Toward An Integrated Approach to Localizing Failures in Community Water Networks

Qing Han*, Phu Nguyen*, Ronald T. Eguchi‡, Kuo-Lin Hsu*, Nalini Venkatasubramanian*
* University of California, Irvine, ‡ ImageCat Inc., CA, US.

***Abstract*—**

We present a cyber-physical-human distributed computing framework, AquaSCALE, for gathering, analyzing and localizing anomalous operations of increasingly failure-prone community water services. Today, detection of pipe breaks/leaks in water networks takes hours to days. AquaSCALE leverages dynamic data from multiple information sources including IoT (Internet of Things) sensing data, geophysical data, human input, and simulation/modeling engines to create a sensor-simulation-data integration platform that can accurately and quickly identify vulnerable spots. We propose a two-phase workflow that begins with robust simulation methods using a commercial grade hydraulic simulator - EPANET, enhanced with the support for IoT sensor and pipe failure modelings. It generates a profile of anomalous events using diverse plug-and-play machine learning techniques. The profile then incorporates with external observations (NOAA weather reports and twitter feeds) to rapidly and reliably isolate broken water pipes. We evaluate the two-phase mechanism in canonical and real-world water networks under different failure scenarios. Our results indicate that the proposed approach with offline learning and online inference can locate multiple simultaneous pipe failures at fine level of granularity (individual pipeline level) with high level of accuracy with detection time reduced by orders of magnitude (from hours/days to minutes).

I. Introduction

Water is a critical resource and a lifeline service to communities worldwide; it is essential for sustaining the economic and social viability of a community [1]. Often the infrastructures that capture, deliver and store water in cities are many decades old. With the rise in urban populations, these infrastructures have become increasingly complex and vulnerable to failures due to natural, technological and man-made events. When human health and safety, and lives are at stake, it is important to quickly isolate faulty regions and prevent ripples into other interdependent infrastructures. Such cascading impact results in community disruptions ranging from temporary interruptions in services to floods, extended loss of business and mass relocation of residents.

Pipe leak is one of the most frequent types of failures in community water networks [2, 3]. Recent reports from Los Angeles Department of Water/Power (LADWP) and Washington Suburban Sanitary Commission (WSSC) indicate that communities are experiencing an unusual increase in pipe beaks, mainly in old pipes that are susceptible to corrosion problems and pipe joint displacements caused by surface deformations. Extreme weather and heavy rainfall (e.g. Hurricane Sandy 2012, El Niño 2016, La Niña 2017) can stress already weakened pipes to the point of causing major pipe breaks and significant increases in leak rates. Additionally, large-scale disasters can cause pipe failures that may drain vital water supplies required for extinguishing fires and other hazards. Note that about 14-18% of water treated in the United States is wasted through damaged pipelines. Quality of water can also be compromised via contaminant propagation through a faulty pipe. A large-scale pipe failures or a pipe burst may cause severe flooding. Those failures in water infrastructure can have implications on other lifelines [4] - water loss often leads to additional energy expenditures for transporting water from natural resources to end users; polluted water can create a serious public health danger; severe flooding can result in transportation network collapse.

Present status of instrumentation: Water is relatively inexpensive resource. Consequently, most water networks are metered only for billing purposes. In the absence of any metering on water pipes, a utility can do little about leak localization except respond to customer complaints.

Unlike buildings or above-ground structures where damage can be visibly seen, damage to underground pipes is often hidden. The only way to confirm break is to observe water that leaks to the surface. With the advent of sophisticated monitoring systems, such as SCADA (supervisory control and data acquisition) [5] and WaterBox [6], it is possible to monitor pressure values and flow rates at key points within the water network e.g. pump stations. However such network-level automatic control is too coarse-grained and cannot identify specific pipes that are suffering the effect of break. Instrumenting the entire system of pipelines with IoT sensors (pressure transducers and/or flow meters) is both unfeasible (inaccessibility of locations) and expensive. Also community water systems are typically densely connected and complex networks with highly correlated measurements. It is therefore non-trivial to isolate anomalies accurately even with a complete observation.

Related localization approaches: One current practice is to use acoustic instruments listening for variation in the reflected signal, yet their effectiveness is only valid within an area around the leak and doing this is expensive [7]. Another approach adopted by utilities is to use a calibrated hydraulic simulator to localize the leak by enumerating possible leaky points for a best match between the simulation result and the inlet and outlet meter data [8]. Although this appears plausible and is also proposed in [9–11], it is computationally expensive or prohibitive for single/multi-leak localization in large-scale water networks. Because the position and severity of a leak jointly affect the hydraulic behavior, making it difficult to enumerate a match. Alternative methods studied in [12–15] are based on fluid transient modelings, since a sudden break often causes a pressure change followed by a transient wave traveling along the pipe. However, the feasibility of this technique is complex due to the difficulty of obtaining a reliable transient model for a pipeline network (rather than a single pipe evaluated in the previous work). Several other techniques using current-flow centrality based approach [16, 17], state estimation [18] or machine learning (ML) based techniques [19–24] have also been investigated. The performance of these techniques, however, are limited by specific contexts (e.g. single leak, a complete observation of the network, very small and simple network topology).

Our study addresses a more realistic case where the available measurement is limited by the type and number of sensors and the objective is to localize multiple concurrent leaks (instead of single failure) of a real-world water network (instead of simple topology) in seconds/minutes (instead of hours/days). To capture the dynamics of complex water network, we introduce AquaSCALE, a computational framework enabling the fusion of multiple different data sources, robust simulation engines and plug-and-play ML techniques. To the best of our knowledge, AquaSCALE is the first cyber-physical-human system (CPHS) enabled platform that (a) models community water distribution infrastructures and pipe breaks/leaks, and (b) supports the integration of various information for identifying multiple pipe failures.

Contributions of this paper:

•Design and development of a CPHS enabled computational framework to integrate multiple data sources and techniques for localizing leaks in community water networks - (Sec. II).

•A novel two-phase process for leak identification using an offline profile generation for quickly identifying potential faulty pipes and online live data integration for accurately localizing damaged pipelines - (Sec. III/IV).

•A plug-and-play analytic engine that enables selection/integration of statistical ML techniques for fault identification and transforms low level pipeline information into higher level impact (e.g. floods) - (Sec. IV).

•Extensive evaluations of the proposed approach under diverse failure scenarios using real-world water network - (Sec. V).

•A prototype implementation of AquaSCALE that integrates multiple sources of information - (Sec. VI).

II. Approach and System Overview

To quickly identify leak events in real-world water networks, we argue that an integrated approach to fusing multiple (incomplete) sources of information is necessary. AquaSCALE is designed as a CPHS system - the architecture aims to integrate multiple technologies and information sources for localizing leak events. IoT sensing data from water infrastructures can track variations in the network and determine pipe breaks based on reduction in pressure heads and increase in flow rates at failure points [9]. The installation of IoT devices is time consuming and expensive; furthermore their measurements are subject to uncertainty due to sensing errors and measurement correlations. To abstract out correct information with limited IoT observations in a timely manner, sophisticated and high performance algorithms are required. External conditions can be used as additional information for failure detection. For example, extremely cold temperature is likely to cause pipe breaks due to ice blockage - this knowledge can be used to capture patterns of changes in pressure heads (increasing first due to pipe freeze and decreasing due to pipe leak). Human leak-related reports can provide deterministic information. The aggregation of external observations can help improve our assessment of leak events.

