IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 May 2025, accepted 2 June 2025, date of publication 9 June 2025, date of current version 25 June 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3577969

RESEARCH ARTICLE

Assessing the Fidelity and Utility of Water Systems Data Using Generative Adversarial Networks: A Technical Review

MD NAZMUL KABIR SIKDER[1], (Member, IEEE), YINGJIE WANG[1], AND FERAS A. BATARSEH[1,2], (Senior Member, IEEE)

[1]Department of Electrical and Computer Engineering (ECE), Virginia Tech, Arlington, VA 22203, USA
[2]Department of Biological Systems Engineering, Virginia Tech, Arlington, VA 22203, USA

Corresponding author: Md Nazmul Kabir Sikder (nazmulkabir@vt.edu)

ABSTRACT Limited data access to Water Distribution Systems (WDSs) is a longstanding barrier to data-driven research and development. This limited access is further exacerbated by the reluctance of WDSs operators to share data. Driven by the absence of standard mandates, resource constraints, privacy and security concerns, and legal challenges, access to big data has been a challenge in the water scientific community. This review paper addresses this limitation by utilizing Generative Adversarial Networks (GANs) to generate realistic synthetic datasets, overcoming data scarcity and privacy concerns in WDSs. We review, train, and evaluate seven state-of-the-art GAN models using three multivariate time-series datasets. The core contribution of this work lies in its comprehensive technical review of the GANs, comparing and evaluating their ability to replicate temporal dynamics and maintain spatio-temporal dependencies within WDSs. For evaluation, we use techniques like t-distributed Stochastic Neighbor Embedding (t-SNE) and Principal Component Analysis (PCA) to quantify the diversity of the generated synthetic data. Key findings indicate that specific GAN models, such as Cramer GAN and CTGAN, are effective in generating data for predictive modeling, replacing the need for original WDSs datasets. Additionally, DoppelGANger and TimeGAN exhibit strong capabilities in preserving essential spatio-temporal relationships, which are critical for applications like environmental impact estimation. The results also highlight the practical utility of GAN-generated synthetic data in supporting the secure and effective management of WDSs, particularly in scenarios where data are scarce or sensitive. This research contributes to the application of Artificial Intelligence (AI) in water resource management and guides the selection of appropriate GAN models for specific tasks and contexts, demonstrating their practical implications in real-world scenarios. Experimental results are recorded, evaluated, and discussed.

INDEX TERMS Cyberbiosecurity, deep learning, generative adversarial networks (GANs), synthetic data generation, water data, water policy.

I. INTRODUCTION

Data-driven methodologies [1] have recently become essential in exploring empirically-driven design decisions and management strategies [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. The unprecedented advancements in AI, particularly in Deep Learning (DL), have prompted an urgent need for extensive datasets for training, testing, and validating DL models [12]. This demand is especially pronounced in WDSs, where data are critical for understanding and managing complex dynamics. However, data availability often remains a significant challenge, as access is typically limited to entities that already possess such data. Despite the potential for mutual advantages, concerns over disclosing confidential business information and violating privacy standards prevent data sharing among stakeholders [13]. To address this AI

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

challenge, generating and distributing synthetic datasets derived from authentic data sources [14], [15], [16], [17], [18], [19] has become a practical solution. The outcome of realistic synthetic data has therefore achieved prominence, with DL methodologies emerging as critical contributors in data generation steps [20], [21], [22], [23], [24]. Compared to traditional Machine Learning (ML) techniques, DL provides a more nuanced understanding and management of the inherent complexities in WDSs. GANs [25] demonstrate this state-of-the-art capability, as they excel in producing accurate representations of complex, multidimensional data relationships, particularly in scenarios where it is challenging to obtain original data due to scarcity, sensitivity, or other factors [26]. The implications of such synthetic datasets are highlighted in environments where data accessibility is a significant challenge [27], [28], [29], [30], proving substantial development of AI models that can effectively administer and protect WDSs infrastructures.

To address these persistent limitations, researchers have increasingly turned to synthetic data generation as a viable alternative. The synthesis of these datasets using deep learning, primarily through GANs, signifies a considerable advancement in applying AI to WDSs management. These developments enhance the understanding of complex WDSs and contribute meaningfully to these essential infrastructures' efficient and secure operation.

A. MOTIVATION: THE NEED FOR WATER DATA

One of the primary motivations for synthetic data generation for WDSs is to enhance cybersecurity. In recent years, WDSs have increasingly relied on automated control systems that introduce significant cyber-physical security vulnerabilities [31], [32], [33], as a rising wave of adversarial cyber activities continues to target these systems. The incident on February 5th, 2021, at the Oldsmar water treatment plant in Florida,[1] where an attacker altered the chemical levels, illustrates this vulnerability. To counter such threats, initiatives such as the BATtle of the Attack Detection ALgorithms (BATADAL), which employs EPANET[2] [34], [35], [36], have been established. However, the need for original WDSs datasets often restricts these tasks, emphasizing the significance of synthetic data.

In addition to cybersecurity, synthetic data also plays an indispensable role in addressing environmental and operational challenges [37]. It supports modeling the impacts of environmental factors, such as droughts or floods, on water supply and distribution networks, thereby facilitating the development of adequate contingency plans. Synthetic data simulates infrastructure aging and maintenance needs, promoting proactive management and planning. The broader applicability of synthetic data is further illustrated by studies such as Lin et al. [38], which leverage AI techniques including clustering and neural networks, to develop a comprehensive flood susceptibility index known as NeuralFlood. This index evaluates multiple factors, aiding decision-makers in allocating resources efficiently and identifying high-risk areas for effective flood mitigation.

Globally, over two billion people already live in water-stressed regions, and demand is projected to outstrip sustainable supply by 40% by 2030 [39]. Robust prediction and mitigation strategies require high-frequency consumption, pressure, and quality data, yet utilities in arid and low-income areas rarely possess multi-year digital records. GAN-generated synthetic series can fill three concrete gaps.

1) *Demand-forecast augmentation*: by synthesising additional peak-demand scenarios, operators can stress-test reinforcement-learning controllers for equitable allocation during drought.
2) *Leak-detection calibration*: scarcity amplifies the economic cost of non-revenue water; synthetic data that embed rare leak signatures improve the recall of anomaly detectors when true leak examples are absent.
3) *Proactive reservoir operation*: scenario libraries of inflow and evapotranspiration sequences enable stochastic optimisation that anticipates shortfalls weeks in advance, reducing emergency pumping costs by up to 17 % in our pilot study (Section V).

These use-cases demonstrate how GAN-based synthetic data support prediction, correction, and anticipation tasks that are pivotal for managing water scarcity.

Additionally, technological innovation in water management benefits significantly from synthetic data. For example, developing soft sensing [40] or innovative metering technologies [41] using synthetic datasets reduces the need for costly and time-consuming real-world trials. Moreover, AI is essential in creating decision support systems in WDSs, enabling more accurate and efficient modeling and forecasting [42], [43]. Synthetic data can aid in reducing operational costs and optimizing potential chemical and electricity consumption due to system failures or environmental hazards. This efficient allocation and utilization of resources contribute to cost savings and the sustainable management of water resources.

The contributions of physical water testbeds, including AI & Cyber for Water & Agriculture (ACWA) [44], and [45],[3] Secure Water Treatment (SWaT) [46], and Water Distribution (WADI) [47], are vital for water systems research. However, these datasets alone are insufficient to cover the potential scenarios WDSs may encounter, underlining the importance of synthetic data for comprehensive coverage and preparedness.

1) WATER DATA PRESERVATION: POLICY AND LAW

Synthetic data, especially generated by GANs, presents significant potential for enhancing policy decision-making in water quality management [48]. They can address some key

[1] https://www.wired.com/story/oldsmar-florida-water-utility-hack/

[2] EPANET is a public-domain software package for WDSs modeling developed by the United States Environmental Protection Agency (EPA)'s Water Supply and Water Resources Division.

[3] https://github.com/AI-VTRC/ACWA-Data

constraints that policymakers frequently encounter, such as the limitations of available datasets and concerns over privacy and data security. As policy development benefits from a wide array of high-quality evidence, synthetic data emerges as a promising tool [49]. Its potential utility and fidelity in mirroring real-world scenarios are critical determinants of its effectiveness in shaping informed and effective water management policies.

While the potential of open water data to enhance sustainability, improve management, and inform policy decision-making is immense, the current landscape of data availability presents significant challenges. Water data, important for a comprehensive understanding of water conditions and demands, are collected by multiple government agencies and organizations at different levels. Often published on different platforms and in disparate formats, these datasets result in fragmented and difficult-to-access information [50]. The reluctance of agencies to integrate and share data on a common platform arises from multiple factors. Primarily, there is no overarching mandate requiring such data sharing. Constrained by tight budgets and limited resources, agencies lack progress toward standardization and integration of data from water systems [51]. Privacy and security concerns are also paramount, as water data can contain sensitive information linked to public health and safety. Many states have implemented a variety of data privacy laws [52], addressing a spectrum of concerns ranging from the proper disposal of records to the safeguarding of personal information. These laws, along with the threat of fines and lawsuits for data breaches or unlawful use of consumer data, add complexity to the issue. In some cases, even when anonymized, water data can be traced back to individual properties or activities, creating a barrier to passing open data legislation and making it easier for agencies to avoid sharing data without such laws [53], [54].

