4. Interdisciplinary Conference on Electrics and Computer (INTCEC 2024)
11-13 June 2024, Chicago-USA

AQUATWIN: A Digital Twin Framework for Early Detection of Water Contamination

Md Rafiul Kabir
Department of ECE, University of Florida,
Gainesville, FL 32611, USA
kabirm@ufl.edu

Sumaiya Afroz Mila
Department of ECE, University of Florida,
Gainesville, FL 32611, USA
mila.s@ufl.edu

Sandip Ray
Department of ECE, University of Florida,
Gainesville, FL 32611, USA
sandip@ece.ufl.edu

2024 4th Interdisciplinary Conference on Electrics and Computer (INTCEC) | 979-8-3503-4945-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/INTCEC61833.2024.10603171

***Abstract*—The stability and well-being of human societies and communities depend on having access to clean, safe water. The sustainability of our civilization rests on our ability to guarantee that the water bodies we utilize are free of harmful contaminants. Digital twins are emerging as a powerful tool for managing complex systems in a variety of Internet of Things applications, including water quality monitoring. In this paper, we discuss the development and implementation of a digital twin for water quality monitoring that mimics sensory data integrated with a structural model of the water quality system to create a multipurpose simulation of water quality conditions. The model can be used to explore changes in water quality based on various scenarios. It employs establishing effective communication and coordination between computational processes, facilitating the utilization of waste inputs for conducting data analysis through machine learning techniques.**

***Index Terms*—Digital twin, Water contamination, Simulation, Machine learning, TCP, LSTM**

I. INTRODUCTION

The health and stability of human communities and societies depend critically on the availability of clean, safe water. Our ability to maintain clean, unpolluted water sources for drinking, domestic use, recreation, and agriculture is critical to the survival of our civilization. Unfortunately, these water bodies are particularly susceptible to contamination from various sources. Examples of common contamination sources include underground migration, fertilizer runoff from agricultural practices, septic tank leaks, stormwater runoff from streets, as well as the discharge of waste from industrial and municipal sources, among others. According to a recent study, water contamination alone was responsible for causing 1.4 million deaths in the year 2019. Each year, over 1 billion people fall ill due to waterborne pathogens present in contaminated drinking water, leading to the spread of diseases such as cholera, giardiasis, and typhoid. The impact of contamination is disproportionately severe on low-income communities, as they often reside close to the most polluting industries. A recent study conducted by the Natural Resources Defense Council discovered that between June 1, 2016, and May 31, 2019, there were 170,959 violations of the Safe Drinking Water Act (established in 1974) across 24,133 community water systems in the United States [1]. Alarmingly, nearly 40% of the U.S. population relied on drinking water systems that failed to comply with this law.

The current methods of detecting contamination, particularly in contexts like environmental monitoring, food safety, and water quality, are marked by significant inadequacies. Traditional methods predominantly rely on laboratory testing, which takes time and often identifies contaminants only after they have already caused harm. The absence of real-time or early detection systems allows contamination to spread unchecked, missing crucial early intervention opportunities. Thus, there is a pressing need for more advanced, rapid, and proactive detection technologies that can identify contaminants early, ideally before they pose a significant threat.

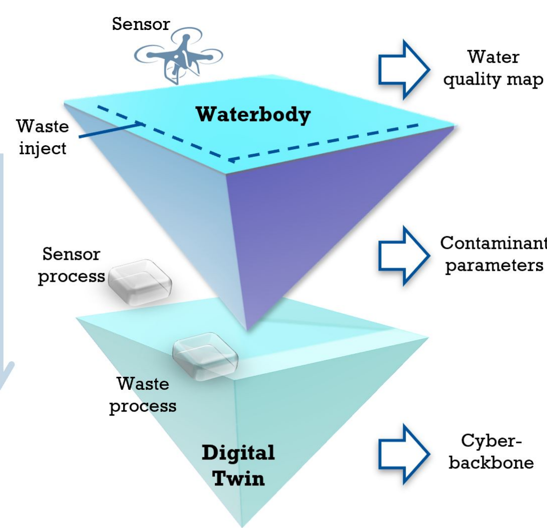


Fig. 1. A Vision for Continuous Water Quality Monitoring

Our goal is to develop an innovative system that can detect abnormal contamination in surface water bodies, explicitly focusing on rural-urban interfaces. To ensure the viability of the cyber-infrastructure, it is essential to thoroughly explore, test, and resolve any issues prior to its deployment. This involves establishing effective communication and coordination between sensors and data analysis, particularly in response to contamination caused by various wastes. To achieve this, we design a digital twin [2] framework AQUATWIN (**A**quatic **Qu**ality Analysis through Digit**a**l **Twin**), that is targeted toward a specific application using core concepts of Internet of Things (IoT). Fig. 1 shows the overall vision for the