As shown in Fig. 1, the core of AquaSCALE framework is a data-driven simulation engine that executes a logical observe-analyze-adapt loop. The input to the analyzer is derived from *Observations* gathered from diverse data sources, and stored in the data management module. The *Analytics* module subsumes models and techniques developed by domain experts and operates on live data to generate higher level awareness for specific application tasks (e.g. leak detection and flood prediction). The awareness then triggers corresponding logical *Adaptations* within the framework (e.g. visualization tools for decision support, actuation and control of water infrastructures). To realize this observe-analyze-adapt loop, AquaSCALE is designed as a workflow based system comprised of multiple modules described in Sec. VI.

In the paper, we apply AquaSCALE for pipe leak identification. AquaSCALE supports a novel two-phase approach for managing water workflows at multiple levels of observation and control. In the first phase, statistical approaches are used to drive the offline creation of a profile model of faults and their impact to help rapid identification of the problem in near real-time. While this initial phase significantly reduces the online detection time, the second phase exploits the availability of dynamic data and compensates for the limitation of the offline model to improve accuracy and efficiency. To support a flexible suite of methods for leak events detection, AquaSCALE incorporates a plug-and-play analytic engine that enables the selection/integration of statistical techniques for improved identification of faults. Statistical based data integration algorithms are used to incorporate IoT measurements with additional observations. This analytic engine facilitates the discovery of an efficient composition of techniques for failure localization in a given water network. In our prototype, robust simulations using an enhanced version of a commercial grade hydraulic simulator EPANET (with added support for IoT sensors and failure modelings) are used offline to train a profile model of anomalous events. The profile and multiple information sources are then used for online rapid coarse fault isolation and fine-grained fault localization (i.e. leak detection).

III. Modeling Resilience In Water Infrastructure

Pipe leaks or breaks, as one of the most frequent types of failures, represent a very high cost vulnerability and is associated with public health implications and wastage of limited resources. It is often caused by operation degradation of pipelines, extremely cold temperature, and large-scale disasters (e.g. earthquake). Leak events may be identified through diverse information sources - an unexpected reduction on pressure heads; an abnormal increase in flow rates; leak-related

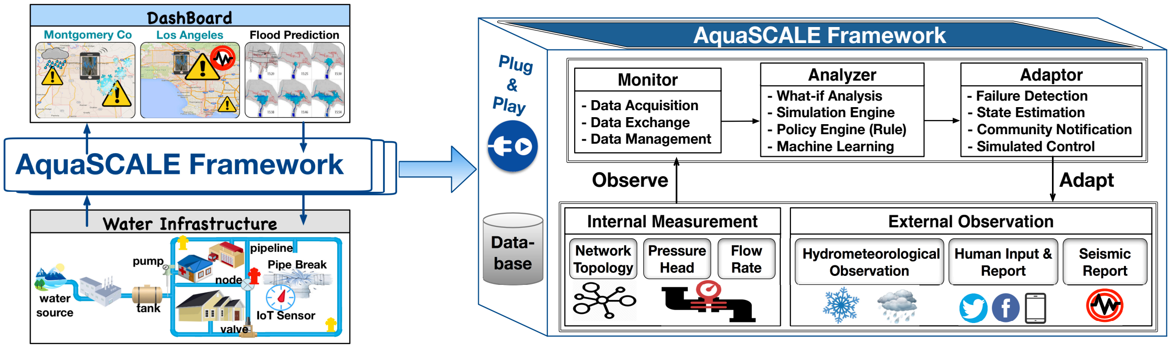


Fig. 1: AquaSCALE prototype architecture. AquaSCALE enables the identification of anomalous events at two layers, a higher service layer that determines water service availability and a lower layer that determines built infrastructure availability, via information integration and plug-and-play capability.

messages posted on social media platforms. Thus we introduce multiple information sources into AquaSCALE, and evaluate its efficacy in the treatment of pipe leaks.

This section introduces the modeling of leak event, IoT measurement in water infrastructures, weather information, and human input. Our experience indicates that IoT measurement alone may work well to identify a single leak event, but, as we explain later in Sec. III-A, it is not sufficiently accurate to isolate multiple concurrent failures. The combination of diverse information sources provides new possibilities for enabling detection of multiple pipe failures. In the real world, extremely low temperatures can cause ice forming in a pipe that leads to complete ice blockage, and continued freezing and expansion inside the pipe increase water pressure heads that leads to pipe breaks. Thus ambient temperature, though is coarse-grained (city-level information), can provide an additional pattern of pressure changing for leak localization. The damage to underground infrastructures is often hidden, and most pipe failures are silent until noticed by people. In the case where IoT measurement is unavailable, human reports on leak events provide indispensable information. Such weather temperature and human input when integrated with IoT measurement can help improve the detection outcome with a higher accuracy in a shorter amount of time.

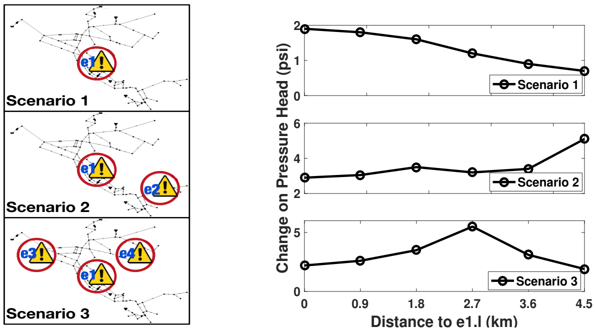
A. Modeling Leak Events

A water system is represented as an undirected graph $G(\mathcal{V}, E)$ (water can flow in both directions) with vertices $\mathcal{V}$ that represent nodes (the joint of pipes), and edges E that represent pipelines. $|\cdot|$ denotes the cardinality of a set. The leak event is denoted as $\mathbf{e} = \{e\}$, where an event $e = (l, s, t)$ is identified by location $e.l$, size $e.s$, and starting time slot $e.t$. The goal is to locate $e.l$ for $\forall e \in \mathbf{e}$. We use and enhance EPANET with the support for IoT devices and failure modelings, named EPANET++. In EPANET++, pipe failures are simulated by emitter that is device associated with node to model the flow through a nozzle or orifice that discharges to the atmosphere [25]. Leakage continuously increases with pressure, and it is often computed using (1) in civil engineering domain [26–28]. More detail refer to [29]. The pipe leak is modeled by

$$Q = EC \cdot p^{\beta} \tag{1}$$

where Q is discharge flow rate at the leak point, EC is effective leak area depending on the discharge coefficient and leak area, p is current pressure head at the leaky node and β is pressure exponent. β typically varies between 0.5 and 2.5 depending on the leak type, and we set it to 0.5 for general purpose [29]. EC indicates the leak size, i.e. $e.s$, and the greater EC the more severity of a leak event. In single leak context, a node will be assigned as an emitter with a EC and a time stamp where the node is leak location ($e.l$), EC is leak size ($e.s$), and time stamp is leak starting point ($e.t$). In multi-leak case, one or more nodes will be assigned as emitters with different EC but same time stamp, to simulate multiple concurrent leaks.

Compared with single leak identification, multiple pipe failures become much more complex to detect and locate. By executing EPANET++, our empirical results show that the changes on pressure head and flow rate are easy to be captured in single failure case (Fig. 2). In scenario 1 where there is a single leak, the total change on pressure values of nodes in a certain distance range of e_1 decreases with increasing distance to $e_1.l$ (Fig. 2b - Scenario 1), and similarly for flow rates that is not shown in the paper for saving space. Here the distance refers to the shortest path between two nodes, and the distance between two adjacent nodes is the length of the connection pipeline. This is because a sudden pipe leak often causes a pressure decrease and a flow rate increase which is followed by a transient wave traveling along the pipe [12, 30]. This pattern can be learnt and captured to identify a leak event. However it is hard to follow a certain changing pattern when multiple failures occur simultaneously, as shown in Scenarios 2 and 3. Multiple leak events interact with each other and jointly affect the hydraulic behavior, resulting in a set of highly correlated observations that makes it difficult to extract correct message in a timely manner. In this case, external data sources and a hybrid of ML based techniques are used to compensate for the limitation of individual information and improve the localization performance. It is worth noting that multiple failures refer to multiple concurrent leak events where the interval between the occurrence of any two events is



(a) Scenarios (b) Changes on pressure head

Fig. 2: Failure scenarios with corresponding changes on pressure head. (a) Three failure scenarios with single leak event $\boldsymbol{e} = \{e_1\}$; two events $\boldsymbol{e} = \{e_1, e_2\}$; three events $\boldsymbol{e} = \{e_1, e_3, e_4\}$. (b) The sum of changes on pressure heads of nodes within a certain range of the location of e_1 along with increasing distance to $e_1.l$ for each scenario.

less than the sampling frequency of IoT devices. The problem then cannot be reduced to single failure detection because leak events cannot be separated by time series.

