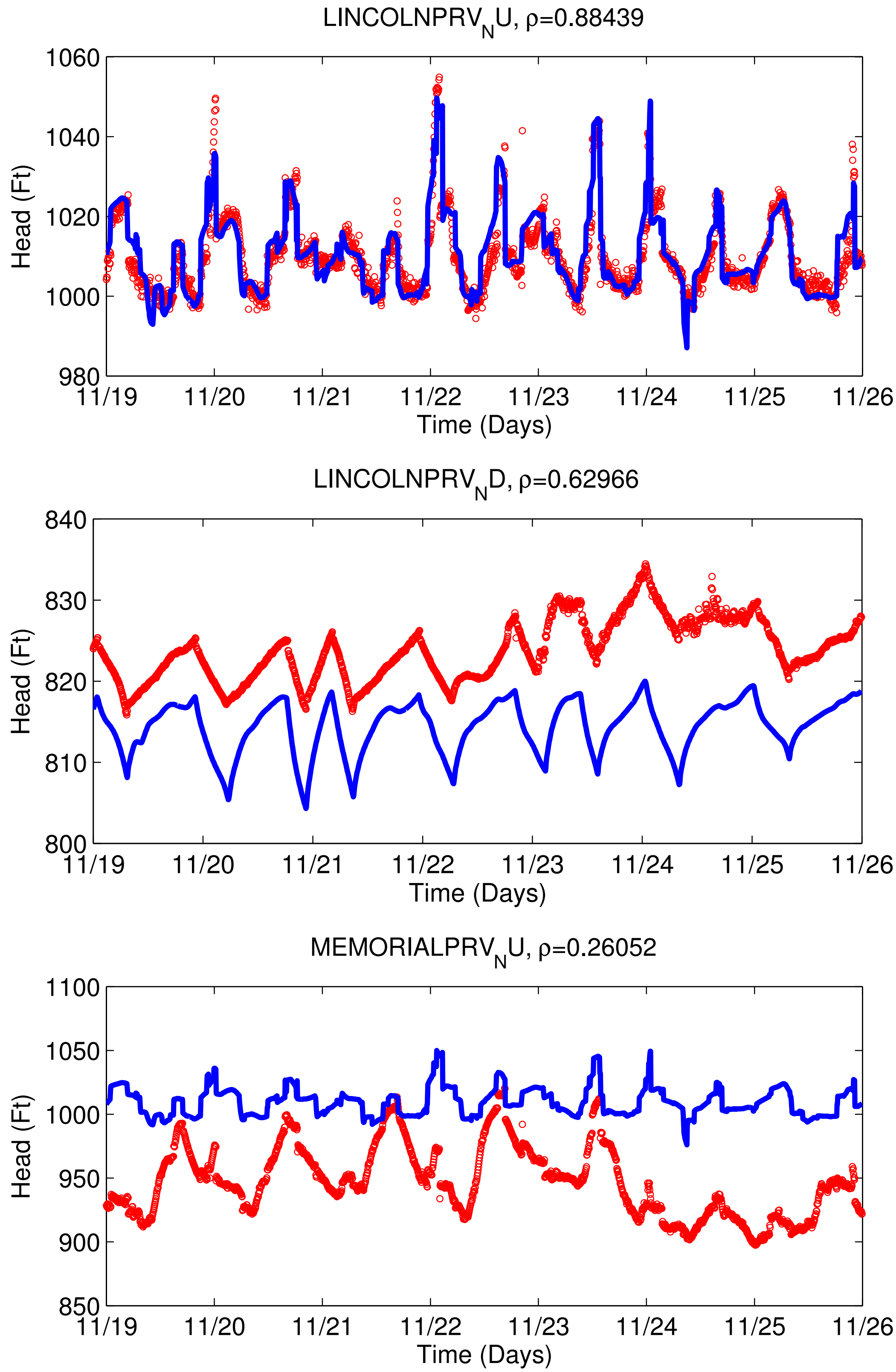

Figure 27: Measured and real-time model simulated heads.

