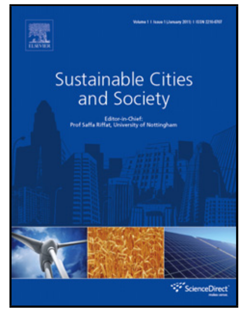
Accepted Manuscript



Title: Enhancing water system models by integrating big data

Author: M. Ehsan Shafiee Zachary Barker Amin Rasekh

PII: S2210-6707(17)30384-0
DOI: https://doi.org/doi:10.1016/j.scs.2017.11.042
Reference: SCS 872

To appear in:

Received date: 10-4-2017
Revised date: 28-11-2017
Accepted date: 28-11-2017

Please cite this article as: M. Ehsan Shafiee, Zachary Barker, Amin Rasekh, Enhancing water system models by integrating big data, *<![CDATA[Sustainable Cities and Society]]>* (2017), https://doi.org/10.1016/j.scs.2017.11.042

Enhancing water system models by integrating big data

M. Ehsan Shafiee[a], Zachary Barker [b] , Amin Rasekh [c]

[a] *Corresponding author, Hydraulic Engineer/ Data Scientist, Sensus USA Inc. ehsan.shafiee@xyleminc.com*
[b] *Hydraulic Engineer/ Data Scientist, Sensus USA Inc., zachary.barker@xyleminc.com*
[c] *Hydraulic Engineer/ Data Scientist, Sensus USA Inc., amin.rasekh@xyleminc.com*

Abstract

The past quarter century has witnessed development of advanced modeling approaches, such as stochastic and agent-based modeling, to sustainably manage water systems in the presence of deep uncertainty and complexity. However, all too often data inputs for these powerful models are sparse and outdated, yielding unreliable results. Advancements in sensor and communication technologies have allowed for the ubiquitous deployment of sensors in water resources systems and beyond, providing high-frequency data. Processing the large amount of heterogeneous data collected is non-trivial and exceeds the capacity of traditional data warehousing and processing approaches. In the past decade, significant advances have been made in the storage, distribution, querying, and analysis of big data. Many tools have been developed by computer and data scientists to facilitate the manipulation of large datasets and create pipelines to transmit the data from data warehouses to computational analytic tools. A generic framework is presented to complete the data cycle for a water system. The data cycle presents an approach for integrating high-frequency data into existing water-related models and analyses, while highlighting some of the more helpful data management tools. The data tools are helpful to make sustainable decisions, which satisfy the objectives of a society. Data analytics distribution tool Spark is introduced through the illustrative application of coupling high-frequency demand metering data with a water distribution model. By updating the model in near real-time, the analysis is more accurate and can expose serious misinterpretations.

Keywords:
water systems, modeling, big data, automation, Hadoop, Apache Spark, cloud computing

1. Introduction

The water resources community relies on computer models to conceptualize and reproduce behavior of systems, aiding in planning, design, and analysis.

The use of computer models is growing due to the need for deeper insights into water systems and providing sustainable solutions for smart cities [1]. Models are formulated by developing a set of mathematical equations and rules, which mimic the real behavior of the system and decisions of stakeholders, and can be executed in an iterative fashion. These equations represent universal laws while parameters represent local systems. Parameters are typically characterized using averages, probability distributions to specify the likelihood of parameters at different states, and assumptions. Model parameters are updated to best reflect the actual system, often done manually when results deviate from field data. This fashion of updating models is time-consuming. Further, due to the speed at which some spatially heterogeneous variables (e.g. water demands and precipitation) change, it is nearly infeasible to manually update with fine resolution.

Engineering advances in sensor and communication devices allow for the continuous monitoring of many systems including water systems. The purposes of these devices are to record and relay time series data with high frequency. Pertinent parameters measured by such devices include flow, quality, and stage; all of which are *in situ*. Technological advancements allow many sites to be monitored in near real-time with very little oversight. This type of measurement creates so-called big data, which relates to the collection in the data cycle – also including, storage, purification, and analysis of large-size data sets [2, 3].

The technological advances in acquisition, processing, and storage of this big data, are poised to greatly advance water systems modeling. The efforts to update models in real-time using large datasets require engineering involvement and discretization. The typical practice is to acquire and format new data so that model parameters can be updated. This two-step practice is time-consuming, insufficient and may introduce many errors that subsequently increase the computational efforts to calibrate these models [4, 5, 6]. In this process, the term *real-time* modeling is overused. Truly real-time models automate the entire process from remote sensing to model output, completing the data cycle.

The authors describe a more thorough integration of high-frequency data with water simulation models. The benefits and challenges are discussed along with examples of integrating big data and models. This work emphasizes the necessity for the collaboration of industry and academic sectors in developing such processes. A generic framework is proposed for the processing of large-size data, collecting valuable information from data, and furthermore, using data to enhance water computer models. Done correctly, these automated models can form the *nervous system* for smart resource management; addressing the resiliency and reliability of water systems in near real-time. This study envisions the process of integrating big data with models and discussing the challenges along with the benefits.