B. Modeling IoT Measurements

The variation in pressure heads and flow rates due to pipe leaks can be used to obtain critical information on which parts of the system are suffering from the effects of water leaks. To model IoT measurements, a set of pressure and/or flow rate sensors $\mathcal{A}$ are simulated using EPANET++, where $\mathcal{A} \subseteq \mathcal{V} \cup E$ since pressure head is measured on node while flow rate is measured on pipeline. The hydraulic time step, time interval between re-computation of system hydraulics, is used to simulate the sampling frequency of IoT devices. The IoT observations are filtered out based on the pre-defined sensor set $\mathcal{A}$ from the computed results of all nodes and links during the simulation time period.

We consider X as a set of IoT measurements collected from sensors, and Y as a set of event variables, i.e. the states of each node (leak or not) that we wish to identify. An arbitrary assignment to X is denoted by a vector $\mathbf{x} = \{x_a : a \in \mathcal{A}\}$. Similarly for Y, an assignment $\mathbf{y} = \{y_v : v \in \mathcal{V}\}$ is a vector of labels taking from the label set $\mathcal{L} = \{0, 1\}$ where $y_v = 1$ indicates a leakage at node v. Note that the leak event is assumed to occur at node (the joint of pipes), since the interconnect points are more risky than others [1]. In our implementation, leak locations are arbitrarily assigned meaning that the structure of labels is independent, therefore the conditional probability $p(y_v|\mathbf{x})$ can be modeled and trained by using supervised ML based techniques [31]. This is a multi-output classification problem since the dimension of the output is more than one. Due to the mutual independence of labels, the problem is then transformed to multiple binary classifications where a binary classifier is trained for each node independently [32]. The goal is to maximize the number of correctly classified labels by learning a set of classifiers that maps $\mathbf{x} \rightarrow \mathbf{y}$, which is

$$\hat{y}_v = \arg\max_{y_v \in \mathbf{y}} f_{v \in \mathcal{V}}(\mathbf{x}, y_v) \tag{2}$$

Each f_v is a compatibility function indicating how well y_v fits the input $\mathbf{x}$, and $\hat{y}_v$ is the prediction for $\mathbf{x}$ that maximizes the compatibility. Given the knowledge of a water network, ML based techniques can be used to find a solution to (2), however the prediction capability may be limited by the uncertainty of IoT measurements due to noise and interference, incomplete observations due to inaccessible locations and high cost, and highly overlapped observations due to tightly interconnected network structures.

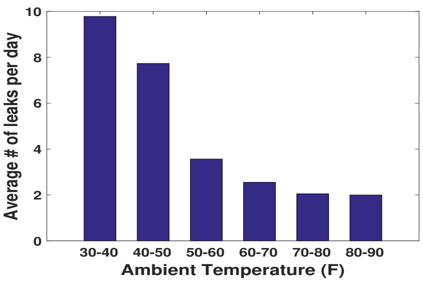


Fig. 3: Average number of pipe breaks per day along with ambient temperatures in the regions of Prince George's and Montgomery County's for recent five years (2012-2016).

C. Modeling Weather Information

When the ambient temperature falls to 20 degrees F or below, pipes may be subject to freezing in the event of extremely cold weather [33]. According to general manager's report from WSSC and weather report from National Oceanic and Atmospheric Administration (NOAA), pipes become more brittle during the winter and the chance of water main breaks rises significantly as the temperature drops (Fig. 3). Cold weather can be a root cause of pipe breaks. The measurement based study in [33] addresses that frozen pipe itself does not typically cause a break. Instead continued freezing and expansion inside the pipe increase water pressure that can dramatically increase stress on a pipe and cause the pipe break following a pressure decrease. Therefore, the pattern of a pressure increase followed by a decrease can help isolate a faulty pipe. That is if the ambient temperature is below $20°F$, a time series of pressure values will be processed that may provide additional information for a faulty pipe. As mentioned, the paper is to evaluate the effect on leak localization by integrating multiple information sources and techniques. Thus weather information is modeled straightforward using probability representation. For each node v, we define $p_v(\text{freeze})$ as the probability of freezing if the temperature is below $20°F$, and $p_v(\text{leak}|\text{freeze})$ as the conditional probability of leak due to freezing. Markov chain will be studied for the modeling of weather information in the future.

D. Modeling Human Inputs

To leverage human inputs, we bring in social media, Online Social Network (OSN), for the incorporation of human sensing. OSN has become a major platform for information sharing in which we can mine interested patterns. Human reports on pipe leak events, such as leak messages posted on Twitter, can help identify the potential damaged region by extracting the associated geographic information. Compared with IoT measurement, human input is considered as deterministic information because it is highly likely to have pipe breaks in a region if people living around report it on OSN. The more reports the higher the level of confidence on the event. Therefore, we incorporate human inputs with the predicted outcome from IoT measurements to improve the detection performance.

Data collected from Twitter represents a previously untapped resource for detecting a pipe break and locating the failure. Human input to AquaSCALE is enabled by integrating a novel Tweet Acquisition System (TAS) [34] developed at UC Irvine, which enhances the monitoring of tweets based on client/application needs in an online adaptive manner such that the quality and quantity of results improve over time. Given a group of interested patterns, TAS can extract related tweets that are then used to help track and locate leak. Twitter users are "sensors" and the posted message with a mention of water pipe break such as "*Pipe bursts @ Sunset Boulevard north of the UCLA campus.*" is an indicator of leak event. To model the human inputs, let $\mathcal{C} = \{c : c = \{v : |l_c - l_v| < \gamma \wedge v \in \mathcal{V}\}\}$ represent a set of subsets of $\mathcal{V}$ (i.e. a set of cliques) inferred from tweets. Here, a clique c is associated with the location l_c where people post the event, and $|l_c - l_v| < \gamma$ means that the distance between l_v (the location of node v) and l_c is less than threshold γ. The threshold γ is a pre-defined parameter indicating the coarseness of the collected Twitter data. For example, if γ is set to 1 km, nodes within 1 km distance to l_c are considered to be likely to leak and will be added into the clique c.

Although there is a high probability for a region to have a pipe break if leak message posted on Twitter, a tweet can be erroneously treated as an indicator of a leak event. For example, tweets like this "*LeakFinderST - innovative leak detection and location in water pipes.*" may be collected but it does not relate to a leak event that we wish to identify. Thus we define a probability of false positive error as p_e, i.e. the likelihood of a tweet that is improperly considered to be relevant, where $0 < p_e < 1$. The confidence that there is a leak within a certain region is represented by

$$p_t = 1 - (p_e)^k \tag{3}$$

where k is the number of tweets collected over a period of time, and with more tweets collected the confidence p_t increases. To model the number of messages received along with the time, we use Poisson distribution that is popular for modeling the number of times an event occurs in an interval of time or space. The human input is assumed to arrive independently of the time. The average number of human reports received in a sampling interval (of IoT devices) is designated as λ that is called arrival rate. The probability of receiving k reports in n elapsed time slots is given by the equation

$$P(\text{k reports in n intervals}) = \frac{(n\lambda)^k e^{-(n\lambda)}}{k!} \tag{4}$$

where e is the Euler's number and $n \in \mathbb{N}$. Combine (3) and (4), the confidence that there is a pipe leak in an area can be computed.