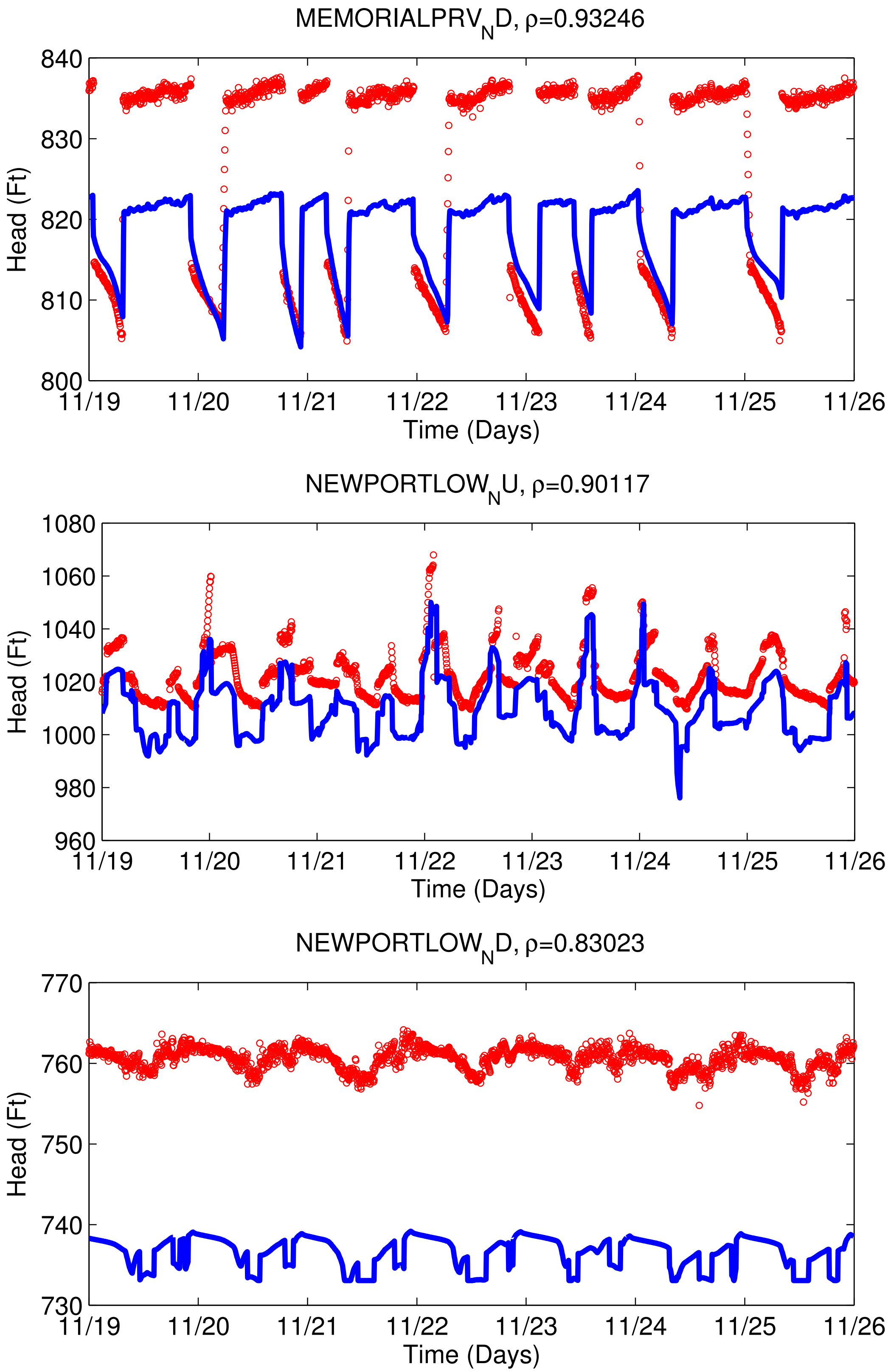

Figure 28: Measured and real-time model simulated heads.