Authorized licensed use limited to: University of Florida. Downloaded on January 24,2025 at 03:40:30 UTC from IEEE Xplore. Restrictions apply.

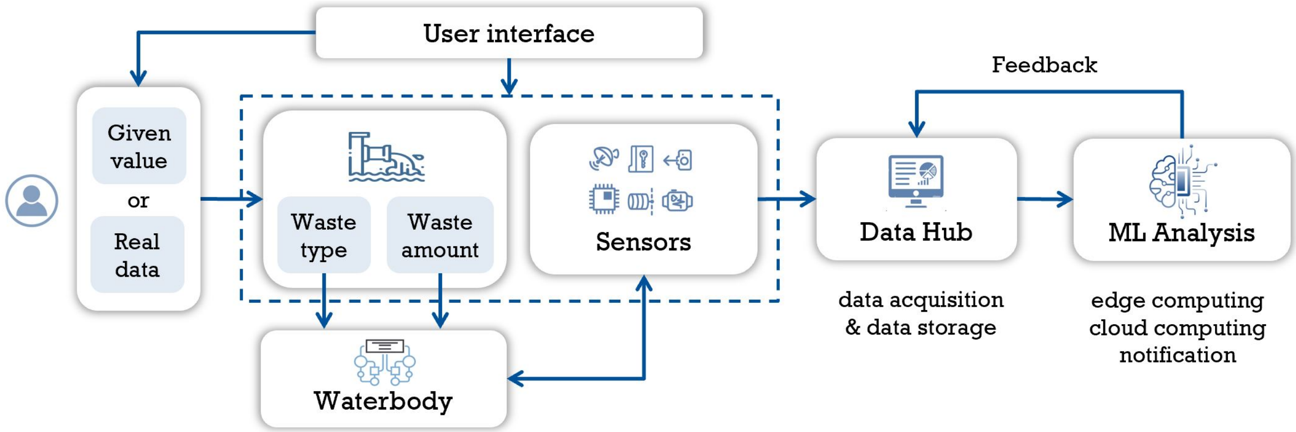


Fig. 2. AQUATWIN high-level architecture

framework. The system enables the exploration of various coordination mechanisms and the integration of relevant sensory systems, thereby providing an initial testing ground for the early exploration of the system.

The remainder of the paper is organized as follows. Section II provides the relevant background and summarizes related research in this area. Section III presents the AQUATWIN platform architecture and details of all components with cyber-backbone. In Section IV we discuss a use case implemented with simulation choices. In Section V, we discuss the machine learning analysis. We conclude in Section VI.

II. Background and Related Work

A. Digital Twins

Digital Twins are characterized as virtual prototypes or computer-modeled constructs that imitate, replicate, mirror, or serve as a "twin" to the existence of a tangible entity, which might encompass an object or a process [3]. However, Digital Twins extend beyond the boundaries of mere simulations or virtual models [4]. They represent an intelligent, dynamic digital replica of a physical entity, closely tracing the lifecycle of its physical counterpart in order to oversee and manage a range of functionalities. Digital twin frameworks applied to intelligent systems might be seen as a rebranding of the concept of "simulating your system design" [5]. Therefore, modeling and simulation are core parts of building this technology. Numerous significant industries, spanning from manufacturing and connected autonomous vehicles to healthcare, energy, urban planning, and beyond, are undergoing transformative shifts facilitated by Digital Twin technologies [6].