IV. A Composite Leak Identification Algorithm

To enable an accurate and timely leak events identification, we discuss a two-phase approach where the profile model is generated offline by learning an extensive amount of measurements in water infrastructures (Phase I) and the additional observations are integrated with the predicted results from the profile model when live data coming in (Phase II). The proposed composite algorithm reduces the detection time by orders of magnitude by generating a profile offline and improve the detection accuracy by incorporating multiple data sources.

A. Phase I: Training Profile Model Using Measurements In Water Infrastructures

In Phase I, the objective is to train a set of classifiers $f_{v \in \mathcal{V}}$ in (2) to generate a robust profile model f using a great amount of measurements collected in water infrastructures. Here we drop the subscript v and use f to represent a set of trained $f_{v \in \mathcal{V}}$ as the profile model. We first discuss the generation of training features and samples that are then input into the classifiers for a profile model generation. The enabling of plug and play ML based techniques allows us to explore the knowledge of which technologies work well in terms of speed and accuracy under different configurations.

Features of internal measurements in water infrastructures include the topology of the network and IoT observations. The basic topology information, denoted as T, includes node elevation, pipe length, diameter and roughness coefficient, which are static parameters for a given water network. Dynamic IoT measurements X collected from IoT sensors depend on the type and location of the devices. Techniques based on the measurements from pressure and flow rate devices allow a more effective and less costly search in situ [35]. Thus we use pressure transducers and flow meters in the paper. Water pipe leak identification is based on the premise that leakage in one or more locations of the network involves local liquid outflow at leaky points, which will change the pressure head and flow rate at monitoring points [35]. Therefore, we use the difference between two sets of consecutive readings from IoT devices as the features of X. That is x_a is the change on pressure head or flow rate of sensor a. The dynamic IoT observations X aggregated with the static topology T are then the features of a training sample. AquaSCALE in conjunction with EPANET++ enables the selection of a sensor set $\mathcal{A}$ giving the type and number of IoT devices. It allows the study of sensor placement by evaluating different sensor configurations. The problem of identifying an optimal sensor placement for leak detection will be studied in future work. In this paper, given the number of available devices, we use *k*-medoids algorithm to select a group of locations as the sensor set. *K*-medoids is a clustering algorithm related to *k*-means, but it is more robust to noise and outliers [36, 37]. k-medoids partitions $|\mathcal{V}| + |E|$ potential sensor locations into certain number of clusters and assigns cluster centers as the sensor locations, based on the pressure head and flow rate read from nodes and pipes.

As discussed in Section III-B, this multi-output classification problem is transformed to multiple binary classification problems where the classifier is trained separately for each potential leak location v using same datasets and its true labels denoted as Y_v. The profile model $f : T \cup X \rightarrow \mathbb{R}$ can be an ensemble of a set of linear/nonlinear predictors, decision trees, or weak learners, and the parameters of f can be learnt by ML based techniques on the basis of the analysis of pressure and flow rate variations produced by the leak. Note that the performance of specific techniques depends on the structure of water networks, the type and number of IoT devices and their deployment. AquaSCALE allows to test different techniques in isolation or combination, and a hybrid approach may improve the performance since it is thought of as a way to compensate the limitations of individual algorithms.

In the paper, we used scikit-learn package for data processing and analysis [38], and compared multiple well-known ML algorithms including Linear Regression (LinearR), Logistic Regression (LogisticR), Gradient Boosting (GB), Random Forest (RF) and Support Vector Machine (SVM). We proposed a hybrid approach named HybridRSL, a combination of RF and SVM via LogisticR, because RF and SVM remain robust with decreasing number of IoT sensors, and LogisticR has low variances and is less prone to overfitting. As shown in Fig. 4, the same dataset is trained and predicted by RF and SVM separately, and their predicted results, i.e. leak probabilities for each node, are then aggregated as a new feature set and input into LogisticR for further learning. Algorithm 1 shows how classifiers are trained and updated to generate the profile model for Phase II.

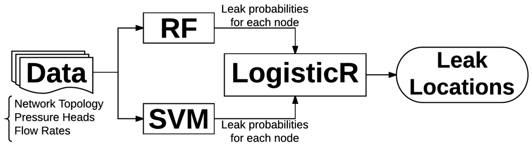


Fig. 4: A sketch of the workflow of HybridRSL approach.

Algorithm 1 Training the Profile Model

```
Input water network topology $T$, IoT measurements $X$, true leak events $Y_v$ and classifiers $f_v$ for $v \in \mathcal{V}$
Output the profile model $f = \{f_v : v \in \mathcal{V}\}$
Objective update $f_v$ to best fit training samples

for $v$ in $V$ do
    $f_v$.fit$(T, X, Y_v)$
end for
```

B. Phase II: Inferring Leak Locations Using Live Ingress Data

In Phase II, we sequentially aggregate multiple data sources to infer the leak locations. Compared with human inputs, IoT measurements and ambient temperatures are relatively stable data sources. We can expect telemetry readings from these two sources at a certain interval once the sensing devices are deployed. Due to dynamic and complex social behavior, however, human reports on leak event maybe not available. Therefore, we first use IoT and temperature streams for event prediction, and use additional human inputs for event tuning.

The live IoT observations $\mathbf{x} = \{x_a : a \in \mathcal{A}\}$ together with the topology information T are firstly learnt by the profile model f. It uses predict_proba and predict methods built in the scikit-learn package, whose outcomes are the score/probability of leak for each node, i.e. $P = \{p_v(i) = \mathsf{score}(y_v = i) : 0 \leqslant p_v(i) \leqslant 1 \wedge \sum_i p_v(i) = 1 \wedge i = \{0, 1\} \wedge \forall v \in \mathcal{V}\}$ with $y_v = 1$ indicating having a leak at node v, and a subset of $\mathcal{V}$ that are predicted to leak, i.e. $\mathcal{S} = \{v : p_v(1) > p_v(0) \wedge v \in \mathcal{V}\}$. If the ambient temperature is below $20°F$ and a location v is detected to be frozen, its predicted leak probability $p_v(1)$ will aggregate with $p_v(\text{leak}|\text{freeze})$ based on Bayes' theorem [39]. This is a well-known method to combine probability distributions from experts in risk analysis, and to apply it into AquaSCALE, we simply consider each information source as an expert. The updated leak probability at node v is

$$p_v^*(1) = \frac{q_v^*(1)}{1 + q_v^*(1)} \tag{5}$$

where

$$q_v^*(1) = \prod_{j=1}^{n} \frac{g_{1j}(p_j|q_v = 1)}{g_{0j}(p_j|q_v = 0)} \tag{6}$$

$q_v^*(1)$ is the posterior odds of the occurrence of leakage at node v; g_{1j} (g_{0j}) represents the probability of source j giving probability p_j conditional on the occurrence (non-occurrence) of leakage at node v. Here, the predicted probabilities come from two information sources, IoT measurements and weather data. In this manner (5), more sources of information means more certainty. For example, if the probability of leak is 0.6 predicted by both two sources, then $p_v^*(1)$ will tend to be much higher than 0.6. The aggregated results then updated P and S correspondingly. In set $\mathcal{S}$, potential faulty pipes are identified. However ML based techniques with noisy IoT sensing data work on the predictive perspective whose output is probabilistic. We use entropy to measure the uncertainty of a predicted event (leak or not) at node v on the basis of its leak probability, which is defined as

$$H(y_v) = -\sum_{i=0}^{1} p_v(i) \log p_v(i) \tag{7}$$

The corresponding uncertainty function is given by

$$E[\mathbf{y}] = \sum_{v \in \mathcal{V}} H(y_v) \tag{8}$$

In order to minimize (8), AquaSCALE integrates additional human input to help to enhance the knowledge of leaks and increase the determinacy of the predicted events.