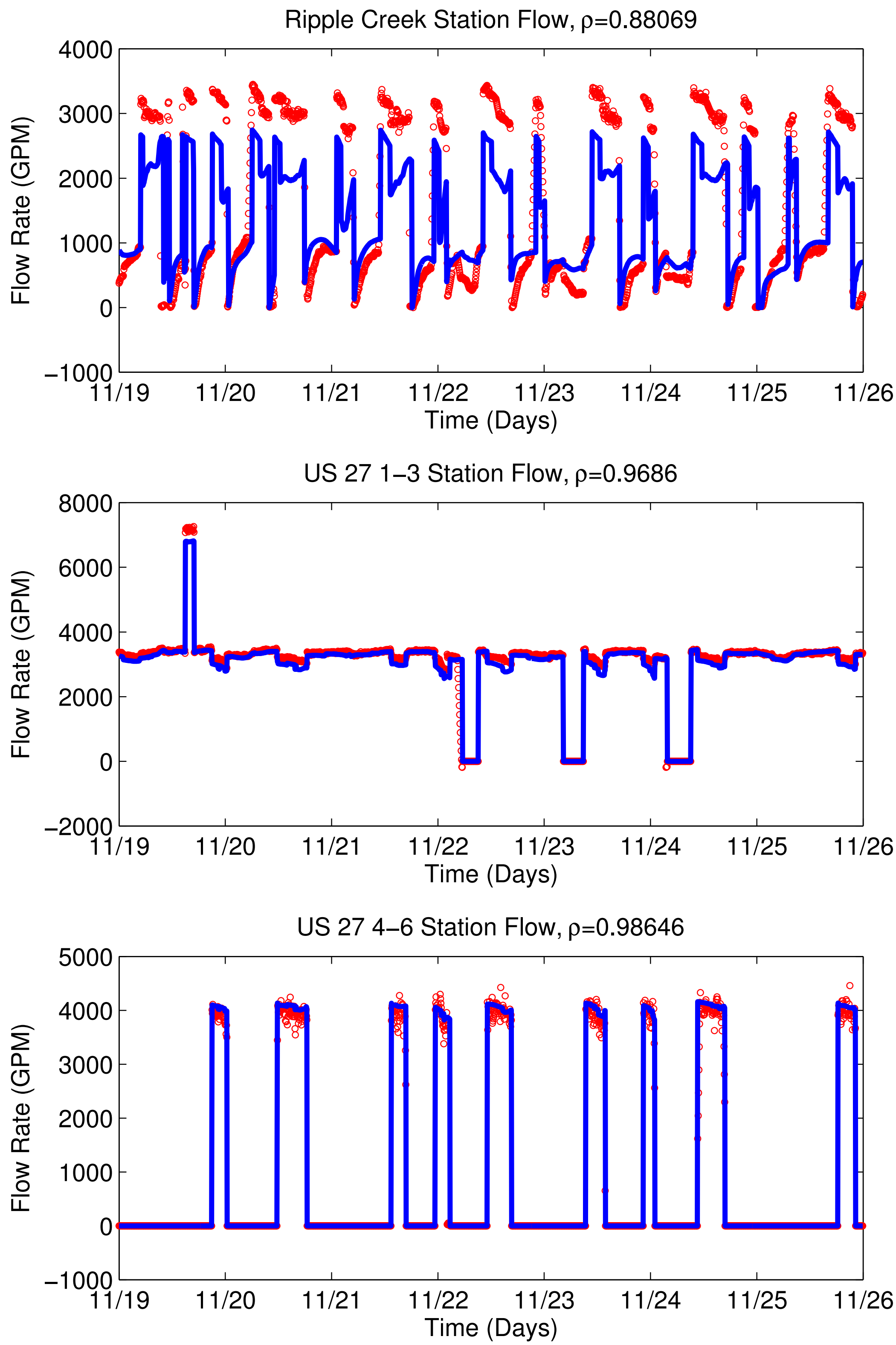

Figure 29: Measured and real-time model simulated flows.

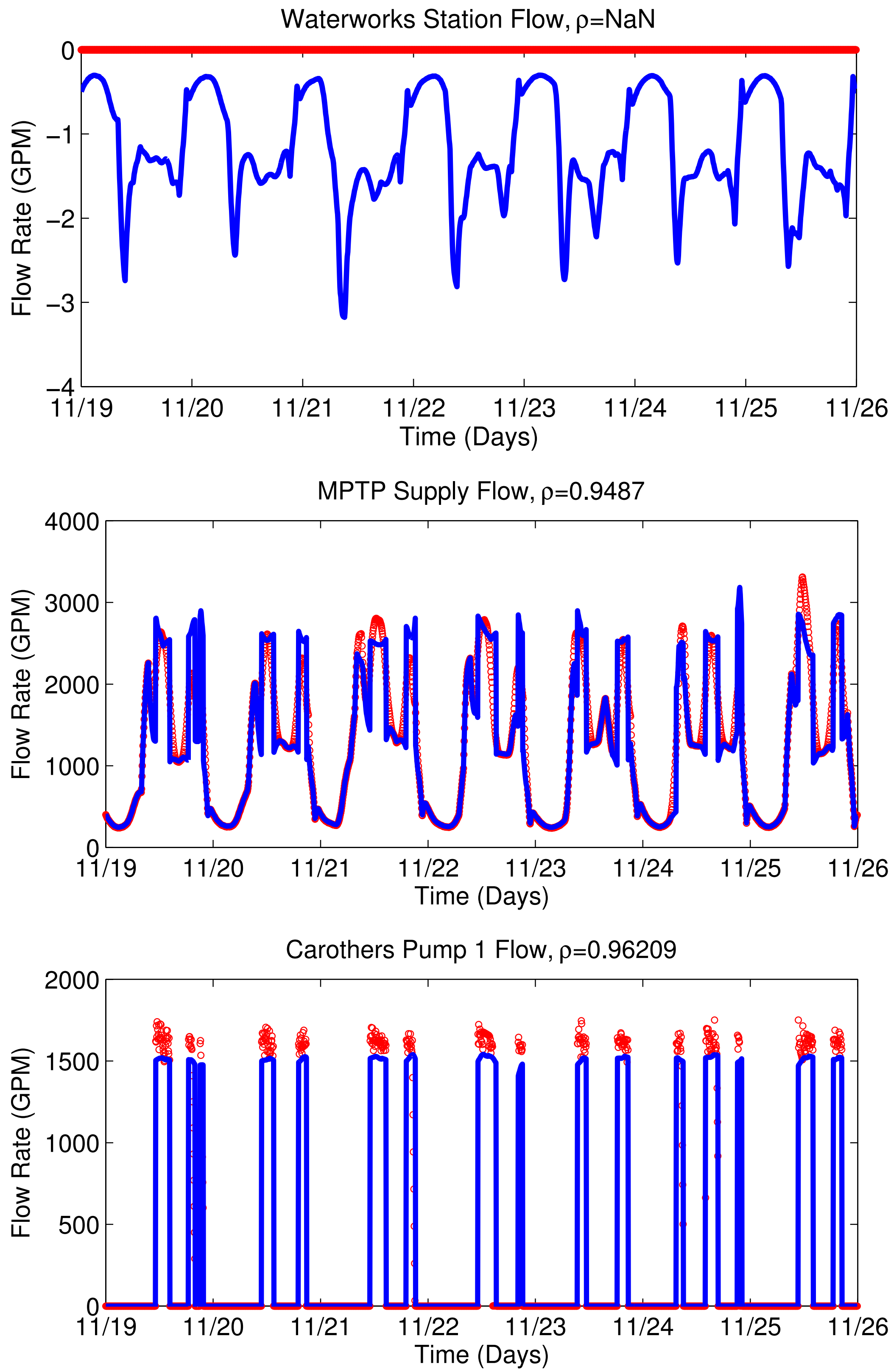

Figure 30: Measured and real-time model simulated flows.