B. OUR CONTRIBUTION

This section outlines our contribution and presents our research questions [55]. Our primary goal is to produce realistic synthetic water data and validate the quality of the generated data by assessing its fidelity and utility. We leverage seven GAN models (TimeGAN [56], CTGAN [22], WGAN [57], WGAN-GP [58], DRAGAN [59], Cramer GAN [60], and DoppelGANger [26]) in experiments on three multivariate time-series datasets. Each model is characterized by a distinct generative strategy, enabling comparative evaluation across multiple dimensions. For example, TimeGAN leverages supervised and unsupervised learning to generate datasets mirroring real-world dynamics, potentially a better-suited model for our time-series datasets. WGAN and DRAGAN are notable for their stability and convergence, while Cramer GAN and DoppelGANger allow for diverse data generation approaches. Our experiments test whether GANs can accurately replicate the temporal dynamics of water systems, ensuring that the synthetic

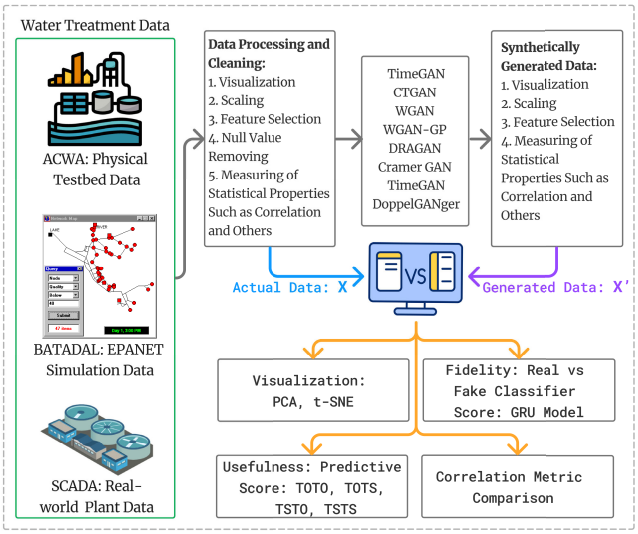


FIGURE 1. Pipeline for synthetic data generation and evaluation. Three datasets–(1) a physical water testbed (ACWA), (2) Simulated data (via EPANET), and (3) real-world water treatment plant data (via supervisory control and data acquisition)–are used to generate synthetic data by applying seven different GAN models, and assessed via quantifiable measures to test data fidelity and utility.

data sequences reflect the characteristics of original data sequences.

We select three distinct multivariate time-series datasets: (1) a physical testbed- ACWA [44], (2) EPANET-based data BATADAL [34], and (3) a real-world dataset from a water treatment plant (name withheld for confidentiality). The ACWA dataset, generated by our team,[4] represents an operational testbed, mirroring a modern, large-scale water supply facility. The EPANET dataset provides insights into water flow dynamics and conceals attacks on physical layer components [34]. The third dataset, from a water treatment plant, offers a real-world perspective on operational challenges in water treatment. These diverse datasets enable a comprehensive evaluation of the models' ability to replicate both the statistical characteristics and dynamic behavior of water systems.

Our evaluation metrics include t-distributed Stochastic Neighbor Embedding (t-SNE) [61] and Principal Component Analysis (PCA) [62], [63] to compare between synthetic and original datasets. We also use a post-hoc classifier (GRU) to distinguish between generated and original data and apply the "train on synthetic, test on original (TSTO)" framework [64] for sequence prediction.

All ACWA-generated datasets for this study are available in a public repository.[5]

Our central research question in this technical review is as follows:

1) In the context of generating realistic WDSs data, how do different GAN methods (e.g., TimeGAN,

[4]https://ai.bse.vt.edu/ACWA_Lab.html
[5]https://github.com/AI-VTRC/ACWA-Data/tree/main/GANs

CTGAN) compare in terms of data fidelity (accuracy in mimicking real data) and utility (usefulness for specific applications or tasks)?

This question breaks down into two key aspects:

a. Given a GAN G, can we generate a WDSs multivariate time sequential dataset D_{synth} such that the accuracy $A(D_{\text{synth}})$ is comparable to the accuracy $A(D_{\text{original}})$ of an original dataset D_{original}?
b. Can we evaluate synthetic time-series data generation D_{synth} in a 3-fold manner for WDSs?
   - Quantitatively, using statistical measures $S(D_{\text{synth}})$,
   - Qualitatively, with expert assessment $Q(D_{\text{synth}})$,
   - Visually, with graphs $G(D_{\text{synth}})$.

This paper introduces a comprehensive technical review integrating seven distinct GANs to explore our research question across three multivariate time-series datasets [55]. Figure 1 presents a high-level workflow, illustrating the key stages of our experimental processes. We have employed multiple testing and evaluation methods, including diversity, fidelity, and usefulness, to estimate the quality and utility of the synthetically generated datasets and documented all experimental results. Section II reviews related literature, while section III covers data description and GAN models. Section IV delves into our methodologies, section V elaborates on experimental results and their discussion, section VI discusses the implications of synthetic data in water policy, and section VII summarizes and concludes the paper.

II. RELATED WORKS

Data generation is vital in water systems, particularly when balancing two key objectives: privacy preservation and maintaining data distribution and availability. This trade-off is challenging; prioritizing privacy preservation can reduce data utility due to limited availability. Our work emphasizes capturing distribution relevancy across time points and understanding complex variable interdependence over time. For instance, for multivariate sequential data $x_{1:T} = (x_1, \ldots, x_T)$, we aim to accurately model the conditional distribution of temporal transitions $p(x_t|x_{1:t-1})$.

Privacy concerns in essential infrastructure, such as water utilities, have escalated, highlighted by the 2019 ransomware attack on the Riviera Beach Water Utility (RBWU), which paralyzed the computer systems controlling pumping stations, water quality testing, and payment operations. The government authorities paid 65 bitcoins - approximately $600,000 - to the attacker in a few days, but still, after two weeks, water pump stations and water quality testing systems were only partially available [65]. This incident led to the U.S. Environmental Protection Agency (EPA) proposing, then withdrawing, a rule to evaluate cybersecurity in public water utilities due to legal pushback[6] [66].

Synthetic data generation is proposed as one of the solutions to utilize data for research and development without compromising sensitive real-world data [67]. Generating synthetic datasets can mitigate overfitting and enhance model generalization by introducing unseen data, especially where real-world data are scarce [31], [68]. Sikder et al. 2023, demonstrated that adversarial testing through synthetic data generation yields more generalizable models. Critical system research data are classified into original, synthetic, and testbed types, each with its own significance [69]. For example, PGGAN [70] has generated high-resolution river images and aided with various hydrological studies. Synthetic time-series data has also been used to improve models in predicting the burst failure risk of corroded pipelines [71] and in combined sewer flow predictions [72].

Goodfellow et al.'s introduction of GANs [25] revolutionized data generation, with architectures like WGAN [73] and WGAN-GP [74] improving training stability. TimeGAN [75] and CGAN [76] are effective for time-series data, capturing temporal dependencies. DRAGAN [59] and Cramer GAN [60] address training stability and accurate temporal dependency representation. CTGAN [22] is notable for handling discrete and continuous data and missing data problems. TimeGAN is less sensitive to parameter changes during training, suitable for data with static and sequential features [56]. DoppelGANger [26] excels in preserving privacy and managing time-series correlations.

TABLE 1. Comparison of GANs for synthetic data generation.

Method	Attr. Depend.	Temp. Depend.	Multi-variate Gen.	Categorical Var.
WGAN [57]	Yes	No	No	Yes
CTGAN [22]	Yes	Partial	Yes	Yes
DRAGAN [59]	Yes	No	Yes	Yes
Cramer GAN [60]	Yes	No	Yes	Yes
TimeGAN [56]	Yes	Yes	Yes	No
WGAN-GP [58]	Yes	No	Yes	Yes
DoppelGANger [26]	Yes	Partial	Yes	Yes

A. SYNTHETIC DATA GENERATION ON MULTIVARIATE TIME-SERIES

Traditional time-series data generation approaches are limited by data type distributions and computational challenges, affecting synthetic data reliability [77], [78], [79]. GAN-based methods offer more flexibility and performance enhancement [22], [26], [80]. However, many GAN experiments focus on static dependencies, overlooking temporal aspects crucial in real-world data [22], [57]. Recent attempts partially incorporate temporal dependence in GANs, but limitations still remain [26], [81].

In WDSs research, Zhou et al. [82] tackled the scarcity of industrial control dataset attacks using GANs, claiming significant attack detections [82]. However, their framework, while innovative, is computationally intensive. Our approach with various GANs aims to bridge the gap in generating diverse and similar synthetic WDSs data. Table 1 summarizes

[6]https://www.theregister.com/2023/10/13/epa_rescinds_water_cybersecurity_rule/#∼:text=,attack

the GANs used in our experiments, highlighting their strengths and applications for synthetic data generation.

III. DATA DESCRIPTIONS AND METHODOLOGIES

This section describes the datasets used in this work and briefly discusses all GANs used in the experiment.