B. Related Work

The field of digital twins for water contamination detection remains relatively underexplored. A majority of existing efforts have concentrated either on evaluating environmental repercussions or water infrastructure-related aspects. Zekri *et al.* [7] proposed a framework based on principles of multi-agent systems and the digital twin paradigm to address the existing gaps within the loop of data acquisition and processing in water management. Conejos *et al.* [8] focused on implementing a dedicated digital twin for the water distribution phase. It served as a comprehensive solution to a wide array of challenges inherent in distribution network management, encompassing everything from the initial design stages to the intricacies of daily operational tasks. In another study, Wang *et al.* [9] concentrated on delineating a Cyber-Physical System (CPS) tailored for the oversight of key water quality parameters. As outlined by the authors, the CPS envisioned necessitates: a) the incorporation of sensors and a robust communication network, b) the integration of computing technologies to construct models, oversee data, and execute data analysis, and c) the incorporation of predictive control technologies. Abhijith *et al.* [10] introduced an innovative approach to water quality modeling that integrates realistic water demand models, with hydraulic solvers, and water quality solvers. Bartos *et al.* [11] presents the concept of *pipedream* — an all-encompassing simulation engine designed for real-time modeling and state estimation in both natural and urban drainage networks. This engine seamlessly integrates two core components: (i) an innovative hydraulic solver based on the one-dimensional Saint-Venant equations and (ii) a proficient Kalman filtering scheme responsible for dynamically updating hydraulic states using observed data.

However, it's worth noting that none of these platforms have thus far offered a dedicated simulation framework that is configurable and versatile enough for a seamless water quality exploration. The application of digital twin technology specifically in the domain of water contamination detection is an area that still holds substantial potential for further investigation and development.

III. AQuaTwin Architecture

AQUATWIN appears as a versatile digital twin platform that aims to develop a targeted prototyping solution for a water quality monitoring system. The emphasis here is on the coordination and communication of various components of the system rather than its physical characteristics. The platform

Authorized licensed use limited to: University of Florida. Downloaded on January 24,2025 at 03:40:30 UTC from IEEE Xplore. Restrictions apply.

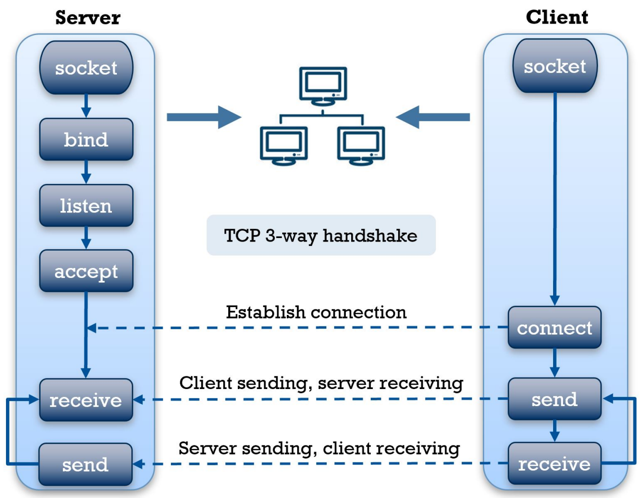


Fig. 3. Client-server socket flow for the cyber backbone

offers an infrastructure that supports the integration of any water-related model. Fig. 2 shows the high-level architecture of the platform.

A. *Cyber-backbone*

The communication system among the processes acts as the cyber backbone for real-time simulation. It is the initial foundation of our digital twin that establishes proper connections among all computational processes. The platform employs Transmission Control Protocol (TCP) sockets to facilitate interaction between all processes. TCP ensures reliable transmission of data packets over the network, albeit with slightly higher latency. The socket programming follows the client-server model (see Fig. 3), where the client initiates communication while the server waits passively to respond to the client's requests [12]. To transmit and receive messages through the socket, the data is encapsulated within a byte array frame.

Remark 1: (Implementation Note) We utilize TCP due to its ability to guarantee the successful transmission of data packets throughout the network, albeit with slightly higher latency. By employing TCP, we ensure secure and efficient data transfer among multiple hosts over an IP network. This protocol enables reliable and orderly communication, making it a preferred choice for ensuring the integrity and accuracy of transmitted data. Moreover, TCP is a better choice when using multiple devices for more computational capability.