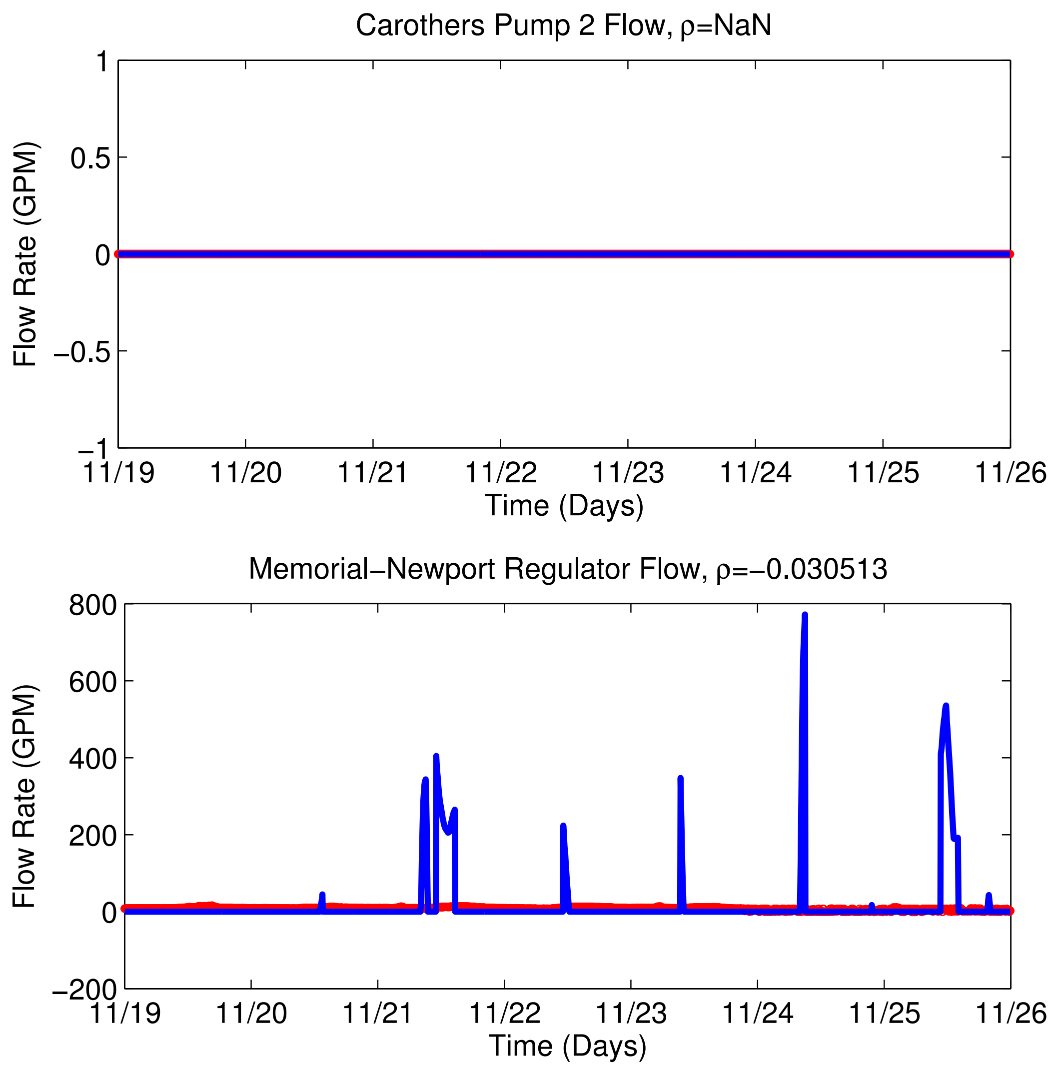

Figure 31: Measured and real-time model simulated flows.