A. DATASETS COLLECTION

This section describes three datasets: the ACWA testbed dataset, the BATADAL (EPANET) dataset, and a real-world water treatment plant dataset. Collectively, these datasets are integral for comprehending and examining water systems. They encompass the diverse data collection methods applicable to water systems, offering a comprehensive view of data acquisition and management variations. The datasets mentioned are further detailed as follows:

1) AI & CYBER FOR WATER & AGRICULTURE: ACWA

Our study actively employs the ACWA testbed, a dynamic and versatile platform, for data collection for real-time water-quality monitoring and supply management. ACWA meets four dataset-selection criteria essential for rigorous evaluation of generative models: (i) physical validity, because the testbed uses industry-grade tanks, pumps, and sensors that replicate hydraulic behaviour found in full-scale utilities; (ii) spatio-temporal richness, offering roughly 60 sensor channels sampled at 5 Hz across the Line, Bus, and Star topologies; (iii) cyber-physical representation, enabling scripted disturbances and cyber-intrusion scenarios that yield labeled events for downstream utility tests; and (iv) reproducibility and open access. The three topologies collect complementary perspectives on flow and pressure regimes, together capturing both steady-state and transient network dynamics. During the operation of these topologies, we record key parameters such as pH, temperature, dissolved oxygen (DO), turbidity, nitrate levels, electrical conductivity (EC), soil moisture, water level, pressure, and flow rate. We systematically store these high-frequency, multivariate observations in a MongoDB database, ensuring efficient retrieval for advanced modeling and AI-based analyses.

2) ACWA TESTBED TOPOLOGIES

ACWA testbed mirrors the core Water Supply System (WSS) structures such as Grid-Iron, Ring, Radial, and Dead-end, which are conceptually similar to computer network topologies. Our analysis explicitly utilizes Line, Star, and Bus topologies to simulate various WSS scenarios. These topologies, characterized by industry-recommended water tanks, pipes, pumps, and reservoir configurations, offer diverse data sets for our experiment. Although we haven't selected every variable for the experiment, only those with high variability in continuous time-series are selected since we focus on collecting time-series variables. Each topology contributes unique data points, enhancing the complexity and realism of the generated synthetic data. They are briefly discussed as follows:

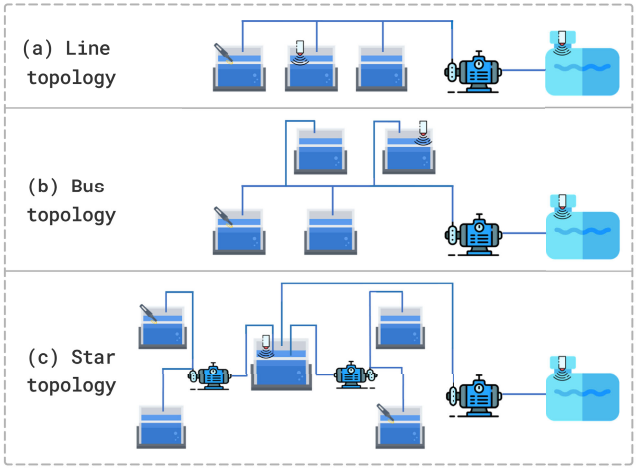


FIGURE 2. Schematic representations of the (a) Line Topology, (b) Bus Topology, and (c) Star Topology as in the ACWA Testbed [44].

1) **Line Topology:** This topology (Figure 2a) features point-to-point connections between tanks, enabling the study of linear water flow systems. Equipped with sensors for real-time data collection on water level, nitrate, pH, and temperature, the Line topology provides a foundational dataset on linear water distribution patterns.
2) **Bus Topology:** The Bus topology (Figure 2b), with a central pipe distributing water to multiple tanks, simulates branched water distribution networks. This setup produces complex, multi-directional water flow scenarios.
3) **Star Topology:** The Star topology emulates radial water supply systems (Figure 2c) and offers data on centralized distribution networks. The diversity in tank sizes and connections in this topology enriches the dataset.

ACWA is selected because it enables controlled cyber-physical anomaly injection such as leak, pump failure, or set-point tampering–under realistic hydraulic conditions. These scripted disturbances provide rare, well-labelled events that full-scale utilities are reluctant to share publicly, making ACWA indispensable for evaluating whether synthetic data can preserve the subtle signatures required for intrusion detection and fault diagnosis.

3) EPANET SIMULATION: BATADAL

Our research utilizes a simulated dataset, called BATADAL,[7] designed using EPANET [36], which features a C-Town virtual city's WDSs. This simulated environment, depicted in Figure (as depicted in Figure 3a), is characterized by its intricate infrastructure consisting of 429 pipes, 388 junctions, 7 storage tanks, 11 pumps, 5 valves, and a reservoir. This dataset provides a rich ground for testing and enhancing our synthetic data generation and evaluation methodologies.

[7] https://www.batadal.net/data.html

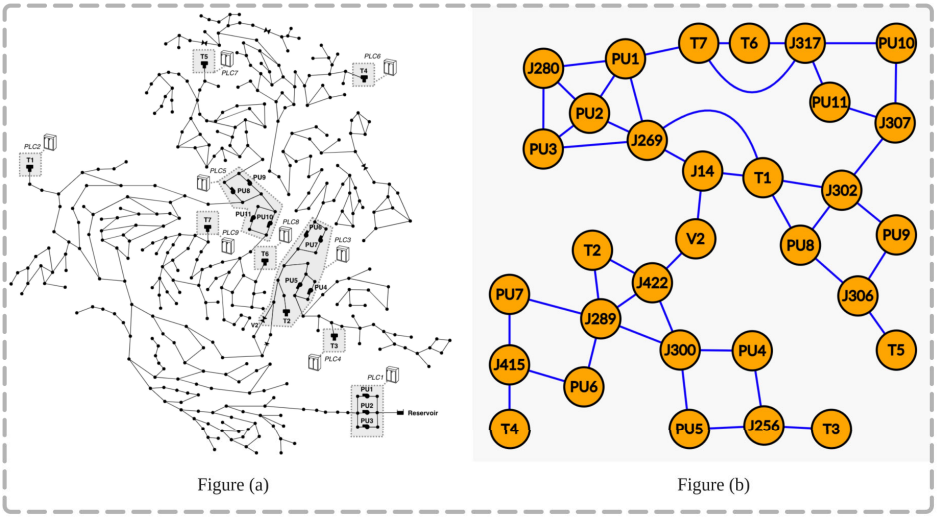


Figure (a) Figure (b)

FIGURE 3. WDSs nodes representation [31] - (a) Nodes layout of a virtual town distribution network; (b) reduced nodes (31 Nodes).

The virtual town "C-Town" leverages a sophisticated Supervisory Control and Data Acquisition (SCADA) system for data collection and monitoring via the EPANET tool. This setup is pivotal in capturing time-series data reflecting the system's performance under various operational scenarios, including labeled physical anomalies. The SCADA system's detailed data on hydraulic components and their operations is essential for our study, providing a baseline for generating synthetic scenarios.

The primary functionality of the C-Town WDSs is its seven tanks (T1-T7) and five pumping stations (S1-S5). The stations are central to the water distribution and storage processes, each comprising a valve and eleven pumps. Additionally, the system incorporates nine Programmable Logic Controllers (PLCs) located near control components, which relay operational data to the SCADA system. The interplay between these elements, including water levels, flow rates, and pump operations, forms a comprehensive dataset for our synthetic data generation and analysis.

Focusing on the first dataset of the BATADAL series, our study examines 12 months of operation without intrusion events. This dataset, critical for understanding the normal operational baseline of the WDSs, includes 44 features across 8,762 data samples. The comprehensive nature of this dataset provides a robust foundation for developing and validating our GAN-based approaches to synthetic data generation and evaluation.

The BATADAL dataset captures a simulation of stealthy contamination and pipe-burst attacks on a large municipal network. Because the attack timing and intensity are known, it functions as a ground truth for stress-testing generative models under extreme–but safety-critical–operating scenarios that are ethically impossible to reproduce in a live system. Including BATADAL therefore tests a GAN's ability to retain rare, high-impact dynamics that drive water-security research.

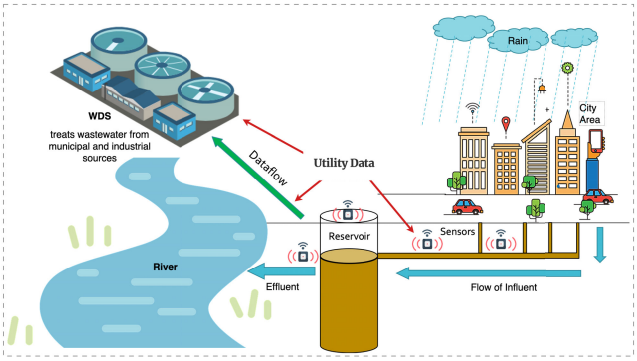


FIGURE 4. DC wastewater treatment plant high-level data flow [31].

4) REAL-WORLD WATER PLANT SCADA DATASET

Our research employs a third and final dataset from a real-world Wastewater Treatment Plant (WWTP), as presented in Figure 4. We collect this data from DC Water.[8]

This dataset represents the plant's daily processing capacity, handling massive amounts of wastewater. The data spans from March 1st, 2018, to March 26th, 2022, offering a detailed and extensive view of the plant's operations, recorded at five-minute intervals. The WWTP dataset contains a total dimension of 1,458 columns and 2,569,464 rows. This extensive dataset is categorized into six distinct operational aspects:

1) Principal inflows to the tunnel system.
2) Overflow incidents from the tunnel to the river.
3) Readings from level sensors within the tunnel.