2. Big Water Data

Big data is systematically characterized with three parameters: *Volume, Velocity, & Variety* [3]. Water data possess these three characteristics. Big water data is being generated constantly at unprecedentedly high temporal and spatial resolutions by ubiquitous sensors embedded in the environment, from smart water meters in our houses to satellite-based spectrometer in Earth's orbit.

Millions of smart meters are already deployed, with many more to come according to the reported projections. IHS Markit estimates that over 2 million units were shipped globally in 2015, and this number is projected to double by 2022 [7]. Many utilities are considering or already have plans to install smart meters, such as the City of San Diego, which revamps the master plan to install more than 200,000 meters during the next three years [8]. With these massive number of smart meters and sensors sending measurements of flow, pressure, and many other parameters every second, minute, or hour, water utilities have already begun to have large amounts of data at their disposal.

Other water resources domains have seen similar trends of collecting more data. NOAA alone generates tens of terabytes of hydro-climatic data everyday day from satellites, planes, ships, and other sources [9], which represents a significant untapped opportunity for water resources researchers and professionals. To better manage the challenges of collection and analyses of big water data, NOAA established a new National Water Center in the University of Alabama. Further, NASA's Moderate Resolution Spectroradiometer (MODIS) generates new data at 1.2 MB/s rate, the National Centers for Environmental Information stores more than 25 petabytes of data, and water data are generated at diverse spatiotemporal scales by many separate entities, monitoring different variables [3].

Advanced technologies facilitate processes to store data [10], to mine big data [11, 12], and to make analytical conclusions about the status quo of systems [13]. To process collected data, database technologies were developed to store relational (e.g., SQL) and non-relational (e.g., Hadoop Distributed File System – hdfs) datasets and execute analytics on data using a distributed and non-distributed computational features. In addition to data collection capabilities, machine learning technologies were developed and embedded to facilitate analytical workflows and integrating with cluster computing platforms such as Apache Spark to run analytics at scale [12].

3. Benefits

Integrating big data into water systems introduces technical challenges but we argue these challenges are outweighed by the following benefits:

3.1. Big Data Reduces Model Assumptions

In the most basic terms, big data leads to more information about systems and increase the insight towards the system. Big data can close a number of existing knowledge gaps about the system. In recent years, our understanding about the water systems has been discontinuous such that the stakeholders typically observe systems at the time of planning. Collecting data at the real-time basis using big data techniques enables stakeholders to understand the trend of the systems and make decisions accordingly. Following benefits illustrates the benefit of using big data to reduce model assumptions such as:

1. The conservation polices and regulations, such as rebate programs and water tariff changes, influence water use behavior of individual citizens based on their social attributes such as income and education. Studies addressed the water conservation policies by understanding the social behavior and creating meaningful statistical and mathematical linkages between water usages and social attributes. Using the hourly water consumptions can remove making unnecessary assumptions for designing the water conservation strategies. For example, the Singapore's National Water Agency gains insight into the comparative effectiveness of its engagement strategies, ranging from traditional water tariffs to modern gamification methods, by analyzing the high-resolution water usage data collected by its new advanced metering infrastructure [14]. Such insights and business intelligence may not be obtained using accumulated monthly usage numbers provided by traditional meters. Using traditional meters, the utility had to make assumptions about the water usage response of customers to new tariffs. However, with the benefit of the new technology, the utility was able to adjust water tariff policies as the water is consumed to meet water usage goals.

2. Managing ecological systems requires identifying and understanding underlying significant factors, in addition to creating a model to represent the systems. The Great Lakes ecosystem was studied by collecting the wind speed and water temperature accurately. The high spatiotemporal variability and the sparsity of the in-situ sensors [15] leveraged an unprecedented collection of one million unique measurements made by volunteer ships on the Great Lakes from 2006 to 2014 to obtain the high spatiotemporal variability and the sparsity of these factors. Using these datasets, they were able to fill some gaps that have not been observed before the study.

With more data, engineers can reduce model assumptions (such as the effectiveness of water conservation strategies) and better determine boundary conditions (such as the nodal demands in an hydraulic models of a water network). These benefits come from three types of high-resolution data: spatial, temporal, and unstructured. High-resolution spatial data (e.g., DEM, LiDAR) allow for the heterogeneity of physical features to be considered. Temporal data aids in the ability to consider variables that are in constant flux such as temperature, precipitation, and user demands. Many models account for some temporal changes using patterns or distributions, but also assume longer term stationarity. These models fail to capture changes in land use, climate, and human impacts [16]. In water systems, physical properties such as pipe roughness, flow (rate and uniformity), and channel depth are in constant flux but are often assumed static. Integrating streaming sensor data into models allows engineers to forgo stationary assumptions.

3.2. Big Data Helps to collect social data

In the world of social science, it is a common practice to collect social attributes by conducting surveys. What if the social attributes can be derived by processing unstructured data. The unstructured data refers to data sources that are neither spatial nor temporal, such as human-generated data on social media. Use of social media posts as a means of crowd-sourcing, data acquisition, and uncertainty reduction is already under investigation in many disciplines, such as for water quality data crowd-sourcing using the iPhone camera [17], real-time description of urban emergency events [18], earthquakes detection and notification using twitter posts [19], spatiotemporal evolution understanding of super-storms [20]. Social media posts offer the advantages of being abundant and accessible, but their lack of official legitimacy could introduce new uncertainties to the models, possibly resulting in misleading results. However, in certain applications, mining social media posts would provide timely, valuable information. Such as in the event of a possible water-related outbreak, when the tracking of observations and complaints posted on social media by affected populations might provide the decision makers with more information about the likelihood, scale, and severity of the possible incident.