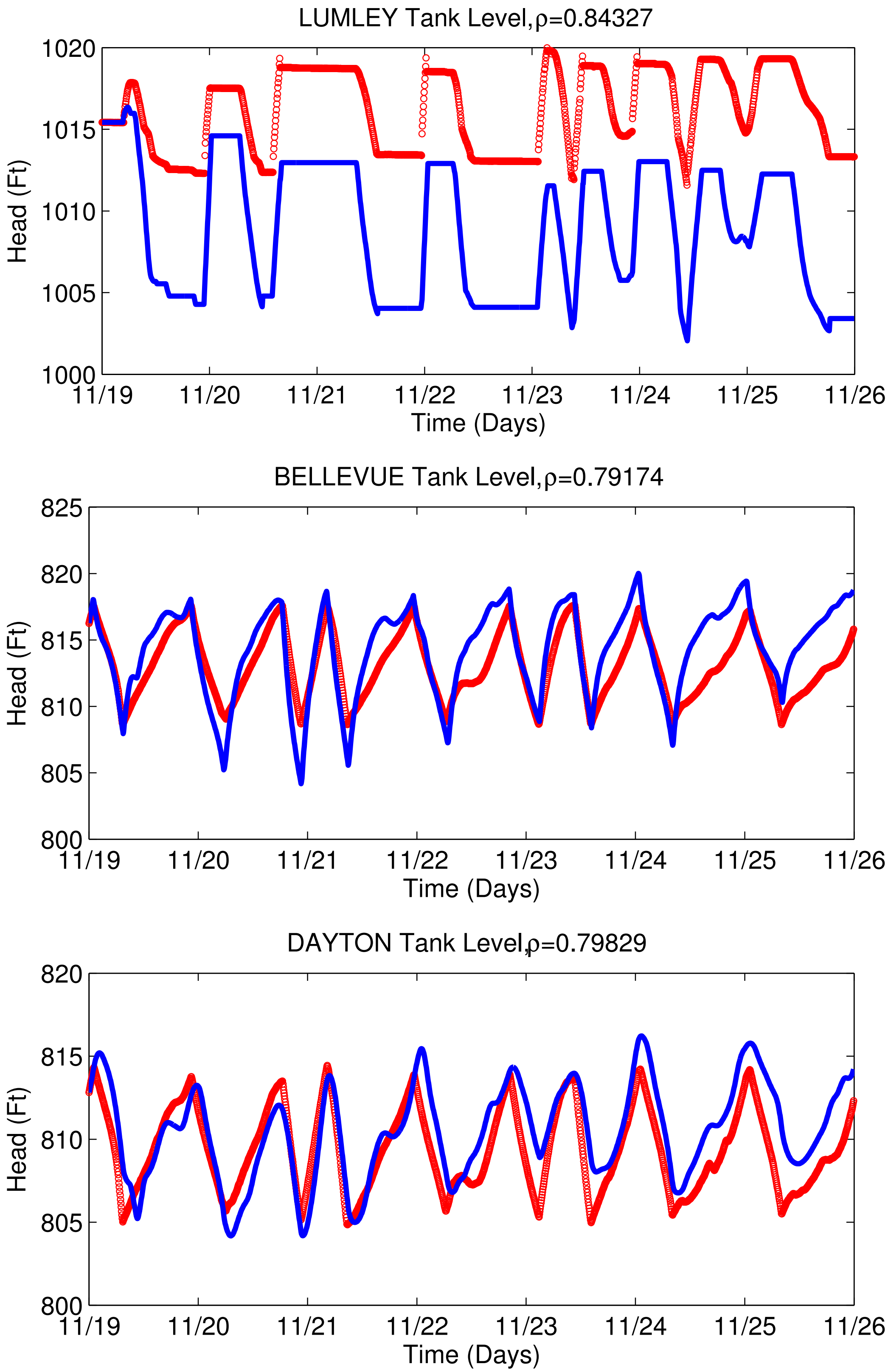

Figure 32: Measured and real-time model simulated tank levels.