Human reports on leak events as deterministic information are able to correctly reflect pipe failures within a certain region, but are unable to specify an exact damaged position due to various social behaviors. Therefore, the human input is used as an additional subzone-level information, working with the pipeline-level outcomes P and $\mathcal{S}$, to enforce the event consistency and improve the prediction results. The event consistency here refers to the consistency of the pipeline-level and subzone-level predictions. An inconsistent event means that none of pipes in the subzone identified by human inputs is currently predicted to leak. To leverage the human inputs, we apply the higher order potential concepts used in the image segmentation problems, which is used to enforce label consistency in image regions [40]. We define a higher order potential function $\Phi_c : \mathcal{L}^{|c|} \rightarrow \mathbb{R}$ on clique c to assign a cost to each possible configurations (or labelings) of $\mathbf{y}$. By incorporating human inputs, (8) can now be written as

$$E[\mathbf{y}] = \sum_{v \in \mathcal{V}} H(y_v) + \sum_{c \in \mathcal{C}} \Phi_c \tag{9}$$

that is the energy function to be minimized. Because the effects of human inputs is considered to be non-negative in the paper, Φ_c can assign a very high cost to clique c if none of nodes in c is currently predicted to leak, i.e. $\nexists v \in \mathcal{S}$ for $\forall v \in c$. In this case, the node in clique c with the highest entropy (uncertainty) will be selected for further processing.

The higher order potential used by us can be written as

$$\Phi_c = \begin{cases} 0 & if \quad \exists v \in \mathcal{S} \;\; for \;\; v \in c \\ 0 & else\,if \;\; H(y_v) < \Gamma \;\; for \;\; \forall v \in c \\ Inf & else \end{cases} \tag{10}$$

Here we introduce a threshold Γ to decide if a pipeline-level prediction is considered to be determinate enough to ignoring the subzone-level information. That is if the entropy for node v is less than the threshold Γ meaning that the current predicted event is likely to occur, then the leak information on node v will not be updated by human inputs. According to (9) and (10), an inconsistent event can push the energy to the infinity. In order to minimize (9), in Algorithm 2, a set of leak locations S is firstly identified by the profile model f and then updated based on clique c by adding a candidate v^* if $\Phi_c = Inf$ and $v^* = \arg\max_{v \in c} H(y_v)$. Correspondingly, $p_{v^*}(0)$ and $p_{v^*}(1)$ will be updated to 0 and 1, and $H(y_{v^*})$ will be 0. In this manner, the inconsistent events will be forced to change and the total energy will be reduced because the infinite potentials are eliminated and the entropy of certain nodes are reduced.

Algorithm 2 Inferring Leak Events

```
Input water network topology $T$, IoT measurements $\mathbf{x}$, profile model $f$, leak probability due to frozen $p_{v \in \mathcal{V}}(\text{leak}|\text{freeze})$ and human inputs $\mathcal{C}$
Output an updated set of leak locations $\mathcal{S}$
Objective minimize $E[\mathbf{y}]$ in (9)

/* Event Prediction */
$P = f.\text{predict\_proba}(T, \mathbf{x})$; $\mathcal{S} = f.\text{predict}(T, \mathbf{x})$
for $v$ in $\mathcal{V}$ do
    if $v$ is detected to be frozen then
        $q_v^*(1) = \frac{p_v(1)}{p_v(0)} * \frac{p_v(\text{leak}|\text{freeze})}{1-p_v(\text{leak}|\text{freeze})}$
        $p_v(1) = \frac{q_v^*(1)}{1+q_v^*(1)}$
        $p_v(0) = 1 - p_v(1)$
        $\mathcal{S} = \mathcal{S} \cup \{v\}$ if $p_v(1) > p_v(0)$
    end if
end for

/* Event Tuning */
$\mathcal{C} = \{c : c = \{v : |l_c - l_v| < \gamma \wedge v \in \mathcal{V}\}\}$; $\Phi_{c \in \mathcal{C}} = \text{Inf}$
for $c$ in $\mathcal{C}$ do
    if $\exists v \in \mathcal{S}$ for $\forall v \in c$ then
        $\Phi_c = 0$, break
    end if
    if $\Phi_c \neq 0$ then
        $v^* = \arg\max_{v \in c} H(y_v)$
        if $H(y_{v^*}) > \Gamma$ then
            $p_{v^*}(1) = 1$, $p_{v^*}(0) = 0$, $\mathcal{S} = \mathcal{S} \cup \{v^*\}$
        end if
    end if
end for
```

V. Experimental Study - Using AquaSCALE for Leak Event Identification

In this section, we evaluated the proposed identification approach on single- and multi-failure scenarios, tested multiple ML based techniques in isolation and combination, and examined the impact of incorporating IoT measurements and additional observations. We begin by describing the setup and datasets under which the experiments are conducted, and introduce the performance metrics and the results.

A. Experimental Setup and Datasets Generation

Water Networks. AquaSCALE is evaluated using two water networks - a canonical water network provided by the EPANET (named EPA-NET) and a real subzone of WSSC water service area provided by WSSC (named WSSC-SUBNET). A graph representation of EPA-NET with $|\mathcal{V}| = 96$ and $|E| = 118$, and WSSC-SUBNET with $|\mathcal{V}| = 299$ and $|E| = 316$ is shown in Figure 5. The elevation of pipes varies with the topography, and each pipe has four attributes - length, diameter, roughness coefficient, and status (open or close controlled by a valve). Each node has a pattern of time variation of the demand (i.e. consumption), and leak events are simulated at nodes. EPANET++ is used to perform extended period simulation of hydraulic behavior, which computes pressure heads at nodes and flow rates at pipes.

IoT Sensing Data. Extensive simulations are run on these two networks using EPANET++. Given the number of devices, we first identify a set of sensor locations, and generate a great amount of IoT measurements for profile training. As mentioned, features of a training sample are the topology of a water network and changes on sensing values. The number of training and testing samples are $20,000$ and $2,000$ respectively. For each simulation run, there is at least one and at most 5 leak events, and the number of events follows the uniform distribution i.e. $U(1, 5)$. The leak events are generated with arbitrary locations and sizes but same starting time since we aim to study concurrent failures that are harder

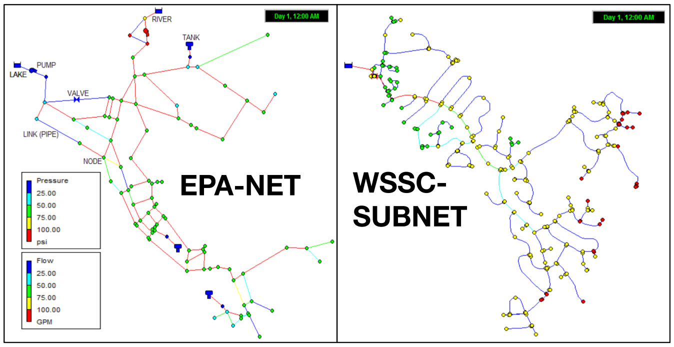


Fig. 5: A graph representation of EPA-NET - a canonical water network provided by EPANET with 96 nodes, 118 pipes, 2 pumps, one valve, 3 tanks and 2 water sources, and WSSC-SUBNET - a subzone of WSSC service area with 299 nodes, 316 pipes, 2 valves and one water source.

to pinpoint. The sensor set $\mathcal{A}$ is selected using k-medoids algorithm based on the given information of IoT devices. The sampling frequency of IoT devices is 15 minutes. Since the goal is to identify leak locations, we assume that the leak starting time $e.t$ is known. The change on pressure heads and flow rates is then computed by taking the differences between the sensing values at $e.t-1$ and $e.t+n$, where n is the number of elapsed time slots after leaking, as in (4).