In addition to social media, with the help of Internet of Things (IoT), new information can be collected as sensors measure environmental factors that contribute to households and environment. For example, it is foreseeable to collect the indoor temperature to relatewith the water usage with. It becomes more plausible to sense the type of water usages in each household by deploying smart devices such as Amazon Echo.

3.3. Big data reduces risk and increases resilience

Risk is directly related to uncertainty. Risk is higher in a more uncertain environment, whether this uncertainty be in possible failure scenarios, loads, capacities, or consequences [21]. Therefore, the reduction in the uncertainties achieved by the integration of high-resolution data in models and decision support systems leads into lower risks and more informed decisions. For instance, the use of high-resolution hydro-climatic data resulted in a realistic simulation of the average discharge regime in the Upper Danube [22]. A narrower flood intensity probability distribution derived using more data, consequently, results in a lower, more accurate failure probability for a given flood control system capacity, and therefore, a lower risk [23]. A design study for flood diversion system of Bakhtiari Dam in Iran demonstrates how the availability of more data enables achieving lower risk for a fixed construction budget [23].

Big data can reduce risk by revealing system weaknesses and enabling allocation of limited resources to the critical weaknesses. In the event of a failure, big data also can accelerate and improve response and selection of mitigation strategy by elucidating the state of emergency and the effectiveness of alternate scenarios to the decision makers. Collection of adequate data in timely fashion leads into a proper selection of response strategy as decision trees are typically developed off-line and require critical data to select the right decision, for example, to flush contaminated water during a water pollution event, the water quality sensor data are valuable information to effectively flush the network [24, 25]

During and aftermath of the super-storm Sandy in 2013, Stafford Township, New Jersey, water utility was able collect and analyze real-time data from various smart sensors and gain a critical view of a utility's infrastructure for strategizing recovery efforts [26]. Smart meters, for example, helped the utility identify, locate, and repair widespread pipes breaks and leakages promptly. Given the fact that many people still had not returned to their property, this success would have been very difficult or impossible to achieve in the absence of the high-resolution data provided autonomously by the ubiquitous smart sensors.

The Las Vegas Valley Water District provides another example of using data to increase resilience. By integrating real-time, high-resolution data with their water distribution model, they improved response times during planned and emergency outages by reducing the time spent setting the model boundary conditions [27]. The hydraulic model is set up with all current operating conditions and pumping schedules and this allows immediate what-if analysis. Emergency outage situations do not conform to the norm of the system, in which the boundary conditions of the model (e.g., consumer nodes' demands) are traditionally set to a handful of generic demand profiles. But with high-resolution, real-time data feed integrated with the hydraulic model, a true image of the current system conditions and its projections under different possible response and recovery scenarios is provided.

In addition, as rivers may become polluted after storms due to new long-term hydrologic regime, identifying the source of a river's pollution is a great concern for decision-makers.To address this concern in the city of Newburgh, the city benefited from a big data application and was able to characterize 13.1 million gallons of overflow at a site over a three-month period by deploying a real-time, high-resolution level monitoring system [28]. Remote field units provided accurate start time, stop time, and overflow volume of combined sewer overflows, reducing the pollution sources uncertainties caused by the combined sewer outfalls being submerged in the Hudson River.

3.4. Big data enables advanced modeling

Human populations are in constant and intertwined interaction with natural and built water systems [29, 30, 31]. A complex adaptive simulation model [32] that couples the human and water systems, therefore, has the immense potential to provide a more accurate image of the reality, as have been proven on modeling drinking water contamination emergencies [33, 34], hydrological systems [35, 36, 37], flood warning [38], amongst others.

Relaxing the unrealistic homogeneity, stationarity, and independency assumptions made possible by the complex adaptive models, nevertheless, has the side effects of the models becoming data-intensive and computationally-expensive. For instance, in a water contamination research study, simulation of a single sociotechnical simulation required 600 seconds, whereas a single engineering simulation took 15 seconds [39].

The advent of big data analytics platforms and the increasing availability of high-resolution data helps resolving both of the data and computation challenges. Researchers have already succeeded to substantially reduce the runtime of sociotechnical models by using Hadoop clusters; for example, from 42 days on desktop computers down to just 2 hours for a large-scale socio-hydrological simulation [13, 40]. Advances in computational social science [41] together with the increased availability of behavioral data from sensors [42], surveys [43], and social media [44, 45] enable quantifying heterogeneity in human behaviors in coupled human-water systems models. Commercial products are already rolled out by companies like WaterSmart Software and Advizzo that interface with the public and harness the power of behavioral data for enhancing consumers satisfaction, water conservation, and beyond. As agent-based modeling has provided the platform for integrated modeling [46, 47, 34], big data stands to replace the agents behavioral assumptions with more accurate profiles of individuals.