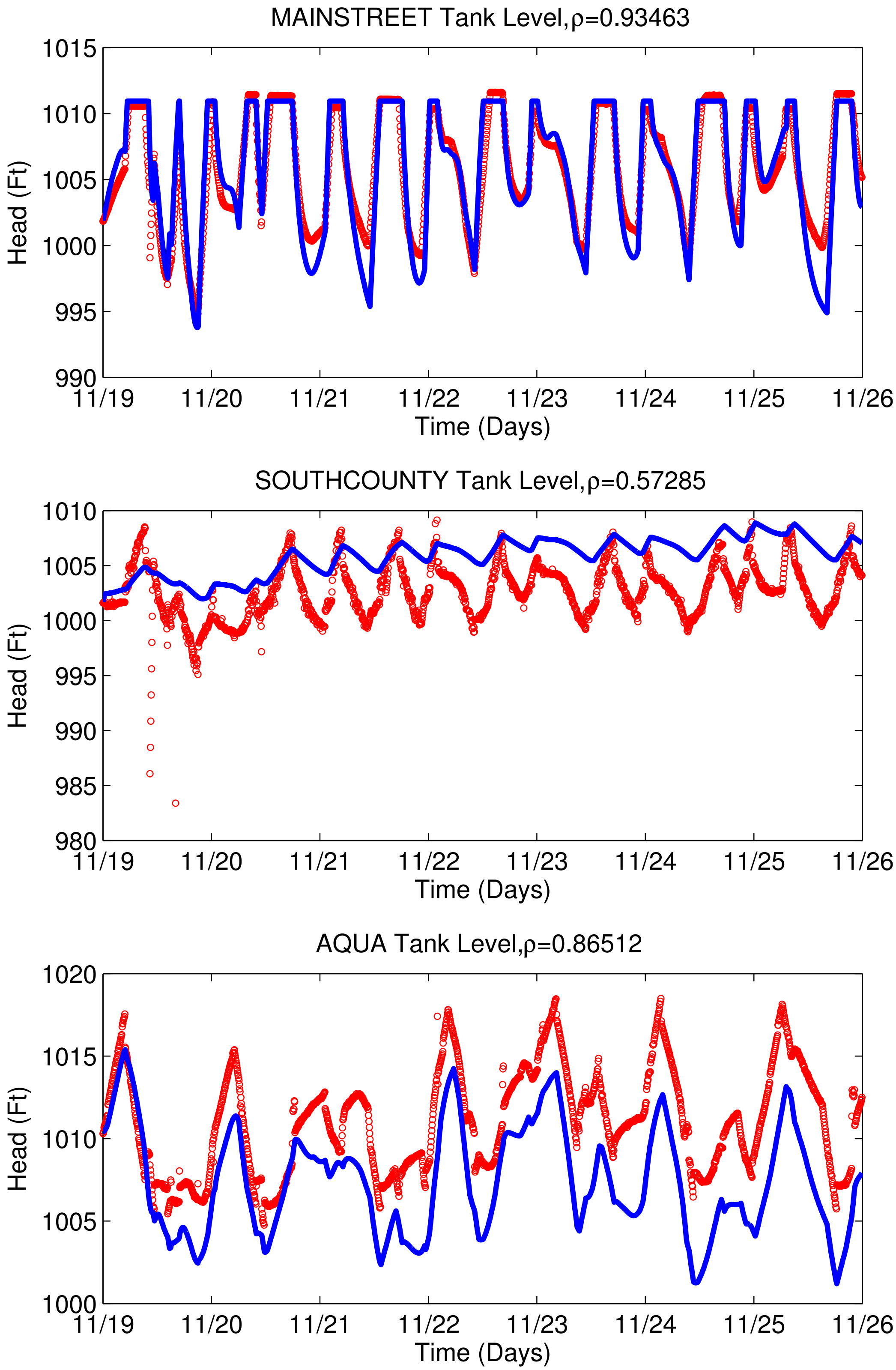

Figure 33: Measured and real-time model simulated tank levels.

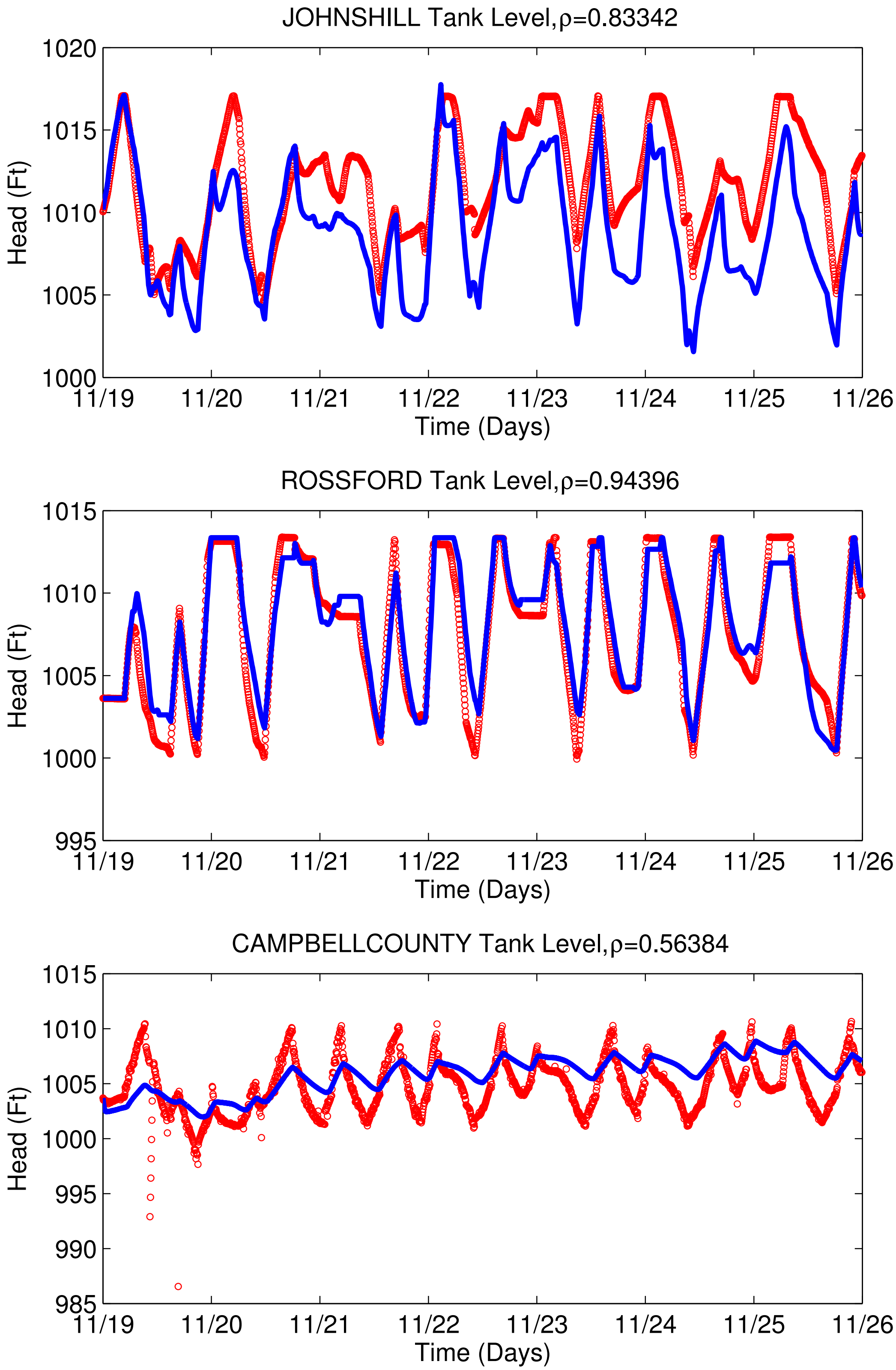

Figure 34: Measured and real-time model simulated tank levels.