B. *Computational Blocks*

The key elements of the Digital Twin functionality besides the cyber-backbone are the individual computational processes representing some extent of the real work elements in the contamination scenario. The primary components we simulate are as follows:

1) Waste Injection: Our platform provides software processes to produce artificially generated computational data simulating the behavior of waste injection on a water body. We utilize a wide range of waste types, to accurately simulate the process of waste injection. By employing different waste categories and quantities, we aim to enhance the fidelity and realism of our simulations. This approach allows us to accurately model the injection process and study its effects on the environment, enabling us to make informed decisions and implement effective waste management strategies. For simulation and visualization purposes, we will be demonstrating three waste types and three quantity categories.

2) Waterbody Flow: Our primary objective is to accurately simulate the flow of water body contamination rather than focusing on the intricate physics-based behavior of the water itself. To achieve this, we have developed an application-specific digital twin that specifically caters to contamination activities. In our simulation framework, we have created a dedicated water body computational block that incorporates both client and server models. This computational block is designed to showcase the progression and dissemination of contamination based on the inputs of waste materials.

3) Sensor: Sensors from the real world (*e.g.*, drones, mobile sensors, etc.) are represented by computational processes that sense and receive data from the waterbody process regarding waste information. The process then sends all the data to the Data Hub Process.

4) Data Hub: The Data Hub process serves as a pivotal node, facilitating coordination between the primary contamination process flow and the data analysis infrastructure. It efficiently handles incoming data from sensors, receiving it intermittently or in chunks. The received data is stored and transmitted to the edge device or cloud as per specific requirements, *i.e.*, for further analysis, machine learning predictions, and other relevant purposes.

C. *User Interface and Extensibility*

We make use of a front-end visualization as a GUI-based interface for users to explore the platform. Initially, the window gives the user options to play with the parameters of the water injection and sensor GUIs. An important feature of AQUATWIN lies in its robust support for *extensibility*, signifying the effortless integration of novel use cases and functionality (even those that may interface with the existing ones). To extend AQUATWIN with additional functionalities, the user needs to provide the following information:

- Waterbody process; add either 1) new types of waterbodies with distinct characteristics or 2) functionalities to the existing waterbody *i.e.*, making it multi-functional.
- Waste inputs; add more waste types, concentrations, and new domains to comprehend contaminants better
- Sensors; add sensors mimicking new components *e.g.*, buoy, drone, smart sensors, etc.
- ML models: employ new machine learning algorithms with the data hub process

The user interface manages all extensibility options including hardware interfacing and scheduling. After adding a new use case, the user can exercise several use cases simultane-

Authorized licensed use limited to: University of Florida. Downloaded on January 24,2025 at 03:40:30 UTC from IEEE Xplore. Restrictions apply.

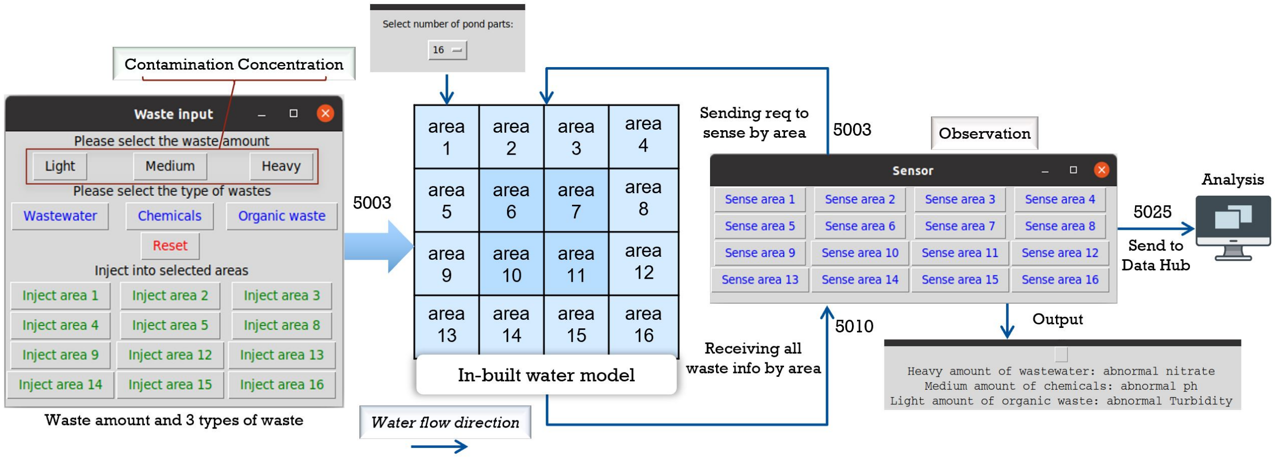


Fig. 4. Digital twin use case implementation

ously, as the cyber backbone can handle several interactions at the same time.