4. Challenges

The benefits gained by automating the integration of big data with models are not realized without overcoming some challenges:

4.1. Data may contain gaps or errors

The quality of data that is stored and transmitted to different databases is a concern in big data. Errors can be introduced and propagated by in-situ sensors and processes that store, reshape, and transmit data among databases. Malfunctioning of advanced technologies– including hardware, firmware, and communication devices– in sensors increase likelihood of having gaps in time series data. Missing-data imputation is not guaranteed to recapture the status of transient data.

4.2. Data heterogeneity necessitates advanced warehousing

Environmental sources of data are heterogeneous, which creates complexities in storage and retrieval. A number of studies have been performed by leading technology companies on the effect of data heterogeneity on databases [48]. Data warehouses require significant engineering efforts to store and purge data, tune the computation system, and to maintain the database. The traditional data warehouses are not effective with real-time data, as they are defined by static structures of their schema and relationships between data. The synchronization between transactional data and data warehouses should be redefined for real-time data to support any dynamics in their structure and contents [49]. As more data from heterogeneous sources and dependencies are incorporated into the models, the potential for time lags to affect data currency becomes more prevalent. These challenges are being addressed by computer scientists. However, efforts are necessary to minimize the knowledge gap among civil engineers when real-time water models are deployed.

4.3. Data is prone to confidentiality, integrity, and availability attacks

The proliferated dependency on cloud and network-based assets demands vast, constant temporal and spatial accessibility. This leaves the cyber-infrastructure open to malicious penetration and data manipulation, introducing new risks [50]. A malevolent attempt to sabotage data and compromise its integrity may be staged at any point from data acquisition to deployment in the data cycle. An outsider attack may compromise chlorine sensors to report lower-than-real concentrations, misleading the network's feed-back disinfection controller, and consequently cause potable water over-chlorination and public poisoning [51]. Additionally, data manipulation by insiders has been observed, as evidenced by the Walkerton E. coli Outbreak [52]. Therefore, along with data confidentiality and availability, a data-reliant water system must be safeguarded against data integrity attacks that might be staged.

5. Proposed framework

Utilizing sensing and computation, engineers have greatly improved the modeling and management of water systems. The current state of the flow of data is illustrated in Figure 1 as the white objects. Sensors are deployed in the environment; data are collected, cleaned, then used as inputs for models. Engineers and decision makers can manipulate the models to receive information, understand state of the environment and, using scenario analysis, make decisions concerning the future. The most valuable piece of the process is the interaction with the model to better inform decisions. However, the preceding steps are very time-consuming when done manually. The gray objects represent the proposed data infrastructure that should be adopted to facilitate automated data integration into models.

5.1. Water Data Lake

The Water Data Lake, Figure 1A, stores data from every step in the process. This data lake should be distributed and redundant in order to facilitate quick querying and reduce data loss. Hadoop-based technologies, along with a handful of components and applications, provide the necessary framework for storing big data. Hadoop is a distributed computing environment that supports the processing and storage of large data sets. A Hadoop-based technology is a customized process that uses the Hadoop environment to perform an application.

5.2. Analytics

Analytical tools (Fig. 1B) are connected to the data lake. The purpose of these tools are to scrub data, fill in missing values, and filter out bad data. Additional analytics can be performed at this step such as statistical summaries and forecasting. Today these processes are often done manually. However, studies show the advantages of automated analytics for scientific discoveries [53, 54, 55]. As the amount of data continues to increase, we will need to employ automated methods. In conjunction with the distributed nature of the data lake, software which allows for distributed computation, such as Apache Spark, should be employed to make computationally-expensive analytics and simulations possible. Scenario analysis for short-term predictive control decisions, for instance, requires next-day hourly demand forecast for the all tens or hundreds of thousands of endpoints in a city to be available for the simulation model. Given the computational expense of accurate time-series forecast methods, such extent of computation easily exceeds the capacity of centralized computers, demanding distributed computing tools.
The Analytics box in Figure 1, therefore, hosts two separate but interfaced libraries: 1) an algorithms library, which acts as a repository for all the data transform functions (e.g., ARIMA for forecast), and 2) a distribution library, which hosts a distribution tool (e.g., Spark) for distributing a collection of independent data transform tasks on a computer cluster.

Apache Spark is a general-purpose platform for distributing independent tasks on a cluster. It has emerged as a popular open-source engine since its inception in 2010 [11]. It provides API's in Java, Scala, Python, and R, and also has a rich set of high-performance, built-in libraries, such as MLlib for scalable machine learning [12] and GraphX for graph-parallel computation [56].

The basic abstraction in Spark is that of a resilient distributed dataset (RDD), which allows users perform in-memory computations on computer clusters in a fault-tolerant manner. A RDD is a set of objects partitioned across nodes in a cluster that can be reconstructed if a partition is lost [11].