Human Sensing Data. From January 6, 2016 to April 1, 2016, the east coast of the US experienced extremely cold temperatures, while the west coast experienced high precipitation due to El Niño effects. We collected 30 million "leak-related" tweets posted in the US during this period using TAS system. Since this data contains significant noise, it was treated as described (Sec. III-D). Based on the result statistics, the arrival rate λ of human inputs is set to 1 per 15 minutes, and the false positive error p_e is set to 0.3. The coarseness parameter γ, determining the clique c, is set to different values to test the impact of incorporating with human inputs. More in-depth analysis of those tweets, such as the distribution among different facilities and how soon after the leaks are the tweet posted, will be discussed in future work. In Algorithm 2, the node in c with the highest entropy will be considered as the most risky point, and it will be predicted to leak if the entropy is greater than threshold Γ. Here Γ is set to 0 to always consider human effect. For each simulation run, given an elapsed time slot n, a random number between 0 and 1 is generated for obtaining the number of received tweets k based on (4), and the confidence probability p_t can then be computed based on (3). With the lapsed of time, more human reports can be collected to help identify pipe failures.

Environmental information - Ambient Temperature. In the paper, the probability $p_v(\text{freeze})$ and $p_v(\text{leak}|\text{freeze})$ are set to 0.8 and 0.9 respectively for all $v \in \mathcal{V}$. It might be different for every node since the vulnerability to low temperature depends on a variety of factors, e.g. material, age, location, which will be studied in future work. For each simulation run, a random number between 0 and 1 is generated for each node and it will be used to decide if the connected pipe is frozen based on the pre-defined probabilities. It is likely to have more pipe failures under extreme cold temperatures, which will be used to drive failure scenarios.

Failure Scenarios. We evaluate the proposed composite algorithm of pipe leak identification through two-failure scenarios over different evaluation strategies. We generated $20,000$ single- and multi-failure scenarios separately for training and $2,000$ for each for testing. *Single Pipe Failure* represents that there is only one leak event, which is denoted as $e = \{l, s, t\}$. While multiple pipe failures represents that multiple leak events occur simultaneously, denoted as a set of events $\mathbf{e} = \{e_i : i = 1, ..., m\}$ where m is the number of leaky points and $e_i = \{l_i, s_i, t_i\}$. Multi-failure is often caused by the ice blockage in winter, thus *Pipe Failures due to Low Temperature* is considered as the use case of multiple leaks. The faulty pipes will be located by using different strategies - measurements in water infrastructures with diverse ML based techniques, weather information, and/or human inputs.

B. Performance Metrics

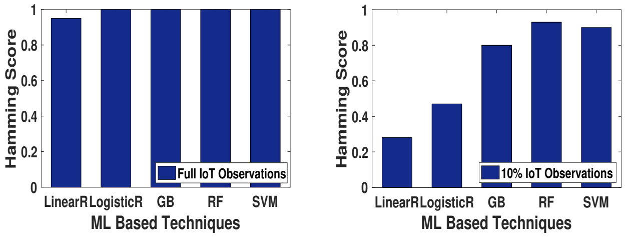
The effectiveness of the proposed algorithm is evaluated in terms of following metrics. **Hamming Score** is defined as $\sum_{v\in\mathcal{V}} \frac{\mathbb{1}[\hat{y}_v=1\wedge y_v=1]}{\mathbb{1}[\hat{y}_v=1\vee y_v=1]}$ where $\mathbb{1}$ is an indicator function. It is the number of leak events correctly predicted divided by the union of predicted and true leak events. The score is bounded by 1 and the higher the score the greater number of leaks that are identified. **Percentage of IoT Observations** is the percentage of IoT deployment penetration. In practice, we want to reduce the number of devices since the installation and maintenance are very expensive. Here $\mathcal{A} = \mathcal{V} \cup E$ with $|\mathcal{A}| = |\mathcal{V}| + |E|$ refers to the full (100%) IoT observations. **Elapsed Time Slot** is the number of time slots elapsing after the leak event, denoted as n. A time slot is a 15 minutes time interval, determined by the sampling frequency of IoT device. With n increasing, on one side, more observations including IoT data and human input will be collected, which may provide more information. On the other side, it may also waste more water, and increase the risk to public health due to water contamination and to other infrastructures due to cascading events.

C. Experimental Results

In this section, the proposed approach for leak event identifications is validated through a detailed simulation study. We begin by plugging and playing several ML based techniques for both single- and multi-failure scenarios using EPA-NET network, and apply Hybrid-RSL technique that outperforms others for following experiments. The effectiveness of integrating diverse data sources for failure detection is evaluated by running extensive simulations on both EPA-NET and WSSC-SUBNET networks. Flood as a cascading event is modeled and predicted to help explore the impact of pipe failures.

Figure 6 illustrates the comparison of different ML based techniques for single leak identification. Those techniques have similar high hamming scores as using 100% IoT observations (Fig. 6a), while RF and SVM can keep a better performance even with 10% IoT (Fig. 6b). Figure 7a/7b show the comparison of RF, SVM and Hybrid-RSL in terms of hamming score for single- and multi-failure scenarios. With a lower percentage of IoT observations, RF yields a higher score compared with SVM. With more IoT data available, SVM outperforms RF as using around 70% IoT in multi-failure scenarios. With the aggregation of RF and SVM, HybridRSL has the best performance in both single- and multi-failure cases. It also shows that multi-failure is much harder to locate. Other ML and data fusion techniques can also be plugged and tested using AquaSCALE.

In the following result, HybridRSL with the highest score will be used, and the integration of multiple data sources will be examined. Here the distance threshold γ for human inputs is set to 30 meters. Figure 8 describes how much do weather and human data together contribute toward identifying *Multiple Failures due to Low Temperature* using real-world WSSC-SUBNET. In Fig. 8a, only IoT data is applied, and the result obtained by aggregating temperature and human input is shown in Fig. 8b. The plotted surface shows how the hamming score varies with the percentage of IoT observations and the elapsed time slots. It clearly illustrates that AquaSCALE with the integration of weather and human information is robust for locating leak events even with limited IoT data. Incorporating with human input can increase the score, however, more human inputs as the time elapsing do not provide significant improvement. Because the false positive error of human data is small in the simulation. Figure 7c/8c present the increment on hamming score by incorporating weather and human data, and the incrementation is more significant with less IoT data. Figure 9 shows that the efficacy of incorporating with human input decreases with the coarser Twitter data. By adding temperature information, however, it can compensate the impact of loose human data and keep the score higher. In Fig. 10, detection using only IoT data is sensitive to the maximum number of leak events, but the aggregation of additional information can help locate failures and output a better result.



(a) Full IoT Observations

(b) 10% IoT Observations

Fig. 6: EPA-NET with *Single Failure* - Comparison of ML techniques for single leak identifications using (a) full and (b) 10% IoT observations.

D. Exploring Impact - Flood modeling and prediction

To capture the impacts of pipe failures and improve post-event awareness, AquaSCALE incorporates flood modeling and prediction to study cascading events. We apply BreZo simulator for flood prediction on WSSC-SUBNET water network. BreZo is a hydraulic model and has been successfully applied in simulation of dam breaks [41, 42] and floods [43, 44]. It can efficiently simulate water flows in varying shapes of the earth surface. A detailed description can be seen in [45]. In BreZo, the flood is predicted based on the digital elevation map (DEM), interpolated from node elevations, shown in Fig. 11a. To feed leak information into the flood model, we use (1) to calculate the outflow rate based on pressure readings, which is then input into BreZo for flood simulations. Two leak events are simulated at v_1 and v_2 with different leak sizes but same start time, and Fig. 11b shows that the flood spreads along the earth's surface. This information can be used by water agencies and city planners for damage control, community notifications and evacuation plans.