IV. Implementation and Use Cases

To comprehend potential scenarios, trends, and impacts of water contamination, we deploy various use cases. Each specific use case scenario allows us to explore relevant data, demonstrating the efficacy of machine learning predictions and numerical analysis in understanding these complex environmental challenges. This approach not only enhances our predictive capabilities but also aids in formulating more effective strategies for water quality management.

A. Simulation of Contaminant Dissipation

To estimate how long it will take for the contamination to reach nearby areas in the depicted pond, we would need to know the specific flow dynamics within the pond. If we assume a simplified model where the pond is divided into grid sections and the water (and thus the contaminant) moves uniformly from one section to the adjacent ones in a specific time unit, we could potentially apply a basic diffusion or advection model. However, without specific flow rates or diffusion coefficients, it is not possible to give an exact time. In reality, this process is governed by factors including the velocity of water flow, the mixing patterns of the waterbody (which can be affected by wind, temperature gradients, and shape), and the properties of the contaminant. Typically, this would be modeled using computational fluid dynamics (CFD) theory, which can simulate how substances move in a fluid medium. In the absence of detailed data and modeling capabilities, we use an estimated average flow rate (if known) to calculate the time it would take for the contaminant to travel the distance *e.g.*, between the center of *area 1* to the centers of *areas 2*, and *6*.

As we offer an infrastructure to support any exploration model, we present a simplified scenario, assuming that advection (the transport due to water flow) is the dominant mechanism and diffusion is secondary. We estimate the time it takes for the peak concentration of a contaminant to travel from area 1 to nearby areas using the following advection-based formula that is based on general principles of fluid dynamics, water quality analysis, and engineering practices commonly used in environmental engineering and water quality management to estimate contamination flow rates [13], [14]. We are using a simplified version *i.e.*, $t = v/d$ Where t is the time for the contaminant to reach the target area. d is the distance to the target area (*e.g.*, from *area 1* to *areas 2* and *6*). v is the average flow velocity (related to the water flow rate). This formula assumes that the contaminant travels with the bulk water flow. It doesn't account for the spreading or dilution of the contaminant, which would be influenced by the diffusion coefficient and the geometry of the water body.

B. Use Case Exploration

To discuss and demonstrate the functionality of AQuaTwin, we implement a specific use case involving three types of wastes (wastewater, chemicals, and organic waste) with various quantities. Primarily we present the waterbody as a square ranging from a 3x3 matrix to 5x5 *i.e.*, the user can select the size of the square. We operate under the assumption that waste disposal is permissible in any part of the water body, except for *areas 6, 7, 10* and *11*. This restriction is due to their central locations within the water body, as opposed to peripheral or edge areas, where waste dumping is typically more feasible (see Fig. 4). Let us assume the water flow is constant and directed from left to right, which facilitates the dispersion of waste toward the areas on the right side and their adjacent regions. The following actions occur to simulate a certain scenario:

- From the GUI selection, we choose a 4x4 square where the whole water body is divided into 16 equal areas.
- The computational process representing the waste input generates a GUI from which the user sends wastewater of heavy amounts, chemicals of medium amount, and organic waste of light amount to *area 1* of the water body process via TCP port 5003.

Authorized licensed use limited to: University of Florida. Downloaded on January 24,2025 at 03:40:30 UTC from IEEE Xplore. Restrictions apply.