Some other key concepts that are necessary for any Spark deployment are: 1) Spark Worker – a cluster node that executes a task, 2) Spark Master – a cluster node that coordinates the resources (i.e., collection of worker nodes), 3) Spark Driver – a client application that requests resources from spark master and executes task on worker nodes, and 4) SparkContext – represents the connection to a Spark cluster. A SparkContext enables access to a cluster through a resource manager, which allocates resources across processes. Once connected, Spark acquires executors on computer nodes in the cluster, which are processes that run computations and store data. Next, it first passes the application code (which is defined in the algorithm library) to the executors and then the tasks for them to run.

A Spark cluster can be set up manually using a collection of physical or cloud-based machines. Most cloud service providers also offer services (Elastic MapReduce by Amazon, Dataproc by Google, etc.) that enable configuring and deploying a cloud-based Spark cluster fast and conveniently. The latter option requires little technical knowledge, and together with the basic examples provided on the Apache Spark official website would create a suitable starting point for beginners. For learning purposes, one may also use Spark in the local mode on a single personal computer. In this non-distributed deployment mode, no earlier setup is required to launch Spark applications and the parallelism is done merely on the set of threads available on the single machine.

5.3. Middleware

The Application Program Interfaces (API) for current water computer models are not designed to integrate data as it becomes available. A middleware component (Fig. 1C), that automatically queries new processed data (Fig. 1A.ii) and formats it to model input, should be introduced. The middleware includes any transformation. For example, the processed data might include one-minute intervals but the model requires five-minute averages, therefore averaging would be applied. Additionally, the middleware should validate the data for each parameter before feeding it as inputs to the model.

5.4. Wrapper

Similar to middleware, a wrapper (Fig. 1D) extends the API of the model. The wrapper provides the functionality to receive streaming data and write model results to the data lake (Fig. 1A.iii). In a real-time EPANET model, for instance, the model boundary conditions, such as individual endpoint demands and tank levels, are automatically updated with their current values streaming in from AMI and SCADA. Therefore, the model outputs, such as pressures and flows distribution, are also current [27].

This step also includes calibration algorithms, which are analytical approaches to characterize empirical parameters such as the friction factors in the Darcy-Weisbach equation. After completion of each specified period (e.g., one day), actual and model-predicted values of tanks levels and other monitored network parameters are compared. As models show discrepancies between the observed and simulated values for a parameter, the calibration model adjusts the model parameters. Additional algorithms can be developed and placed to intelligently detect an anomaly, field issue, and identify its source. The calibration is done automatically but manual investigations and verifications may be still conducted periodically.

5.5. Decision making

Engineers, scientists, and stakeholders can explore the model results interactively using visualization tools. Popular visualization tools include Tableau™, D3, and RStudio Shiny™. Additionally, the user can modify model inputs to reflect possible future scenarios. Altogether this automated process decreases the chance of implementing ineffective decisions in the life-time of the water system.

5.6. Data cycle platform

The infrastructure for Fig. 1 can be engineered in house to facilitate the data cycle. Alternatively, it can be hosted on the new cloud-based services such as Amazon Web Services and Google Cloud Platform if they do not bypass the cost and expertise required for in-house servers.

5.7. Computation cost considerations

Data analytics (e.g., demand time series imputation and forecast) and simulation model runs (e.g., for what-if analysis, calibration, and operation optimization) constitute the majority of computation cost. For data analytics, for instance, week-ahead, hourly demand forecast of 15,000 individual water consumers in a medium-sized town in California has been done in about one minute on a 10-node, cloud-based Spark cluster [57]. Simulation model runs are more expensive, but since they are often performed in parallel to investigate different scenarios, they can be also distributed over a cluster by Spark. Given the scalability offered by Spark, distributing the run on a larger cluster is merely a matter of setting the cluster size to a larger number when configuring the cluster on cloud-based service portals. However, this distribution is feasible when the underlying tasks are parallelizable. A run of a single complex adaptive system simulation, for instance, can be only partially parallelizable, given the interdependencies between the agents in the past and present.