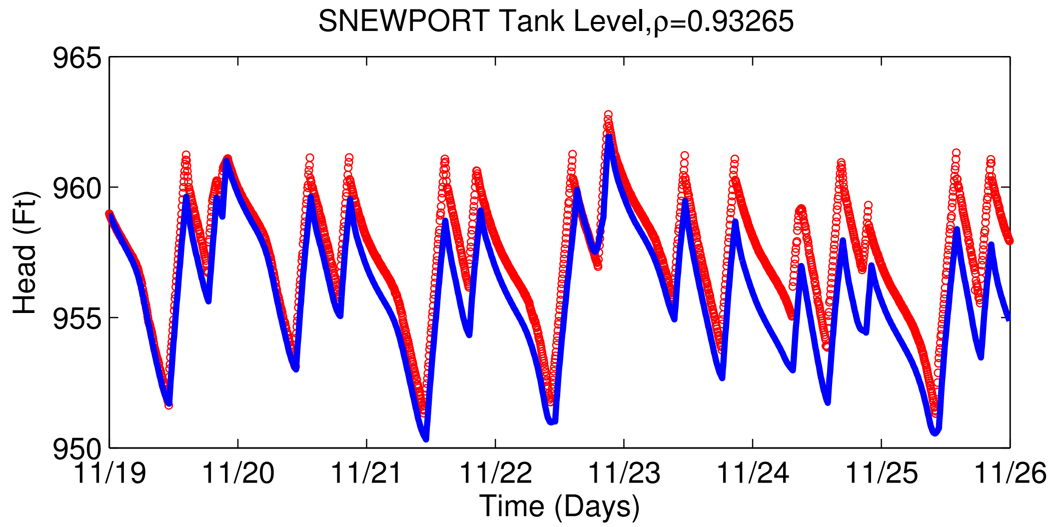

Figure 35: Measured and real-time model simulated tank levels.

South County tanks – exhibit relatively poor correlation. 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 tanks are filled by the Ripple Creek pump station, shown in Figure 29, and the Ripple Creek 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 pipe diameters, incorrect valve statuses, or another cause to be determined.

7 Conclusions

This report has provided the first 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 SCADA data streams in a complex real distribution system. The results were obtained using Epanet-RTX, an open source library of software objects that greatly facilitate the development of real-time environments that connect models and real-time data. The results shown for a one week evaluation period were fully automated by Epanet-RTX data processing algorithms, and prove the feasibility of calculating accurate real-time simulations for complex distribution systems. Given correlation coefficients averaging approximately 0.80 for flows, pressures, and tank levels in the study area – and without complex micro-calibration of system parameters – real-time hydraulic simulation results would be sufficiently accurate that water utilities can contemplate changing some existing work flows, and envisioning new ones.

8 Acknowledgements

The authors wish to thank a variety of individuals for their valuable efforts, without which the study could not have taken place. John Hall of the USEPA National Homeland Security Research Center (NHSRC) helped coordinate the engagement with the NKWD. Lew Rossman and Michael Tryby of the USEPA National Risk Management Research Laboratory provided valuable guidance and technical support for the real-time model development, and the development of Epanet-RTX. We are especially grateful to Robert Janke of NHSRC, who has provided steady encouragement and advice, and carried forward a vision of real-time modeling, without which the project would have stalled. Steve Allgeier, of the USEPA Office of Water, and Terra Haxton, from USEPA/NHSRC contributed valuable suggestions as members of the Epanet-RTX project committee. Hyoungmin Woo and Ernesto Arandia of the University of Cincinnati, and Tom Taxon of Argonne National Lab, participated on the original Epanet-RTX design team, and helped to craft the software architecture that supported this study. Dick Males and Walter Grayman, both private consultants in Cincinnati, OH, gave valuable advice as part of the Epanet-RTX design review committee. Kevin Morley of the American Water Works Association, along with Rob Janke, helped to fund and organize water utility workshops in 2010 that provided valuable industry feedback about the functionality that water utilities would value from a real-time modeling system. Last but not least, individuals from the Northern Kentucky Water District have provided many hours of their time, and full access to information, that was critical for making this project possible. These included Operations Supervisors Amy Matracia and Bill Wulfeck, Information Technology manager William Stewart, and especially Engineering Supervisor Amy Kramer, who led the NKWD support effort. The active support of senior NKWD management was also vitally important, including Vice President of Engineering, Richard Harrison, and CEO Ron Lovan. Finally, we wish to thank Lindell Ormsbee for his guidance as Principal Investigator of the grant under which this work was, in part, undertaken.