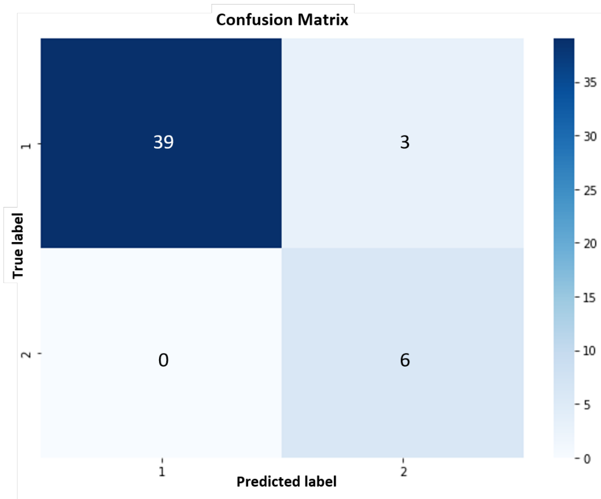


Fig. 5. Confusion matrix for the LSTM model

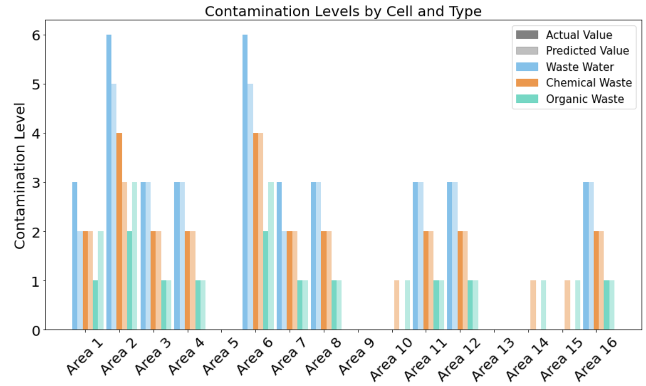


Fig. 6. Bar graph for the comparison of actual versus estimated waste levels

- After a specified duration t, the waste begins to disperse towards *areas 2* and *6*.
- Following the same time interval, the waste further spreads to *areas 3, 7,* and *11*.
- The dispersion eventually ends by reaching *areas 4, 8, 12,* and *16*.
- Throughout this whole waste flow process, the concentration of waste in each area varies over time. Computational processes, which emulate sensors, can monitor any area at any given moment to assess contamination levels and provide graphical user interface (GUI) outputs via ports 5003 and 5010.
- The sensor then can send data to the datahub for machine learning and numerical analysis via port 5025.

Note that, our use case takes into account the direction of water flow. Consequently, some areas, particularly those on the opposite side of the flow direction, are not experiencing contamination as a result.

V. Machine Learning Analysis

One of the primary objectives of developing AQuaTwin is to analyze trends in contamination flow and forecast similar patterns under varying conditions. To accomplish this, we delved into a range of machine learning techniques and numerical analyses, enabling us to determine the most effective approach for our predictions. Eventually, we employed time series analysis to understand trends and patterns in the contamination levels.

A. Data Pre-processing

By simulating various waste input scenarios, we developed a time-series dataset, comprising 27 combinations of waste levels (high, medium, low) distributed across permissible waterbody areas. To understand the trend in contamination flow over time within these areas, we have developed an LSTM (Long Short-Term Memory) model, utilizing a training dataset specifically designed for this purpose. In our dataset, contamination levels range from 0 to 9. These levels are categorized into three classes: values 0 to 3 indicate low contamination (class 1), 4 to 6 signify medium contamination (class 2), and 7 to 9 denote high contamination (class 3).

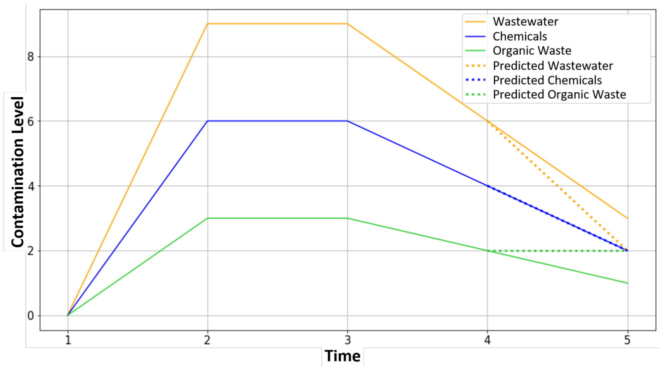
B. Model Training and Analysis

We approach the analysis of contamination trends from two key perspectives: as a classification problem, where the model's objective is to identify the contamination level of each area (high, medium, or low) for each type of waste, and as a forecasting problem, where the model predicts future contamination levels for each type of waste in each area based on their historical data. Fig. 5 illustrates the confusion matrix of our LSTM model. This particular model was tested by disposing of a numerical value representing various types and concentrations in *area 1*, achieving an accuracy of 93%. Fig. 6 presents a comparison of actual versus estimated waste levels for each waste type across 16 areas.