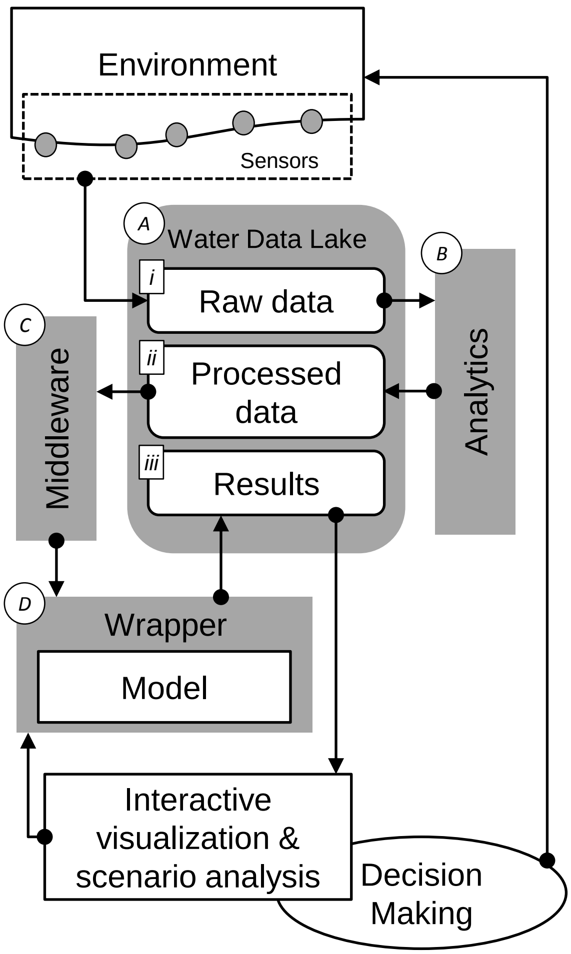


Figure 1: The data cycle for a water system — from collection to decision making — should include a data pipeline that automatically updates a specific model. A) The Water Data Lake stores data during every stage. B) Analytics processes raw data and returns cleaned or forecasted data. C) Middleware pulls, aggregates, and formats data for a model. D) A wrapper provide communication capabilities to a model.

6. Applications

The proposed framework is applicable to many water computer models. The models can be categorized into physical models, that encode mathematical equations governing a water system (e.g. EPANET, SWMM, and MIKE) and policy-related models that encode rights and policies for sharing and uses of water and evaluate the effect of each decisions on water availability (e.g. WRAP and WEAP). Due to accessibility and lower subjectivity, the transition of environmental data into a water model is simpler than the transition of water policies and decisions into these models. This framework can be applied to many models but stands to benefit operational models most. A few examples include water distribution networks, lock and dam operation, treatment plant operation, and storm water management. Below, an illustrative example is briefly explained for integrating high-frequency data with a water distribution model: EPANET [58,].

Traditional use of EPANET involves making assumptions about the demand patterns for customers and rules for pumps and valves. With the use of Advanced Metering Infrastructure (AMI) and Supervisory Control and Data Acquisition (SCADA) data, the hydraulic model can be enhanced by integrating the consumption of each consumer and operations of pumps and tanks. New raw meter reads and SCADA information are stored in the data lake (Fig. 1 *A*). An analytics platform (Fig. 1 *B*) will periodically query and run operations on the data, saving the cleaned data back to the data lake. At each time step, the middleware (Fig. 1 *C*) submits queries to the data lake (Fig. 1 *A*) to check the availability of data for the next time step. The AMI system has transmission latency, therefore, the hydraulic model can be stopped to receive the data. The wrapper (Fig. 1 *D*, which ensures the consumption rate has been stored for each meter and the data is not an error, is checked before running the model and returning the results to the data lake.

7. Discussions and Conclusions

The aim of this manuscript is to encourage development and enhancement of water computer models by integrating big data. High-frequency data is collected from heterogeneous sources across environmental systems. However, the collected data is processed and analyzed at discrete actions. Each action can be thought of as collecting a hunk of data to process and analyzing it to make engineering and scientific discoveries. Despite significant challenges, the data should be integrated with water models in an automated fashion to create real-time models and complete the data cycle for a water system.

A broad framework is proposed to enhance the current water computer models with a new API that enables near real-time dynamic modeling and completes the data cycle. In this way, the model is able to characterize some parameters using data that becomes available in the water data lake. The results of a simulation are also stored in a water data lake for further analysis. The ultimate outcome of this modeling is to enable a stakeholder to gain better understanding on the status quo of a water system and manage this system with more confidence. This type of model enhancement provides ways to encounter water systems as a whole rather than a set of technical, economical, and social systems that are studied separately and in isolation. The outcome of this holistic approach is useful to assess the performance of all aspects of a system.

Most importantly this manuscript emphasizes the increasing importance of computing and analytics in water systems modeling. While many of the challenges are being addressed by the computer science field, future water professionals will need the basic skills to interface with complex database structures and ever evolving API's.

References

[1] Simon Elias Bibri and John Krogstie. On the social shaping dimensions of smart sustainable cities: A study in science, technology, and society. *Sustainable Cities and Society*, 29:219–246, 2017.

[2] Viktor Mayer-Schönberger and Kenneth Cukier. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.

[3] David Hill, Branko Kerkez, Amin Rasekh, Avi Ostfeld, Barbara Minsker, and M. Katherine Banks. Sensing and Cyberinfrastructure for Smarter Water Management: The Promise and Challenge of Ubiquity. *Journal of Water Resources Planning and Management*, 140(7):01814002, 2014.

[4] Leonard F Konikow and John D Bredehoeft. Ground-water models cannot be validated. *Advances in water resources*, 15(1):75–83, 1992.

[5] Jan Seibert and Jeffrey J McDonnell. On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration. *Water Resources Research*, 38(11), 2002.

[6] Calibration of hydrological models on hydrologically unusual events. *Advances in Water Resources*, 38:81 – 91, 2012.

[7] Michael Markides. Water meters 2016 annual report. *Public Relations Review*, 2016.

[8] David Garrick. San diego switching to conservation-friendly smart water meters. *San Diego Union Tribune*, 2017.

[9] Z Flamig, M Patterson, W Wells, and R Grossman. The occ noaa data commons: First year experiences. In *AGU Fall Meeting Abstracts*, 2016.

[10] Claudia Vitolo, Yehia Elkhatib, Dominik Reusser, Christopher J.A. Macleod, and Wouter Buytaert. Web technologies for environmental big data. *Environmental Modelling & Software*, 63:185 – 198, 2015.