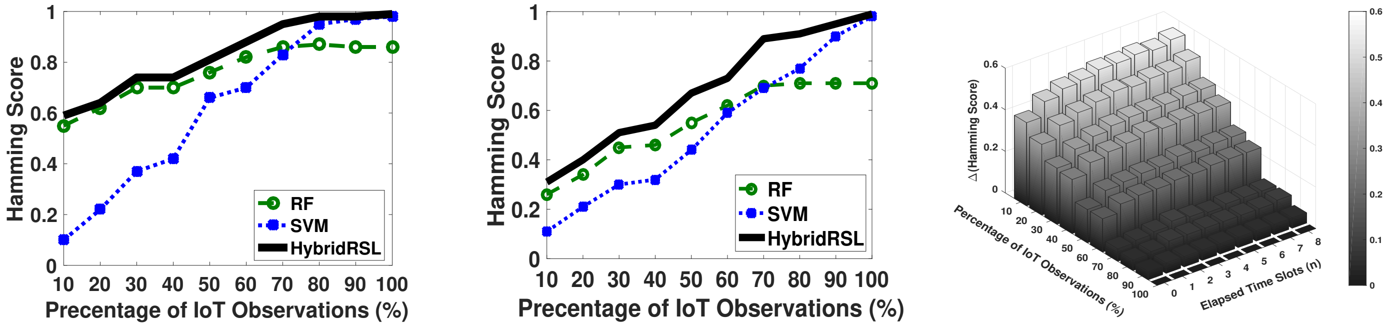
VI. Towards A Prototype Implementation

AquaSCALE framework is a flexible and extensible platform to capture and visualize dynamic community water systems at multiple levels. Our initial implementation of AquaSCALE is designed as a workflow based system comprised of multiple components.

The **Scenario Generation Module** enables water managers and analysts to provide meaningful and diverse water contexts to the framework by generating a range of situations. A user of the tool can start defining a situation by choosing a geographic region, entity elements of interest in that region, and using additional modules to identify hazard, vulnerability, restoration, and impact of the hazard at a temporal and spatial scales of choice. The **Sensor Data Acquisition Module** enables gathering of real-time field information for predefined scenarios by projecting the effects of new updates from the field on simulation outcomes. The **Integrated Simulation and Modeling Engine** executes EPANET++ and BreZo to simulate the dynamic behavior of water networks and interactions between water infrastructures and floods. EPANET++ allows us to model sensor devices and embed them at interested locations, collect pressure heads and flow rates, capture hydraulic and water quality behavior, simulate single/multiple leaks and study impact of damage to infrastructure components. A **Plug and Play Analytics Module** is used to plug and unplug specific information, such as data sets and algorithms, at will depending on the specific context of applications, and to understand the advantage and limitation of diverse strategies in isolation and combination. Users/operators/analysts interact with AquaSCALE using the **Decision Support Module** to manage devices at runtime as they identify vulnerable spots and address accuracy/cost tradeoffs and, to optimize sensor placement for a better performance.

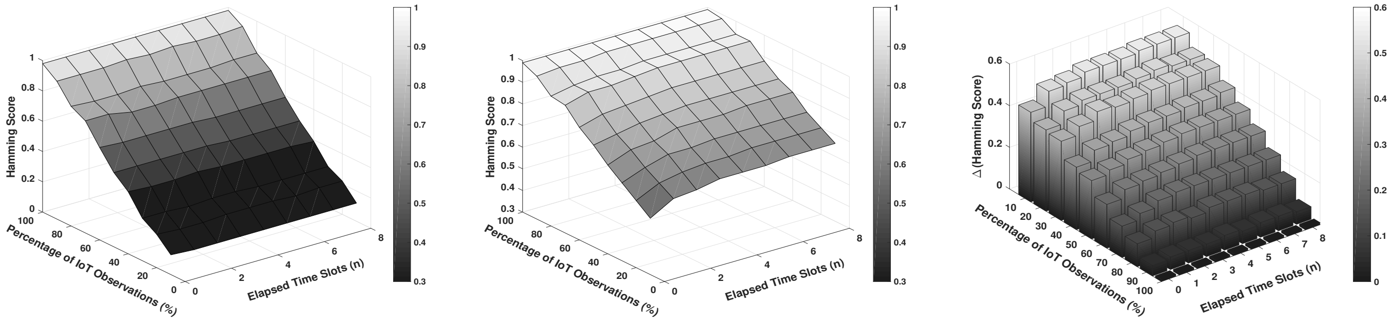
VII. Concluding Remarks

In the paper, we introduced AquaSCALE, a CPHS computational framework, and use it for localizing leaks in community water networks. We formulated multi-leak identification problem, developed an ML based integration mechanism for fusing information from multiple sources, and evaluated it using real-world and synthetic networks. AquaSCALE can be used by water agency operators with expertise in civil



(a) Single Leak (b) Multiple Leaks (c) Hamming Score Increment

Fig. 7: A group of comparisons running on EPA-NET. Comparison of RF, SVM and HybridRSL in terms of hamming score for (a) single- and (b) multi-leak identifications. (c) Average increment on hamming score by adding weather and human inputs.



(a) IoT Data (b) IoT + Weather + Human Data (c) Hamming Score Increment

Fig. 8: WSSC-SUBNET with *Multiple Failures due to Low Temperature* - Average hamming score for multi-leak identifications using (a) IoT and (b) multiple data sources, and (c) increment on hamming score by integrating weather and human data.

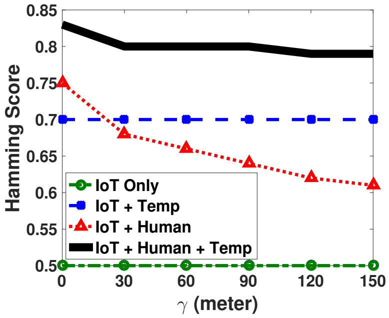


Fig. 9: WSSC-SUBNET with *Multiple Failures due to Low Temperature* - Average hamming score with coarser twitter data using different sources.

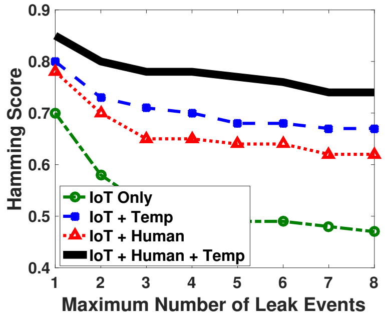
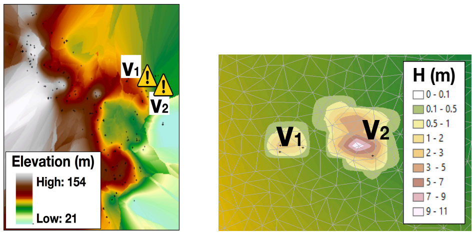


Fig. 10: WSSC-SUBNET - Average hamming score for failure identifications with increasing number of leak events using different sources.



(a) DEM (b) Flooding Map

Fig. 11: Flood prediction based on (a) DEM of WSSC-SUBNET with leaks at v_1 and v_2. (b) Zoom-in flooding map overlaying over the DEM. Flood flows from the center to the outer. H represents the flood depth in meter.

infrastructures to explore problems and solutions in cyberspace before instantiating them into a physical infrastructure. For example, a large section of water systems (usually an entire pressure zone) can be shutdown to prevent cascading failures of pipe burst and to preserve critical water supplies. Such exploration, proactive planning and their effective instantiation during damage/shutdown is relevant in global contexts and is a topic of future research.