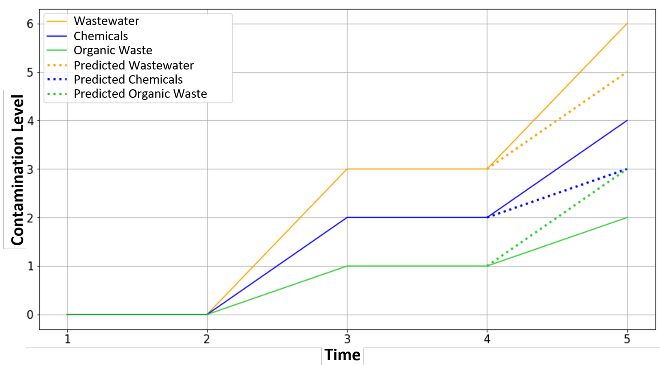
To provide a more focused analysis, we have particularly highlighted the trend of contamination levels (both actual and predicted values for each waste type) over time in *areas 1, 2*, and *6* (see Fig. 7). These areas were selected for detailed observation as they are the most impacted by waste disposal in *area 1*. We present the trends of actual and predicted contamination levels over time for these areas, demonstrating how the model's forecasts align with the observed data.

Remark 2: (Algorithm Choice) We use LSTM because time series data can be of varying lengths, and LSTMs can handle these sequences effectively, which is a common scenario in environmental data. Traditional recurrent neural networks (RNNs) suffer from the vanishing gradient problem, where gradients shrink as they are propagated back through time, making it hard to learn from data points that are far apart in time. LSTMs are designed to *remember* long-term dependencies by using their internal memory cells [15]. This is particularly useful in time series data of our waterbody process where the relevance of previous events can extend far back into the past.

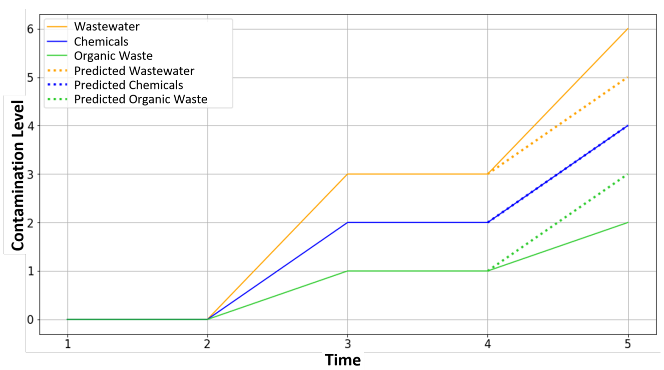
Authorized licensed use limited to: University of Florida. Downloaded on January 24,2025 at 03:40:30 UTC from IEEE Xplore. Restrictions apply.



(a) Area 1



(b) Area 2



(c) Area 6

Fig. 7. Trend of contamination flow over time in three areas for both actual and predicted values for each waste type

VI. Conclusion and Future Work

We developed an innovative digital twin platform dedicated to water quality monitoring, addressing the crucial need for access to clean and safe water. AQUATWIN facilitates the user to tweak various parameters and experiment with different quantities and types of waste across distinct areas of a pond. We further explored a use case and created a dataset from a simplified combination of waste information to enable predictive analysis using a machine learning algorithm.