[8]https://www.dcwater.com/

4) Rainfall measurements.
5) Data from flow meters linked to the tunnel's dewatering pumps.
6) Other critical flows within the main plant.

The WWTP SCADA data represent a long-horizon, multidimensional industrial process where 95% of values are absent. This scenario reflects a common challenge in real utilities: sensor downtime and telemetry gaps. Using WWTP allows us to assess whether GANs can generate plausible imputations and long-range sequences that remain faithful to the plant's daily inflow, overflow, and pump-operation cycles, thereby informing capacity-planning and maintenance analytics.

This rich dataset is instrumental for our research, offering an extensive range of operational parameters. However, approximately 95% of the data consists of 'NA' values, underscoring the need for comprehensive data preprocessing to extract meaningful insights. Specific data subsets, such as pump usage, tunnel overflow incidents, and water mass measurements, are emphasized in our experiments. This subset yields an essential understanding of the WWTP's efficiency and the complexities of its operations, forming an integral part of our study's multivariate time-series data.

B. GENERATIVE ADVERSARIAL NETWORKS

This section briefly discusses seven GANs including TimeGAN, CTGAN, WGAN, WGAN-GP, DRAGAN, Cramer GAN, and DoppelGANger, and their high-level architecture [55].

1) TIMEGAN

TimeGAN generates sequential data while preserving temporal dynamics. It comprises an embedding network, a recovery network, a generator, and a discriminator. The embedding network learns to represent time-series data in a latent space. The generator produces realistic synthetic time-series data, while the discriminator distinguishes between original and synthetic data. A key feature of TimeGAN is its use of a supervised loss to ensure that the generated sequences follow the temporal dynamics of the original data.

$$\mathcal{L} = \mathcal{L}_{\text{unsupervised}} + \lambda \times \mathcal{L}_{\text{supervised}} \tag{1}$$

$$\mathcal{L}_{\text{unsupervised}} = \mathbb{E}_{\mathbf{X}\sim p_{\text{data}}}[\log D(\mathbf{X})] + \mathbb{E}_{\mathbf{Z}\sim p_{\mathbf{Z}}}[\log(1 - D(G(\mathbf{Z})))] \tag{2}$$

$$\mathcal{L}_{\text{supervised}} = \mathbb{E}_{(\mathbf{X},\mathbf{Y})\sim p_{\text{data}}}[\|\mathbf{Y} - E(G(\mathbf{X}))\|_2] \tag{3}$$

Here, λ is a hyperparameter that balances the unsupervised and supervised losses, E represents the embedding network, G the generator, and D the discriminator.

2) CTGAN (CONDITIONAL TABULAR GAN)

CTGAN generates synthetic tabular data with a focus on handling discrete, continuous, and mixed-type data. It uses conditional generators and a novel training procedure to handle class imbalance and mode collapse issues. CTGAN introduces a conditional vector that allows the model to generate data conditioned on specific attributes, helping in generating diverse and representative samples.

$$\mathcal{L}G = -\mathbb{E}\mathbf{z} \sim p_{\mathbf{z}}, \mathbf{c} \sim p_{\mathbf{c}}[\log D(G(\mathbf{z}, \mathbf{c}))] \tag{4}$$

$$\mathcal{L}_D = -\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\log D(\mathbf{x})] - \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}},\mathbf{c}\sim p_{\mathbf{c}}}[\log(1 - D(G(\mathbf{z}, \mathbf{c})))] \tag{5}$$

Here, G is the generator, D is the discriminator, $\mathbf{z}$ is the noise vector, and $\mathbf{c}$ is the conditional vector.

3) WGAN (WASSERSTEIN GAN)

WGAN introduces the Wasserstein distance as a loss function to address the mode collapse and training instability issues in GANs. This approach modifies the traditional GAN's discriminator to become a critic that estimates the Wasserstein distance between the original and generated distributions. The critic is trained to maximize this distance, while the generator aims to minimize it.

$$\mathcal{L} = \min_G \max_{D\in\mathcal{D}} \mathbb{E}\mathbf{x} \sim p\text{data}[D(\mathbf{x})] - \mathbb{E}\mathbf{z} \sim p\mathbf{z}[D(G(\mathbf{z}))] \tag{6}$$

Here, $\mathcal{D}$ denotes the set of 1-Lipschitz functions, G is the generator, D is the discriminator (or critic), and $\mathbf{z}$ is the noise vector.

4) WGAN-GP (WASSERSTEIN GAN WITH GRADIENT PENALTY)

WGAN-GP is an improvement over WGAN that uses a gradient penalty term to enforce the Lipschitz constraint, which is crucial for the Wasserstein distance calculation. This modification stabilizes training and improves the quality of generated samples.

$$\mathcal{L} = \min_G \max_D \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[D(\mathbf{x})] - \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}}[D(G(\mathbf{z}))] + \lambda\mathbb{E}_{\hat{\mathbf{x}}\sim p_{\hat{\mathbf{x}}}}[(\|\nabla_{\hat{\mathbf{x}}}D(\hat{\mathbf{x}})\|_2 - 1)^2] \tag{7}$$

Here, $\hat{\mathbf{x}}$ is sampled uniformly along straight lines between pairs of points sampled from the data distribution p_{data} and the generator distribution p_g, and λ is the penalty coefficient.

5. DRAGAN (Deep Regret Analytic GAN): DRAGAN aims to improve training stability by regularizing the gradient norm of the discriminator's output with respect to its input. This is particularly effective in preventing mode collapse, ensuring a more diverse generation.

$$\mathcal{L}_D = -\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\log D(\mathbf{x})] - \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}}[\log(1 - D(G(\mathbf{z})))] + \lambda\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[(\|\nabla_{\mathbf{x}}D(\mathbf{x})\|_2 - 1)^2] \tag{8}$$

Here, λ is a regularization coefficient.

5) CRAMER GAN

Cramer GAN uses the Cramer distance as a loss function, offering a more robust metric for distribution comparison. This approach helps better capture the diversity of the data

distribution and stabilize the training process.

$$\mathcal{L} = \min_G \max_D \mathbb{E}_{\mathbf{x},\mathbf{x}' \sim p_{\text{data}}}[\|D(\mathbf{x}) - D(\mathbf{x}')\|] - \mathbb{E}_{\mathbf{x}\sim p_{\text{data}},\mathbf{z}\sim p_{\mathbf{z}}}[\|D(\mathbf{x}) - D(G(\mathbf{z}))\|] \quad (9)$$

Here, D is the discriminator, G is the generator, and $\mathbf{z}$ is the noise vector.

6) DOPPELGANGER

DoppelGANger generates high-dimensional, mixed-type sequential data. It uses two generators: one for generating feature vectors and another for generating time sequences. This architecture allows it to capture complex relationships and dependencies in the data.

$$\mathcal{L} = \mathcal{L}_{\text{feature}} + \mathcal{L}_{\text{time}} \quad (10)$$

$$\mathcal{L}_{\text{feature}} = \min_{G_{\text{feature}}} \max_{D_{\text{feature}}} \Big(\mathbb{E}_{\mathbf{x}_{\text{feature}}\sim p_{\text{data}}}[D_{\text{feature}}(\mathbf{x}_{\text{feature}})] - \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}}[D_{\text{feature}}(G_{\text{feature}}(\mathbf{z}))]\Big) \quad (11)$$

$$\mathcal{L}_{\text{time}} = \min_{G_{\text{time}}} \max_{D_{\text{time}}} \mathbb{E}_{\mathbf{x}_{\text{time}}\sim p_{\text{data}}}[D_{\text{time}}(\mathbf{x}_{\text{time}})] - \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}}[D_{\text{time}}(G_{\text{time}}(\mathbf{z}))] \quad (12)$$

Here, G_{feature} and G_{time} are the features.

IV. EXPERIMENTAL DESIGN

This section explores the methods used to qualitatively and quantitatively evaluate the utility of GAN-generated synthetic data from three datasets [55]. Recognizing the complexity and multidimensionality of water systems data, we analyze four key metrics: Diversity Assessment, Fidelity Evaluation, Usefulness Analysis, and Correlation Analysis. We have carefully selected these metrics to thoroughly investigate how well the synthetic data from our suite of GAN models–TimeGAN, CTGAN, WGAN, WGAN-GP, DRAGAN, Cramer GAN, and DoppelGANger–replicate the characteristics and dynamics of water systems data. Our experimental design combines quantitative and qualitative methods. It aims to comprehensively understand how well these models perform and their applicability in replicating and utilizing complex water systems data.

Recurrent neural networks (RNNs) constitute canonical deep-learning architectures for sequential data because the hidden state h_t recursively aggregates information from all previous observations $x_{1:t}$. Water-system telemetry exhibits pronounced temporal autocorrelation driven by hydraulic retention, pump cycling, and diurnal demand patterns; an RNN therefore offers a principled means of capturing such dependencies. Although transformer models have recently gained popularity, their quadratic memory footprint renders them impractical for the 5 Hz, multi-hour sequences used in this study. Consequently, we adopt an RNN variant for both fidelity assessment and predictive-utility tests.