Acknowledgements: We are grateful to Prof. Sharad Mehrotra, Prof. Soorosh Sooroshian, Mehdi Sadri, Dan Hoffman, Lingyi Zhang for providing input, tools and datasets for this work. This work was funded under NSF Awards No. CNS-1143705, CNS-0958520, and CNS-1450768.

References

[1] Community resilience planning guide for buildings and infrastructure systems. *National Institute of Standards and Technology*, 2015.

[2] A.-K. Tariq, A.-T. Ziyad, and A.-O. Abdullah. Wireless sensor networks for leakage detection in underground pipelines: a survey paper. *Procedia Computer Science*, 2013.

[3] W. Li, W. Ling, S. Liu, et al. Development of systems for detection, early warning, and control of pipeline leakage in drinking water distribution: A case study. *Journal of Environmental Sciences*, 2011.

[4] H. Zamenian, D. M. Abraham, and K. Faust. Energy loss modeling of water main breaks: a hybrid system dynamics-agent based modeling approach. 2015.

[5] M. Dobriceanu, A. Bitoleanu, M. Popescu, et al. Scada system for monitoring water supply networks. *WSEAS Trans. on Systems*, 2008.

[6] S. Kartakis, E. Abraham, and J. A. McCann. Waterbox: A testbed for monitoring and controlling smart water networks. In *Cyber-Physical Systems for Smart Water Networks*, 2015.

[7] Y. Gao, M. Brennan, P. Joseph, et al. On the selection of acoustic/vibration sensors for leak detection in plastic water pipes. *Journal of Sound and Vibration*, 2005.

[8] A. Nasir, B.-H. Soong, and S. Ramachandran. Framework of wsn based human centric cyber physical in-pipe water monitoring system. In *Control Auto. Robotics & Vision*, 2010.

[9] Z. Poulakis, D. Valougeorgis, and C. Papadimitriou. Leakage detection in water pipe networks using a bayesian probabilistic framework. *Probabilistic Engineering Mechanics*, 2003.

[10] J. Rougier. Probabilistic leak detection in pipelines using the mass imbalance approach. *Journal of Hydraulic Research*, 2005.

[11] R. Puust, Z. Kapelan, D. Savic, et al. Probabilistic leak detection in pipe networks using the scem-ua algorithm. In *Annual Water Dist. Systs. Analysis Symposium*, 2006.

[12] P. Lee, M. Lambert, A. Simpson, et al. Leak location in single pipelines using transient reflections. *Australian Journal of Water Resources*, 2007.

[13] A. F. Colombo, P. Lee, and B. W. Karney. A selective literature review of transient-based leak detection methods. *Journal of Hydro-environment Research*, 2009.

[14] J. Sun, R. Wang, and H.-F. Duan. Multiple-fault detection in water pipelines using transient-based time-frequency analysis. *Journal of Hydroinformatics*, 2016.

[15] P. J. Lee, H.-F. Duan, M. Ghidaoui, et al. Frequency domain analysis of pipe fluid transient behaviour. *Journal of hydraulic research*, 2013.

[16] I. Narayanan, A. Vasan, V. Sarangan, et al. One meter to find them all-water network leak localization using a single flow meter. In *Information Processing in Sensor Networks*, 2014.

[17] W. Abbas, L. S. Perelman, S. Amin, et al. An efficient approach to fault identification in urban water networks using multilevel sensing. In *Embedded Systems for Energy-Efficient Built Environments*, 2015.

[18] F. Fusco and A. Ba. Fault diagnosis of water distribution networks based on state-estimation and hypothesis testing. In *Communication, Control, and Computing*, 2012.

[19] C. Ai, H. Zhao, R. Ma, et al. Pipeline damage and leak detection based on sound spectrum lpcc and hmm. In *Intelligent Systems Design and Applications*, vol. 1, 2006.

[20] J. Mashford, D. De Silva, D. Marney, et al. An approach to leak detection in pipe networks using analysis of monitored pressure values by support vector machine. In *Network and System Security*, 2009.

[21] G. Hessel, W. Schmitt, K. Van der Vorst, et al. A neutral network approach for acoustic leak monitoring in the vver-440 pressure vessel head. *Progress in Nuclear Energy*, 1999.

[22] R. Jafar, I. Shahrour, and I. Juran. Application of artificial neural networks (ann) to model the failure of urban water mains. *Mathematical and Computer Modelling*, 2010.

[23] S. R. Mounce, A. J. Day, A. S. Wood, et al. A neural network approach to burst detection. *Water science and technology*, 2002.

[24] Y. Wang, J. Cao, W. Li, et al. Mining traffic congestion correlation between road segments on gps trajectories. In *Smart Computing*, 2016.

[25] L. A. Rossman. Epanet 2 users manual. 2000.

[26] J. Muranho, A. Ferreira, J. Sousa, et al. Pressure-dependent demand and leakage modelling with an epanet extension–waternetgen. *Procedia Engineering*, 2014.

[27] O. Giustolisi, D. Savic, and Z. Kapelan. Pressure-driven demand and leakage simulation for water distribution networks. *Journal of Hydraulic Engineering*, 2008.

[28] G. Germanopoulos. A technical note on the inclusion of pressure dependent demand and leakage terms in water supply network models. *Civil Engineering Systems*, 1985.

[29] A. Lambert. What do we know about pressure-leakage relationships in distribution systems. 2001.

[30] D. Misiunas. *Burst Detection and Location in Pipelines and Pipe Networks - With Application in Water Distribution*, 2003.

[31] C. Sutton and A. McCallum. An introduction to conditional random fields. 2010.

[32] G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining multi-label data. In *Data mining and knowledge discovery handbook*, 2009.

[33] Freezing and bursting pipes. *Natural Hazard Mitigation*. A publication of the Institute for Business and Home Safety.

[34] M. Sadri, S. Mehrotra, and Y. Yu. Online adaptive topic focused tweet acquisition. In *Information and Knowledge Mgmt.*, 2016.

[35] R. Sarrate, J. Blesa, F. Nejjari, et al. Sensor placement for leak detection and location in water distribution networks. *Water Science and Technology: Water Supply*, 2014.

[36] Y. Lu, I. Cohen, X. Zhou, et al. Feature selection using principal feature analysis. In *Multimedia*, 2007.

[37] A. Malhi and R. X. Gao. Pca-based feature selection scheme for machine defect classification. *Instrumentation and Measurement*, 2004.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.

[39] R. T. Clemen and R. L. Winkler. Combining probability distributions from experts in risk analysis. *Risk analysis*, 1999.

[40] P. Kohli, P. H. Torr, et al. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 2009.

[41] L. Begnudelli and B. Sanders. Simulation of the St. Francis dam-break flood. *Journal of Eng. Mechanics*, 2007.

[42] L. Begnudelli and B. Sanders. Adaptive Godunov-based model for flood simulation. *Journal of Eng Mechanics*, 2008.

[43] P. Nguyen, A. Thorstensen, S. Sorooshian, et al. Flood forecasting and inundation mapping using HiResFlood-UCI and near-real-time satellite precipitation data: The 2008 iowa flood. *Journal of Hydrometeorology*, 2015.

[44] P. Nguyen, A. Thorstensen, S. Sorooshian, et al. A high resolution coupled hydrologic-hydraulic model HiResFlood-UCI for flash flood modeling. *Journal of Hydrology*, 2015.

[45] S. Bradford and B. Sanders. Finite-volume model for shallow-water flooding of arbitrary topography. *Journal of Hydraulic Eng*, 2002.