In our future work, we aim to substantially enhance the overall system by incorporating real data using drones and static sensors to gather specific parameters like water flow rates, diffusion coefficients, and temperature. With this precise data, our improved model will be able to more accurately simulate the mixing patterns within water bodies, taking into account both advection and diffusion processes in detail. Moreover, we plan to upgrade AQUATWIN by seamlessly integrating it with sensor networks across diverse aquatic environments, leveraging cloud services for robust data management and analysis. This will enable real-time detection and allow our predictive analysis to feed back into the system dynamically, enhancing both the accuracy and responsiveness of our environmental monitoring efforts. Furthermore, AQUATWIN serves as a representative digital twin for exploring environmental issues. Therefore, the core concept of this platform can be utilized to develop other related platforms aimed at investigating a variety of environmental challenges *e.g.*, climate change, biodiversity loss, etc.

References

[1] K. P. Fedinick, S. Taylor, M. Roberts, R. Moore, and E. Olson, “Watered down justice,” *Natural Resource Defense Council Report*, 2019.

[2] D. Wagg, K. Worden, R. Barthorpe, and P. Gardner, “Digital twins: State-of-the-art and future directions for modeling and simulation in engineering dynamics applications,” *ASCE-ASME J Risk and Uncert in Engrg Sys Part B Mech Engrg*, vol. 6, no. 3, 2020.

[3] B. R. Barricelli, E. Casiraghi, and D. Fogli, “A survey on digital twin: definitions, characteristics, applications, and design implications,” *IEEE access*, vol. 7, pp. 167 653–167 671, 2019.

[4] S. Boschert and R. Rosen, “Digital twin—the simulation aspect,” in *Mechatronic futures*. Springer, 2016, pp. 59–74.

[5] S. Mihai, M. Yaqoob, D. V. Hung, W. Davis, P. Towakel, M. Raza, M. Karamanoglu, B. Barn, D. Shetve, R. V. Prasad *et al.*, “Digital twins: a survey on enabling technologies, challenges, trends and future prospects,” *IEEE Communications Surveys & Tutorials*, 2022.

[6] M. R. Kabir and S. Ray, “Virtual prototyping for modern internet-of-things applications: A survey,” *IEEE Access*, vol. 11, pp. 31 384–31 398, 2023.

[7] S. Zekri, N. Jabeur, and H. Gharrad, “Smart water management using intelligent digital twins,” *Computing and Informatics*, vol. 41, no. 1, pp. 135–153, 2022.

[8] P. Conejos Fuertes, F. Martínez Alzamora, M. Hervás Carot, and J. Alonso Campos, “Building and exploiting a digital twin for the management of drinking water distribution networks,” *Urban Water Journal*, vol. 17, no. 8, pp. 704–713, 2020.

[9] Z. Wang, H. Song, D. W. Watkins, K. G. Ong, P. Xue, Q. Yang, and X. Shi, “Cyber-physical systems for water sustainability: challenges and opportunities,” *IEEE Communications Magazine*, vol. 53, no. 5, pp. 216–222, 2015.

[10] G. R. Abhijith, D. B. Steffelbauer, and A. Ostfeld, “Toward digital twins for emerging contaminants in water distribution systems,” in *World Environmental and Water Resources Congress 2023*, pp. 1047–1057.

[11] M. Bartos and B. Kerkez, “Pipedream: An interactive digital twin model for natural and urban drainage systems,” *Environmental Modelling & Software*, vol. 144, p. 105120, 2021.

[12] R. L. R. Maata, R. Cordova, B. Sudramurthy, and A. Halibas, “Design and implementation of client-server based application using socket programming in a distributed computing environment,” in *2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. IEEE, 2017, pp. 1–4.

[13] C. J. Newell, *Calculation and use of first-order rate constants for monitored natural attenuation studies*. US Environmental Protection Agency, National Risk Management Research Laboratory, 2002.

[14] M. Gröshbacher, D. Eckert, O. A. Cirpka, and C. Griebler, “Contaminant concentration versus flow velocity: drivers of biodegradation and microbial growth in groundwater model systems,” *Biodegradation*, vol. 29, no. 3, pp. 211–232, 2018.

[15] S. A. Mila, M. R. Kabir, and S. Ray, “Continuous monitoring for diagnosis: A wearable for covid-19 detection in symptomatic patients,” in *2023 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 2023, pp. 01–06.

Authorized licensed use limited to: University of Florida. Downloaded on January 24,2025 at 03:40:30 UTC from IEEE Xplore. Restrictions apply.