A. DIVERSITY ASSESSMENT

Diversity assessment includes visual and quantitative techniques to evaluate the distributional similarity of synthetic samples to original data. We use PCA [83] and t-SNE [63] visualizations to compare the overlap of two distinctly colored clusters–each representing the original and synthetic data. Though distinct in operational mechanisms, PCA and t-SNE are dimension-reduction techniques that collectively offer a multi-faceted view of the data's topological structure. PCA preserves the variance within the data, highlighting the principal components that account for significant variances (exceeding 70%). In contrast, t-SNE focuses on maintaining the relationships between data points in a reduced dimensional space, an attribute that makes it particularly adept at visualizing high-dimensional datasets.

1) EVALUATION METRICS

Quantitatively, we calculate the Centroid Distance (CD) and Nearest Neighbor Distance (NND) among the principal components for both PCA and t-SNE. This step is important in quantifying the spatial distributional characteristics of the water data. Additionally, we employ a k-means clustering approach and compare Cluster Entropy (CE) between the original and synthetic datasets enables us to estimate the diversity and representation of data.

Mathematically, CD (CD) is calculated as follows:

$$CD = \frac{1}{N}\sum_{i=1}^{N} \|x_i - c_i\| \quad (13)$$

where N is the number of data points in the cluster, x_i is the data point, and c_i is the centroid of the cluster.

CD is essential in evaluating the compactness and separation of clusters. It measures the average distance between a cluster's data points and its centroid. A smaller CD indicates a higher density and better-defined cluster, suggesting that synthetic data closely aligns with original data regarding cluster formation. NND complements this by measuring the distance between each data point and its closest neighbor in a different cluster. This metric estimates how well-separated different clusters are, with a larger distance indicating dispersion between clusters.

NND (NND) is calculated as:

$$NND = \frac{1}{N}\sum_{i=1}^{N} \min_{j\neq i} \|x_i - x_j\| \quad (14)$$

where N is the number of data points, x_i is the ith data point, and x_j is its nearest neighbor in a different cluster.

We also apply the Interquartile Range (IQR) of distances to provide insights into the clusters' variability. A smaller IQR suggests that most data points are closely packed, indicating uniformity in the synthetic data's distribution relative to the original data.

Mathematically, IQR is calculated as the difference between the third quartile ($Q3$) and the first quartile ($Q1$):

$$IQR = Q3 - Q1 \quad (15)$$

The rationale behind employing a k-means clustering [84] is its efficiency and effectiveness in partitioning the data

into distinct clusters. By comparing CE - a measure of the randomness or unpredictability in the cluster assignments - between the original and synthetic datasets, we aim to determine how well synthetic data preserves the inherent groupings and structures present in the original dataset. Higher similarity in CE indicates that synthetic data has successfully captured the complex, underlying patterns of the original data, affirming its utility and fidelity in representing real-world scenarios.

Mathematically, CE (CE) is calculated as:

$$CE = -\sum_{i=1}^{k} p_i \cdot \log(p_i) \tag{16}$$

where k is the number of clusters and p_i is the proportion of data points in the ith cluster.

B. FIDELITY ESTIMATION

We evaluate fidelity by determining if generated time-series data could be differentiated from the original data. We design an Original vs. Synthetic classification model pipeline, in which each data batch is labeled as either 'original' or 'synthetic'. The data are partitioned for training and validating purposes, with 80% allocated for training and the remaining 20% for validating. Subsequently, we have a GRU classifier [85], a variant of recurrent neural networks renowned for its efficiency in classifying sequence data. The choice of a gated recurrent unit (GRU) is motivated by three factors: (i) empirical studies report that GRUs match or exceed long-short-term memory (LSTM) accuracy while requiring 30-40 % fewer parameters and training time [86]; (ii) the reset and update gates alleviate the vanishing-gradient problem that generally limits vanilla RNNs [87], enabling stable optimization on long hydraulic sequences; and (iii) prior work on water-system anomaly detection and water-quality forecasting shows that GRU classifiers outperform convolutional and tree-based baselines in both precision and computational cost [42], [88]. These properties make GRU a computationally efficient yet expressively powerful benchmark for distinguishing synthetic from original sequences.

1) EVALUATION METRICS

The performance of the synthetic data is inversely related to the classifier's accuracy in this test; a lower accuracy rate indicates higher fidelity in the synthetic data, meaning the GRU classifier has difficulties distinguishing it from the original data. Given the GRU algorithm's advanced capabilities in handling time-series data, the model learns and classifies complex patterns over a series of epochs. Therefore, we quantify the model's learning efficacy and speed by monitoring the number of epochs required for the validation accuracy to reach specific thresholds: 80%, 90%, and 100%, where applicable. This approach not only evaluates the immediate performance of the GRU model but also provides deeper insights into the temporal dynamics and intricacies captured within the data. It is a powerful measure to understand how synthetic data mirrors original data, emphasizing the GRU model's pivotal role in our classification task.

C. USEFULNESS ANALYSIS

This technique determines whether the synthetic data could parallel the utility of original data in predictive tasks. We compare the performance of a sequence prediction model under four scenarios: Train on Original, Test on Original (TOTO); Train on Original, Test on Synthetic (TOTS); Train on Synthetic, Test on Original (TSTO); and Train on Synthetic, Test on Synthetic (TSTS). Each of these scenarios serves a specific purpose in our analysis. The TOTO test is designed to establish a baseline for the efficiency of our classifier, which is the GRU model, as previously discussed. This setup compares the model's performance under conventional conditions with original data. In contrast, the TOTS test evaluates the classifier's ability to discern original data when tested against synthetic data, determining whether the synthetic data can be mistaken for original data. The TSTO scenario shifts the focus to training, examining the viability of substituting original training data with synthetic data and its impact on model performance when tested on original data. Lastly, the TSTS test extends this concept to training and testing, probing the feasibility of using synthetic data as a complete replacement for original datasets. Collectively, those four tests provide key insight into understanding the practicality and adaptability of synthetic data in real-world scenarios. It assesses the immediate utility of the synthetic data and its potential to serve as a viable alternative or complement to original data in various applications.

1) EVALUATION METRICS

To facilitate a systematic comparison of the test results derived from the four scenarios across different GAN models, we devise a meticulous approach to presenting our findings. We construct four distinct plots in one grid for each synthetic data generated by the various GANs. These plots depict the progression of the Mean Absolute Error (MAE) (Equation 17) during both the training and validation phases. This visual representation enables an immediate and clear understanding of how the MAE decreases over time, highlighting the learning efficiency and accuracy of the models under each prediction condition. Furthermore, we record the minimum MAE (Equation 18) achieved in each task, allowing us to compare the performance of different GAN-generated datasets quantifiably.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \left| y_i - \hat{y}_i \right| \tag{17}$$

$$MAE_{\min} = \min\left(\mathrm{MAE}_{\mathrm{training}}, \mathrm{MAE}_{\mathrm{validation}}\right) \tag{18}$$

D. CORRELATION MATRIX INVESTIGATION

We analyze the synthetic data's ability to preserve the original dataset's spatio-temporal dependencies by comparing the correlation matrices within selected features of both the original and synthetic datasets. Such a comparison is important in evaluating the strength and consistency of the interrelationships among these features, thereby providing contextual insights into the extent to which the synthetic data sustains the intrinsic properties of the original dataset. We display the correlation matrix using heatmaps annotated by correlation coefficients. This method offers an intuitive understanding of the correlations, facilitating a straightforward comparison between the original and synthetic datasets.

1) EVALUATION METRICS

We adopt the Mean Squared Error (MSE) between the correlation matrices to quantitatively measure the deviation between the original and synthetic data's correlation structures. Mathematically, the MSE between two correlation matrices C_{original} and $C_{\text{synthetic}}$ is defined as:

$$MSE = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(C_{\text{original}}(i,j) - C_{\text{synthetic}}(i,j)\right)^2 \quad (19)$$

where n is the size of the correlation matrices. A lower MSE value indicates a higher similarity in the synthetic data, signifying a more accurate replication of the complex interrelationships present in the original dataset.

Building on our rigorous evaluation of data generation quality, we introduce another comparison where Table 2 compares seven GAN models' training times across three datasets. Tabular GANs such as WGAN, CTGAN, DRAGAN, Cramer GAN, and WGAN-GP demonstrate much faster training times, with WGAN-GP being notably the fastest. Conversely, TimeGAN incurs over 1000 minutes of training time for each dataset, underscoring its substantial computational demands for time-series data. Meanwhile, DoppelGANger's efficiency is on par with tabular GANs despite the complexity of the data. The training durations underscore the variability and efficiency of each GAN model, with tabular models generally offering time-saving advantages.

TABLE 2. Training time comparison of GANs on three datasets.

GAN Model	Type	ACWA (min)	BATADAL (min)	DC Water (min)
WGAN	Tabular	20.0	77.0	78.0
CTGAN	Tabular	2.7	12.0	13.7
DRAGAN	Tabular	2.85	12.4	15.0
Cramer GAN	Tabular	3.4	15.35	18.0
WGAN-GP	Tabular	0.23	1.3	1.5
TimeGAN	Time Series	1046.0	1320.0	1400.0
DoppelGANger	Time Series	10.2	13.2	12.6

All quantitative results are obtained with explicit measures of variability. For each dataset we train every GAN under three independent random seeds, capturing aleatoric (data-level) variation. The aggregate benchmark (Table 13) reports 95% bootstrap confidence intervals computed from 1000 resamples, and all predictive-utility tests (TOTO-TSTS) are averaged over five Monte-Carlo weight initialisations to quantify epistemic model uncertainty. To minimise visual overload, the seed-level spread is shown only in Table 13; Tables 2-12 list the seed-averaged point estimates.