[11] Matei Zaharia, Mosharaf Chowdhury, Michael J Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster computing with working sets. *HotCloud*, 10(10-10):95, 2010.

[12] Xiangrui Meng, Joseph Bradley, B Yuvaz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, D Tsai, Manish Amde, Sean Owen, et al. Mllib: Machine learning in apache spark. *JMLR*, 17(34):1–7, 2016.

[13] Yao Hu, Oscar Garcia-Cabrejo, Ximing Cai, Albert J Valocchi, and Benjamin DuPont. Global sensitivity analysis for large-scale socio-hydrological models using hadoop. *Environmental Modelling & Software*, 73:231–243, 2015.

[14] Singapores National Water Agency. Innovation in water singapore. *Singapores National Water Agency*, 8, 2016.

[15] Kevin Fries and Branko Kerkez. Big ship data: Using vessel measurements to improve estimates of temperature and wind speed on the great lakes. *Water Resources Research*, 53(5):3662–3679, 2017.

[16] P. C. D. Milly, Julio Betancourt, Malin Falkenmark, Robert M. Hirsch, Zbigniew W. Kundzewicz, Dennis P. Lettenmaier, and Ronald J. Stouffer. Stationarity is dead: Whither water management? *Science*, 319(5863):573–574, 2008.

[17] Thomas Leeuw. Crowdsourcing water quality data using the iphone camera. 2014.

[18] Zheng Xu, Yunhuai Liu, Neil Yen, Lin Mei, Xiangfeng Luo, Xiao Wei, and Chuanping Hu. Crowdsourcing based description of urban emergency events using social media big data. *IEEE Transactions on Cloud Computing*, 2016.

[19] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):919–931, 2013.

[20] Stephen P Good, Derek V Mallia, John C Lin, and Gabriel J Bowen. Stable isotope analysis of precipitation samples obtained via crowdsourcing reveals the spatiotemporal evolution of superstorm sandy. *PloS one*, 9(3):e91117, 2014.

[21] Louis Anthony Cox Jr. *Risk analysis foundations, models, and methods*, volume 45. Springer Science & Business Media, 2012.

[22] Rutger Dankers, Ole Bøssing Christensen, Luc Feyen, Milan Kalas, and Ad de Roo. Evaluation of very high-resolution climate model data for simulating flood hazards in the upper danube basin. *Journal of Hydrology*, 347(3):319–331, 2007.

[23] Amin Rasekh, Abbas Afshar, and Mohammad Hadi Afshar. Risk-cost optimization of hydraulic structures: methodology and case study. *Water resources management*, 24(11):2833–2851, 2010.

[24] M Ehsan Shafiee and Emily Zechman Berglund. Real-time guidance for hydrant flushing using sensor-hydrant decision trees. *Journal of Water Resources Planning and Management*, 141(6):04014079, 2014.

[25] M. Ehsan Shafiee and Emily Zechman Berglund. Complex adaptive systems framework to simulate the performance of hydrant flushing rules and broadcasts during a water distribution system contamination event. *Journal of Water Resources Planning and Management*, 143(4):04017001, 2017.

[26] Sensus Inc. Hurricane season 2013: Sensus technology helps utilities build resilient networks. 2013.

[27] Paul F Boulos, Laura B Jacobsen, J Erick Heath, and S R I Kamojjala. Real-time modeling of water distribution systems : A case study. *Journal AWWA*, 106(9):391–401, 2014.

[28] Hannah Rosenstein. Data-driven solutions to u.s. wastewater challenges. *SWAN Forum*, 2016.

[29] Scott Campbell. Green cities, growing cities, just cities?: Urban planning and the contradictions of sustainable development. *Journal of the American Planning Association*, 62(3):296–312, 1996.

[30] Jianguo Liu, Thomas Dietz, Stephen R Carpenter, Carl Folke, Marina Alberti, Charles L Redman, Stephen H Schneider, Elinor Ostrom, Alice N Pell, Jane Lubchenco, William W Taylor, Zhiyun Ouyang, Peter Deadman, Timothy Kratz, and William Provencher. Coupled Human and Natural Systems. *AMBIO: A Journal of the Human Environment*, 36(8):639–649, dec 2007.

[31] Efi Foufoula-Georgiou, Zeinab Takbiri, Jonathan A. Czuba, and Jon Schwenk. The change of nature and the nature of change in agricultural landscapes: Hydrologic regime shifts modulate ecological transitions. *Water Resources Research*, 51(8):6649–6671, 2015.

[32] John H Holland. Complex adaptive systems. *Daedalus*, pages 17–30, 1992.

[33] Emily M Zechman. Agent-based modeling to simulate contamination events and evaluate threat management strategies in water distribution systems. *Risk Analysis*, 31(5):758–772, 2011.

[34] M. Ehsan Shafiee and Emily M. Zechman. An agent-based modeling framework for sociotechnical simulation of water distribution contamination events. *Journal of Hydroinformatics*, 15(3):862, jul 2013.