Funding for this research was provided by the U.S. Department of Homeland Security, Science and Technology Directorate, 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 support is greatly appreciated. Additional financial support was also provided by the U.S. Environmental Protection Agency's (EPA) Office of Research and Development, under 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." The views expressed in this report are those of the authors, however, and do not necessarily reflect the views or policies of EPA. Mention of trade names or commercial products does not constitute endorsement or recommendation for use. Finally, crucial in-kind support was provided by CitiLogics, LLC, Newport, KY, including the use of proprietary software and specialized technical support.

A Catalog of Operational Notes, Model Updates, Recommendations, and Open Issues

A.1 Operational Notes

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

1. TMTP was off line 9/7/12 - 2/5/13. TMHS lift pumps 3,4,5 ran less than one hour combined during the study time range (to clear water diverted to clearwell by FTTP diversion valve, at a calculated average rate of 18 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.

A.2 Model Updates

The following tables summarize structural modifications (Table 6), and parametric modifications (Table 7), that were implemented for the real-time hydraulic simulations described in section 6.

Table 6: Summary of structural model changes.

Model Elements	Description
New FCV	Modeling TMTP diversion as a flow control valve at 18 gpm.
TMTP_DIVERSION_VALVE	Runtime data for TMHS lift pumps, plus assumed 8000 gpm for each pump, gives an average of 18 gpm diverted to the TMTP clearwell from FTTP during the period when TMTP was off line.
W. Covington PS	Modifications to include bypass (as per drawings)

Table 6: Summary of structural model changes (cont.).

Model Elements	Description
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.
8209_hydrant	Added hydrant node for tracer study.
8209_b	Added pipe for tracer study.

Table 7: Summary of parametric model changes.

Model Elements	Description
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.

Table 7: Summary of parametric model changes (cont.)

Model Elements	Description
PRV settings	Site surveys collected upstream and downstream pressures 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 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_Ther_Reg.
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 100-01. (Note: the K coefficients given by Cla-Val are the same dimensionless ones to be used for the model.)

Table 7: Summary of parametric model changes (cont.)

Model Elements	Description
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 RT 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". **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, Indepdence, 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.
TMTP valving	The entire flow coming from TMTP was discovered to be accounted for by the venturi meter tracked in SCADA as TMHS_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 TMTP; model link 4606 carries flow, and link 13326 does not.

Table 7: Summary of parametric model changes (cont.)

Model Elements	Description
Bromley inlet pipe (2798)	Closed because Bromley was out of service during study period; would need to adjust status to reflect actual state when it reopened, although errors will be obvious from tank levels.
Johns Hill Rd. PRV	Reversed direction; was facing uphill.
US 27 Pump 2	Pump curve shifted 24 ft to match operating point.
Ripple Creek Pump 2	Shifted to match operating point.
MPTP Clearwell	Increased the MPTP clearwell bottom elevation from 705 to 721. This is consistent with USGS ground elevation data at the clearwell base, showing about 725. The bottom elevation was set at 721 because when 20 feet of clearwell depth are added, one obtains the 741 maximum elevation which is the name of the pressure zone.
Waterworks Pump Curves	Changed the pump curves assigned in the model to the waterworks PS 1-3 pumps. Was assigned to single point curves but changed to curve that accurately represents documented head-discharge data. This curve was already in the model, even though it was not being used. Note these pumps are variable speed, and there remains the issue of determining their speed from SCADA.

A.3 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. These include issues related to SCADA, model infrastructure and operations data, and real-time model configuration.

1. The normal procedure for testing PRV settings should be reviewed 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. Key PRVs that are SCADA controlled, or used actively for pressure management, should have both stem position and upstream/down stream pressures transmitted via SCADA. This would allow the valve status to be determined, and thus 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. Does not appear that Delta storage mode is reliably configured in historian.

4. Investigate and fix scada measurements for: Bullock Pen meter pit flows 1-3 and pressure; Chesapeake 1 regulator pit flow (critical for 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 stream that record speeds of waterworks variable speed pumps. 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. SCADA data indicate a 10 ft. head loss through the check valve at Ripple Creek PS at a flow of 800 gpm (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. Should be inspected.

7. The Dudley 1080 pump station flow signal is very noisy; it is understood that this sensor has been replaced in 2013.

8. Investigate and confirm piping details surrounding the three gravity feed mains from the two FTTP clear wells. 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 see 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 see essentially equal hydraulic pull on both clearwells, indicating they are not serving separate zones.

9. 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 Kenton Lands tank, which was built at the same time?

10. Confirm MPTP clearwell bottom elevation.

11. Confirm diameters of pipes 8993, 9539, 9540, 8993, 15093.

12. Confirm status of pipe 9778.

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.)

14. Calculated hydraulic heads for Dayton and Bellevue tanks, using the 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 corresponding to November 2011 data

16. Experiment using step interpolation for all resampling of pump station flows that will ultimately be trimmed by the status timeseries. This allows gaps when the pump is on 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.

18. Add bypass PRVs to model.

19. Include measured pressures from field study in 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. MPTP calculated supply flow should probably be changed to include a second moving average on the actiflo flow rate.

22. Investigate MPTP clearwell diameter. Modeled value is 150 ft; Google earth says more like 170 ft.

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).