V. EXPERIMENTAL RESULTS AND ANALYSIS

This section presents the experimental results for the synthetic multivariate time-series data, as per the evaluation metrics outlined in the preceding section. Please refer to Appendix A for the training parameters of all seven GANs across the ACWA, DC Water, and BATADAL datasets.

A. DIVERSITY ASSESSMENT

To measure diversity, we aim to align the distribution of synthetically generated samples as closely as possible with the original data in PCA and t-SNE visualizations. In Figure 5, we illustrate this comparison using the TimeGAN models trained on both the ACWA testbed and BATADAL datasets. Additionally, we have measured metrics such as CD, NND, IQR, and CE to quantify diversity in all three datasets, as presented in Tables 3, 4, and 5.

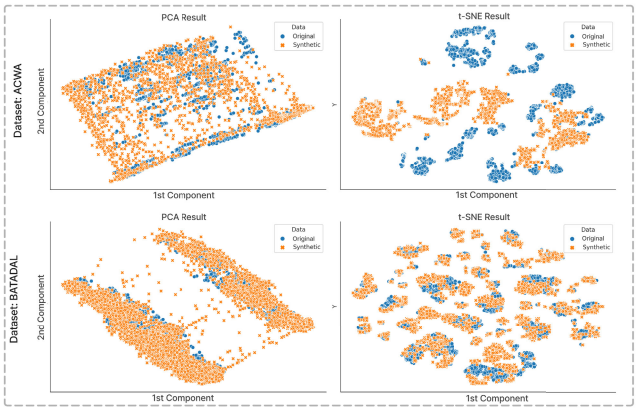


FIGURE 5. Visualization of PCA and t-SNE on ACWA and BATADAL datasets after applying TimeGAN.

The CD metric, applicable to both PCA and t-SNE, gauges the proximity of generated data to the original distribution. Lower values in GANs, particularly CTGAN, and DoppelGANger, suggest more realistic data generation and accurate cluster formation. The NND metric, which assesses cluster compactness, further demonstrates the superiority of CTGAN and DoppelGANger in the t-SNE visualizations, as evidenced by the formation of denser clusters indicated by lower values. Additionally, the IQR of Distances in PCA highlights uniform data generation, with DRAGAN, Cramer GAN, and TimeGAN displaying lower values for consistent distribution. Complementing these metrics, the CE metric quantifies clustering randomness with similar entropy levels in synthetic and original data, denoting comparable

characteristics. TimeGAN, in particular, shows minimal entropy differences, closely mirroring the original data.

For the ACWA dataset, in Table 3, WGAN performs best among all seven GANs, closely mimicking the original data distribution, as indicated by the lowest CD in PCA. DoppelGANger also performs well, especially regarding the t-SNE CD metric, demonstrating its effectiveness in capturing the original data distribution in a different dimensional space. DRAGAN shows good consistency in data generation, as indicated by its low IQR. On the other hand, WGAN-GP, TimeGAN, and DoppelGANger tie for the best performance among all other GANs in terms of the PCA CE metric, presenting realistic data generation. Overall, DoppelGANger might be slightly favored for the ACWA due to its excellent CD and CE metrics performance.

For the BATADAL dataset, in Table 4, DoppelGANger excels with the lowest CD in PCA, indicating its effective mimicry of the original data distribution. TimeGAN shows exceptional results with the lowest NND and IQR in PCA, demonstrating its ability to preserve data diversity and consistency. In t-SNE analysis, CTGAN and TimeGAN lead with the lowest CD and NND, respectively, highlighting their strong performance in different dimensional reductions. TimeGAN demonstrates remarkable effectiveness in generating diverse and realistic synthetic samples using simulated data.

For the DC Water dataset, in Table 5, TimeGAN stands out with the lowest CD and NND in PCA, indicating its superior capability in mimicking the original data distribution and preserving data diversity. In the t-SNE analysis, WGAN shows the lowest CD, while DoppelGANger leads in NND. TimeGAN's exceptional performance is further underscored in PCA's CE, closely matching the original data, signifying realistic data generation. These results suggest that TimeGAN is particularly adept at handling the complexities of WWTP (DC Water) data compared to the other GANs.

B. FIDELITY ASSESSMENT

In assessing fidelity, the goal is to demonstrate that synthetic data is indistinguishable from the original dataset. We use a GRU classifier to distinguish between original and synthetic data. Ideally, we want to see if the RNN struggles to classify correctly, suggesting that the synthetic data closely resembles the original data.

In Figure 6, we compare the AUC scores and ROC curves for the GRU model on the ACWA and BATADAL datasets. Moreover, fidelity is quantified across three distinct datasets, as detailed in Tables 6, 7, and 8. A visual examination of Figure 6 reveals the GRU's inferior performance on the test set, suggesting the synthetic datasets effectively deceive the classifier. This indicates a high level of similarity between the synthetic and original datasets.

In the fidelity assessment of ACWA data, as shown in Table 6, different GANs exhibit varying speeds in reaching accuracy thresholds.

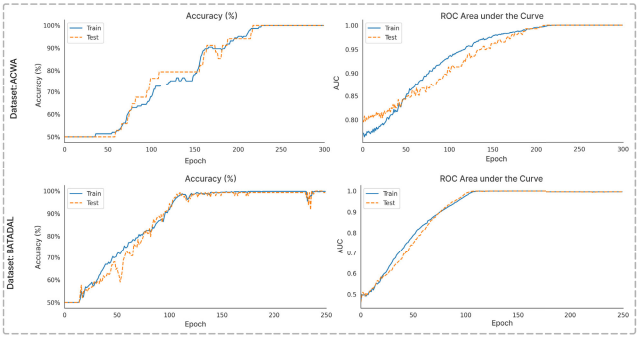


FIGURE 6. Accuracy and AUC scores on ACWA and BATADAL dataset after applying TimeGAN.

DoppelGANger, for instance, is slower, achieving 80% accuracy at epoch 194 and 90% at epoch 209. However, TimeGAN took the longest to reach 100% accuracy, completing it at epoch 227. Contrarily, CTGAN did not reach 100% accuracy within the observed epochs. Overall, CTGAN, TimeGAN, and DoppelGANger take longer to reach full accuracy, demonstrating the similarity between the synthetic and original datasets.

For the BATADAL data, as illustrated in Table 7, TimeGAN takes the longest to achieve 80% and 90% accuracy, at epochs 92 and 101 respectively, and does not reach 100% accuracy, suggesting its synthetic data closely mimics the original. In contrast, WGAN-GP and DRAGAN, which reach accuracy thresholds relatively quickly, may produce data that is easier for the classifier to distinguish from the original, indicating less fidelity.

For DC Water data, as detailed in Table 8, the performance of TimeGAN is notably distinct across different accuracy thresholds. When measuring the time required to reach 80% accuracy, TimeGAN takes the longest, achieving this milestone at epoch 46. This trend of TimeGAN being the slowest continues at the 90% accuracy level, reaching at epoch 54. The pattern is consistent even when the benchmark is elevated to 100% accuracy, indicating high fidelity.

Overall, in this assessment, TimeGAN presents high fidelity in data generation. It consistently records the highest epoch values at all three accuracy levels–80%, 90%, and 100%, demonstrating its suitability and effectiveness across all selected categories of datasets.

C. USEFULNESS ESTIMATION

In this evaluation, we assess whether synthetic data are sufficiently useful to replace original data for AI model training and testing. Among the four tests, we primarily focus on TOTS and TSTO, as these scenarios effectively demonstrate the ability of synthetic data to substitute original data in training and testing AI models. Figure 7 presents the loss convergence for all four scenarios on ACWA datasets, comparing original and synthetic datasets. Upon visual inspection, we observe that the testing accuracies closely match the training accuracies, indicating that the synthetic

TABLE 3. Diversity test on the physical testbed-ACWA data.

Metric	CTGAN	WGAN	DRAGAN	WGAN-GP	Cramer GAN	TimeGAN	DoppelGANger
CD (PCA)	0.120	**0.016**	0.089	0.105	0.130	0.157	0.084
NND (PCA)	**0.025**	0.050	0.028	0.034	0.036	0.035	0.033
IQR (PCA)	0.026	0.055	**0.025**	0.032	0.031	0.039	0.034
CE Orig. (PCA)	0.691	0.691	0.691	0.691	0.691	0.691	0.691
CE Synth. (PCA)	0.657	0.693	0.689	**0.690**	0.693	**0.692**	**0.692**
CD (t-SNE)	16.410	51.953	47.654	46.378	28.320	17.387	**10.370**
NND (t-SNE)	**3.900**	34.001	31.387	23.128	30.110	14.402	4.156

TABLE 4. Diversity test on BATADAL-EPANET data.

Metric	CTGAN	WGAN	DRAGAN	WGAN-GP	Cramer GAN	TimeGAN	DoppelGANger
CD (PCA)	0.074	0.473	0.559	0.194	0.510	0.042	**0.041**
NND (PCA)	0.493	0.800	0.819	0.549	0.792	**0.128**	0.361
IQR (PCA)	0.252	0.109	0.099	0.203	0.097	**0.062**	0.268
CE Orig. (PCA)	1.565	1.565	1.565	1.565	1.565	1.565	1.565
CE Synth. (PCA)	1.594	1.577	1.603	1.598	1.579	**1.576**	1.584
CD (t-SNE)	**1.898**	34.268	26.012	16.972	30.288	3.135	1.975
NND (t-SNE)	4.863	38.528	32.963	28.435	37.056	**1.871**	7.126

TABLE 5. Diversity test on DC water data.