[35] Matthew R Sanderson, Jason S Bergtold, Jessica L Heier Stamm, Marcellus M Caldas, and Steven M Ramsey. Bringing the social into sociohydrology: Conservation policy support in the central great plains of kansas, usa. *Water Resources Research*.

[36] Paul H Nol and Ximing Cai. On the role of individuals in models of coupled human and natural systems. *Environmental Modelling & Software*, 92(C):1–16, 2017.

[37] Alireza Mashhadi Ali, M. Ehsan Shafiee, and Emily Zechman Berglund. Agent-based modeling to simulate the dynamics of urban water supply: Climate, population growth, and water shortages. *Sustainable Cities and Society*, 28:420 – 434, 2017.

[38] Erhu Du, Samuel Rivera, Ximing Cai, Laura Myers, Andrew Ernest, and Barbara Minsker. Impacts of human behavioral heterogeneity on the benefits of probabilistic flood warnings: An agent-based modeling framework. *JAWRA Journal of the American Water Resources Association*, 53(2):316–332, 2017.

[39] Amin Rasekh, M Ehsan Shafiee, Emily Zechman, and Kelly Brumbelow. Sociotechnical risk assessment for water distribution system contamination threats. *Journal of Hydroinformatics*, 16(3):531–549, 2014.

[40] Yao Hu, Ximing Cai, and Benjamin DuPont. Design of a web-based application of the coupled multi-agent system model and environmental model for watershed management analysis using hadoop. *Environmental Modelling & Software*, 70:149–162, 2015.

[41] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

[42] Alessandro Cominola, Matteo Giuliani, Dario Piga, Andrea Castelletti, and Andrea Emilio Rizzoli. Benefits and challenges of using smart meters for advancing residential water demand modeling and management: A review. *Environmental Modelling & Software*, 72:198–214, 2015.

[43] Michael K Lindell, Shih-Kai Huang, and Carla S Prater. Predicting residents responses to the may 1–4, 2010, boston water contamination incident. *Int. J. Mass Emerg. Disasters*, 2016.

[44] Hang Zheng, Hong Yang, Di Long, and Jing Hua. Monitoring surface water quality using social media in the context of citizen science. *Hydrology and Earth System Sciences*, 21(2):949, 2017.

[45] Raechel Johns. Community change: Water management through the use of social media, the case of australia's murray-darling basin. *Public Relations Review*, 40(5):865–867, 2014.

[46] Eric Bonabeau. Agent-based modeling: methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America*, 99 Suppl 3:7280–7, may 2002.

[47] Emily Zechman Berglund. Using Agent-Based Modeling for Water Resources Planning and Management. *Journal of Water Resources Planning and Management*, 141(11):04015025, nov 2015.

[48] Alon Halevy, Anand Rajaraman, and Joann Ordille. Data integration: The teenage years. In *Proceedings of the 32Nd International Conference on Very Large Data Bases*, VLDB '06, pages 9–16. VLDB Endowment, 2006.

[49] Robert M Bruckner, Beate List, and Josef Schiefer. Striving towards near real-time data integration for data warehouses. In *International Conference on Data Warehousing and Knowledge Discovery*, pages 317–326. Springer, 2002.

[50] Amin Rasekh, Amin Hassanzadeh, Shaan Mulchandani, Shimon Modi, and M Katherine Banks. Smart water networks and cyber security, 2016.

[51] Verizon Communication. *Data breach digest": scenarios from the field.* 2016.

[52] Dennis R OConnor. Report of the walkerton inquiry. *Toronto, Ontario: Ontario Ministry of the Attorney General, Queens Printer for Ontario*, 2002.

[53] Shiyong Lu and Jia Zhang. Collaborative scientific workflows. In *Web Services, 2009. ICWS 2009. IEEE International Conference on*, pages 527–534. IEEE, 2009.

[54] Yolanda Gil, Ewa Deelman, Mark Ellisman, Thomas Fahringer, Geoffrey Fox, Dennis Gannon, Carole Goble, Miron Livny, Luc Moreau, and Jim Myers. Examining the challenges of scientific workflows. *Ieee computer*, 40(12):26–34, 2007.

[55] M. Ehsan Shafiee and Emily Berglund. Agent-based modeling approach to evaluate the effect of collaboration among scientists in scientific workflows. *Journal of Simulation*, Submitted, Under Review, 2016.

[56] Reynold S Xin, Joseph E Gonzalez, Michael J Franklin, and Ion Stoica. Graphx: A resilient distributed graph system on spark. In *First International Workshop on Graph Data Management Experiences and Systems*, page 2. ACM, 2013.

[57] Amin Rasekh, Sarin E Chandy, Zachary A Barker, M Ehsan Shafiee, Bruce Campbell, and Travis Smith. Short-term forecast of high-resolution utilities time series at scale. In *World Environmental and Water Resources Congress 2017*.

[58] L. Rossman. Epanet user's manual. *U.S. Environmental Protection Agency Risk Reduction Engineering Lab*, 2000.

The highlights of the study are:

1- Identify and highlights of using the big data—mention the benefit and study the example
2- Identify the challenges along with using the big data for water systems
3- Propose a generic model for integration of water computer models with the big data
4- Support the study and paper with examples