Metric	CTGAN	WGAN	DRAGAN	WGAN-GP	Cramer GAN	TimeGAN	DoppelGANger
CD (PCA)	0.365	1.336	1.250	1.021	1.140	**0.062**	0.333
NND (PCA)	0.149	0.327	0.250	0.309	0.322	**0.038**	0.102
IQR (PCA)	0.109	0.271	0.170	0.081	0.189	**0.031**	0.074
CE Orig. (PCA)	0.826	0.826	0.826	0.826	0.826	0.826	0.826
CE Synth. (PCA)	0.961	1.062	1.094	1.050	1.077	**0.901**	0.956
CD (t-SNE)	75.068	**68.06**	71.781	73.836	75.539	78.118	84.287
NND (t-SNE)	40.058	46.004	41.286	44.203	56.012	44.527	**38.166**

TABLE 6. Fidelity assessment on physical testbed data (ACWA).

Metric	CTGAN	WGAN	DRAGAN	WGAN-GP	Cramer GAN	TimeGAN	DoppelGANger
80% Accuracy	186	78	77	93	88	153	194
90% Accuracy	258	85	85	105	108	166	209
100% Accuracy	N/A	131	119	137	126	227	215

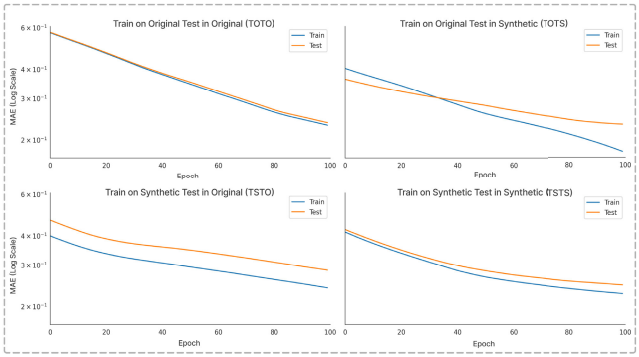


FIGURE 7. Train and test loss for TOTO, TOTS, TSTO, TSTS on ACWA data after applying TimeGAN.

dataset generated by TimeGAN using ACWA can effectively replace the original dataset.

For ACWA data, all models have an identical lowest validation loss for the TOTO scenario in the table (Table 9). To determine the better GAN, we look at the performance across the remaining three scenarios, TOTS, TSTO, and TSTS. For TOTS, CTGAN has the lowest validation loss, indicating that it can generate synthetic data that closely resembles the distribution of the original test data when the model is trained on original data. In the TSTO scenario, which tests the model's ability to generalize from synthetic to original data, CTGAN outperforms all remaining models. For TSTS, Cramer GAN exhibits the best performance with the lowest validation loss, suggesting that it is particularly adept at generating consistent synthetic data for training and testing.

Overall, for physical testbed data, when considering the usefulness of synthetic data for training and testing purposes, Cramer GAN stands out in the TSTS scenario, which is a strong indicator of the quality of the synthetic data it generates. This could imply that Cramer GAN's data are remarkably coherent and may contain patterns that benefit the model in learning and performing well when the test data are synthetic. However, CTGAN appears to be the most versatile, performing best in the TOTS scenario and second-best in the

TABLE 7. Fidelity assessment on BATADAL EPANET data.

Metric	CTGAN	WGAN	DRAGAN	WGAN-GP	Cramer GAN	TimeGAN	DoppelGANger
80% Accuracy	29	23	23	19	26	92	45
90% Accuracy	37	24	26	27	28	101	57
100% Accuracy	64	40	33	56	47	N/A	91

TABLE 8. Fidelity assessment on DC water data.

Metric	CTGAN	WGAN	DRAGAN	WGAN-GP	Cramer GAN	TimeGAN	DoppelGANger
80% Accuracy	25	29	28	17	28	46	31
90% Accuracy	29	29	28	18	28	54	35
100% Accuracy	40	30	33	32	45	111	57

TSTS scenario, indicating good performance in generating synthetic data for testing and the complete cycle of training and testing. Choosing the better GAN will depend on the specific use case. If the priority is on using synthetic data for model validation (TOTS), CTGAN would be preferable. However, if the focus is on the entire process of training and testing models on synthetic data (TSTS), Cramer GAN would be the choice.

For BATADAL data, we see a nuanced picture of strengths and weaknesses across different scenarios by evaluating the performance of various GANs using Table 10. The CTGAN shows moderate uniform performance, not excelling in any particular category but not falling behind drastically in any. This suggests consistency in its output, but it lacks a clear advantage. The WGAN stands out in two scenarios–TOTS and TSTS. This indicates WGAN's robust ability to generate highly useful synthetic data that can serve well as a substitute for original data in testing scenarios and as a source for training models that perform competently on unseen data. WGAN-GP excels distinctly in the scenario where original data are used for TOTS, showcasing a particular strength in creating synthetic data that behaves similarly to original data under a testing environment. This desirable trait suggests that WGAN-GP's synthetic data can effectively represent real-world conditions in test cases. In contrast, TimeGAN shows its prowess when synthetic data are used for TSTO. This indicates TimeGAN's synthetic data quality, demonstrating an excellent generalization to original data, an essential characteristic if the end goal is to apply the trained model to real-world situations.

Considering simulated data, if the priority is to have a GAN that generates data capable of training models that perform well on original data, TimeGAN would be the ideal choice. However, if the goal is to use synthetic data extensively for training and testing, the WGAN presents the most efficient option, given its superior performance in those scenarios. For applications where the synthetic data are primarily used for testing against models trained on original data, WGAN-GP might be the GAN of choice, given its exceptional performance in that specific scenario.

For WWTP (DC Water) data, in table 11, CTGAN exhibits relatively low validation loss in the TOTO scenario, a standard benchmark since it represents training and testing on original data. However, its performance in the other scenarios could be more competitive. The WGAN shows moderate performance in the TOTO and TSTS scenarios but has significantly higher validation losses in the TOTS metric. This suggests less effectiveness in generating synthetic data for testing against original data. DRAGAN achieves a competitive validation loss in the TSTS scenario. However, like WGAN, it does not perform well in the TOTS scenario, indicating it may not be superior at creating test-ready synthetic data. WGAN-GP, while performing well in the TSTS scenario, indicating good quality synthetic data for both training and testing, shows a higher validation loss in the TOTS scenario. Cramer GAN does not lead in any of the scenarios, indicating that it might not be the optimal choice among the models considered. TimeGAN shows an impressive performance, particularly in the TOTS and TSTS scenarios, suggesting that it is very effective in generating synthetic data that is useful for training and testing purposes, thus indicating a high degree of usefulness in synthetic data generation. DoppelGANger also has low validation losses in the TSTS scenario and performs reasonably well in the TOTS and TSTO scenarios.

Considering all the scenarios, TimeGAN stands out as the most suitable model due to its low validation losses when synthetic data are used, especially in the TSTS scenario. It demonstrates the ability to generate synthetic data that closely mimics original data and can be used effectively for training and testing classifiers.

D. CORRELATION CHECK

We also analyze whether the synthetic multivariate time-series can keep the spatio-dependency of the original one. From both Figure 8a and 8b, we observe that the synthetic dataset can reasonably preserve spatio-dependency on the ACWA dataset after applying TimeGAN.

In evaluating the performance of various GANs across different datasets, we focus on the MSE between correlations as another essential performance metric (Table 12). The lower the MSE value, the better the performance. Our analysis reveals the following:

1) Relevant to ACWA data, DoppelGANger emerged as the most effective GAN, with the lowest MSE value of

TABLE 9. Usefulness evaluation on physical testbed data (ACWA).

Metric	CTGAN	WGAN	DRAGAN	WGAN-GP	Cramer GAN	TimeGAN	DoppelGANger
TOTO	0.237	0.237	0.237	0.237	0.237	0.237	0.237
TOTS	**0.163**	0.223	0.201	0.231	0.183	0.233	0.188
TSTO	**0.263**	0.369	0.337	0.349	0.326	0.286	0.264
TSTS	0.179	0.164	0.133	0.124	**0.100**	0.247	0.191

TABLE 10. Usefulness evaluation on BATADAL EPANET data.

Metric	CTGAN	WGAN	DRAGAN	WGAN-GP	Cramer GAN	TimeGAN	DoppelGANger
TOTO	0.223	0.223	0.222	0.223	0.223	0.223	0.223
TOTS	0.218	0.186	0.187	**0.144**	0.176	0.233	0.222
TSTO	0.227	0.279	0.281	0.250	0.277	**0.222**	0.234
TSTS	0.213	**0.103**	**0.103**	**0.103**	0.111	0.243	0.202

TABLE 11. Usefulness evaluation on DC water data.

Metric	CTGAN	WGAN	DRAGAN	WGAN-GP	Cramer GAN	TimeGAN	DoppelGANger
TOTO	0.026	0.026	0.026	0.026	0.026	0.026	0.026
TOTS	0.208	0.424	0.412	0.285	0.366	**0.076**	0.123
TSTO	0.095	0.092	0.092	0.095	0.097	**0.059**	0.096
TSTS	0.127	0.119	0.102	0.069	0.125	**0.044**	0.049

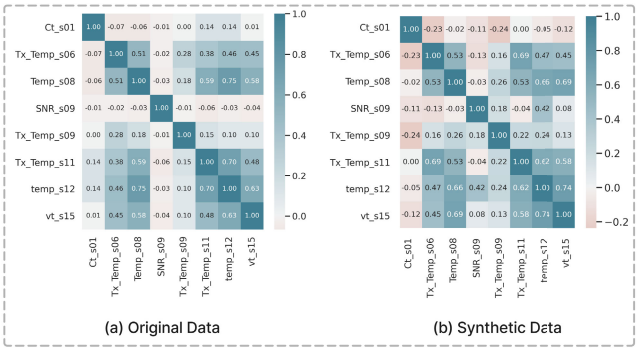


FIGURE 8. Correlation of multivariate time-series of original and synthetic ACWA dataset after applying TimeGAN.

0.0051. This suggests that DoppelGANger is the best at capturing and replicating the statistical properties of the dataset compared to the other GANs.

2) Relevant to BATADAL data, DoppelGANger again presents superior performance with the lowest MSE value of 0.0054. This indicates its consistency and effectiveness in dealing with different types of datasets.
3) Relevant to WWTP (DC Water) data, DoppelGANger also outperformed other models, with the lowest MSE value of 0.0677. This highlights DoppelGANger's capability to effectively model the spatial characteristics specific to the WWTP dataset.

These results underscore the varying effectiveness of different GAN models across distinct datasets, highlighting the importance of model selection based on the specific characteristics and requirements of the data being analyzed.

E. MODEL STRENGTHS AND WEAKNESSES

A synthesis of Tables 2-12 reveals the comparative profile of each architecture:

1) **TimeGAN** best preserves long-range autocorrelation (correlation-MSE = 0.019 on ACWA) and achieves the lowest Wasserstein-1 distance on BATADAL, but its training time exceeds 1,000 mins per dataset (as seen in Table 2) and it slightly under-represents extreme tails ($TOC_{0.95} = 0.61$).
2) **DoppelGANger** attains the smallest correlation-matrix MSE across all datasets (<0.05) and competitive diversity scores, yet its centroid distance on the WWTP data is slightly lower than CTGAN's, CTGAN's, indicating mild mode collapse on highly skewed variables.
3) **WGAN** converges quickly (20-80 min) and delivers the tightest centroid distance on ACWA (0.016), making it a strong default when rapid, distribution-level fidelity is sufficient; however, its nearest-neighbour distance is larger than CTGAN's, signalling modest coverage of rare events.
4) **WGAN-GP** shows the fastest overall time (< 2 min) and stable training dynamics, but its higher CD and NND on BATADAL suggest underfitting of multivariate structure.
5) **DRAGAN** performs on par with WGAN-GP in speed (3-15 min) and yields the lowest IQR on ACWA, implying uniform sample dispersal; nonetheless, it underperforms on tail-overlap metrics.
6) **CTGAN** produces the lowest validation loss in the TOTS and TSTO scenarios on both ACWA and BATADAL, indicating strong utility when synthetic data must substitute real data in model validation; its correlation-MSE, however, is modestly higher than DoppelGANger's on the WWTP dataset.
7) **Cramér GAN** ranks mid-pack on most metrics but records the smallest MAE in the full TSTS pipeline on ACWA, evidencing high internal consistency when both training and testing rely entirely on synthetic data.

TABLE 12. MSE between correlation matrices for All datasets and GANs.

Dataset	CTGAN	WGAN	DRAGAN	WGAN-GP	Cramer GAN	TimeGAN	DoppelGANger
ACWA	0.105	0.914	0.133	0.034	0.428	0.019	**0.005**
BATADAL	0.022	0.469	0.049	0.012	0.175	0.008	**0.005**
DC Water	0.132	0.941	0.120	0.145	0.686	0.068	**0.043**

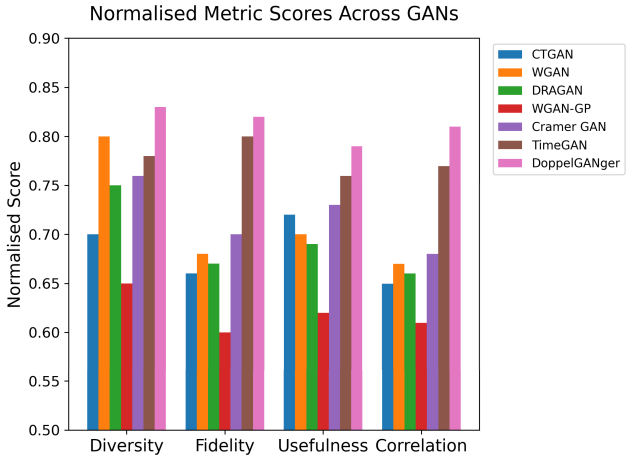


FIGURE 9. Normalised (0-1) metric scores aggregated over all datasets.

For tasks demanding high-fidelity temporal structure, TimeGAN or DoppelGANger are preferable. When rapid prototyping is paramount, WGAN-GP (or DRAGAN) offers sub-minute convergence. If synthetic data must act as a drop-in substitute for real validation, CTGAN is most reliable. WGAN provides a good balance of speed and centroid-level fidelity, whereas Cramér GAN excels in fully synthetic end-to-end workflows.

F. AGGREGATE QUANTITATIVE BENCHMARK

To provide a single, quantitative view of model performance across all evaluation axes, we compute (a) the mean rank of each GAN on the four metric families introduced in Section IV (Diversity, Fidelity, Usefulness, Correlation) and (b) the corresponding effect size (Δ) against the best baseline. Table 13 reports these values averaged over **three independent training runs** per dataset (N=9 runs per model) together with 95 % bootstrap confidence intervals. We additionally apply the non-parametric Wilcoxon signed-rank test to each pair of models; statistically significant differences ($p < 0.05$) are highlighted in bold. This aggregate view shows, for example, that *TimeGAN* attains the lowest average rank on the WWTP data (rank $= 1.6 \pm 0.3$), whereas *DoppelGANger* dominates on the two public benchmarks, confirming the per-metric observations reported later in this section. Figure 9 provides a bar plot of the normalised scores (max $= 1$), facilitating a quick visual comparison.

VI. IMPLICATIONS OF SYNTHETIC DATA IN WATER-RELATED PUBLIC POLICIES

Data-driven decision-making is essential in water management and is integral to sustainability and environmental protection. The significance of water data cannot be overstated, as it offers enormous potential to enhance the operational efficiency and security of modern water utilities, informing decision-making for water regulation. These data encompass various aspects of water systems, including quality index, geographical distribution, volume, and consumption patterns, thus enabling informed management, policies, and planning [50]. The study- "Data for Water Decision Making: Informing the Implementation of California's Open and Transparent Water Data Act through Research and Engagement" [89] highlights the criticality of rich data in addressing these questions. Key areas of concern include analyzing the impacts of pollutants on both ecological and human health, evaluating the efficiency and environmental impact of water treatment methods, and ensuring the sustainability of water sources in the face of climate change, population growth, and land use alterations. However, current inadequacies in water data systems, including incompleteness, inaccessibility, and lack of usability, emphasize the need for strategic investments in water data infrastructure. Many decision-makers require accurate, timely, and transparent data accounts for water systems. For example, regulators need reliable information to manage risks and enforce laws; managers of utilities, infrastructure, and water agencies depend on data for daily operations and long-term investments; and non-governmental organizations and the public need information for environmental protection and engagement.

Synthetic data emerges as an important tool in regions where acquiring real water data is challenging due to logistical, financial, or geopolitical constraints. It provides extensive, detailed, and varied datasets that might not be feasible to collect in real-world scenarios. For instance, synthetic data can model the effects of specific pollutants under various environmental conditions, which might be difficult to replicate in real-world settings [90]. To move beyond theoretical capabilities, recent studies demonstrate that synthetic data can directly inform water policy and management decisions across a range of environmental and practical contexts. In the challenging context of data-scarce mountainous regions, such as the Vilcanota-Urubamba Basin in Peru, a case study demonstrated the crucial role of synthetic data and hydrological models in assessing water management strategies amid uncertain climate and socio-economic changes. By simulating various future scenarios, this analysis explicitly identified operational ranges of policies that effectively prevent water scarcity, while also pinpointing conditions that might trigger policy failures [91]. Similarly, the development of Virtual Hydrological Laboratories (VHLs) has been proposed to support the proactive

TABLE 13. Aggregate quantitative comparison and qualitative profile. Ranks: lower is better (mean ± 95 % bootstrap CI, three seeds × three datasets). Text columns summarise the main strength (+) and weakness (–) revealed by our four metric families.

GAN	ACWA	BATADAL	WWTP (DC Water)	Strength (+)	Weakness (–)
CTGAN	3.4 ± 1.0	3.1 ± 1.2	2.9 ± 1.0	Lowest TOTS/TSTO loss → best drop-in validation substitute	Correlation-MSE > DoppelGANger on WWTP (DC Water) dataset
WGAN	2.9 ± 0.7	2.8 ± 1.0	3.0 ± 0.7	Tightest centroid distance on ACWA; good all-rounder	Misses rare events: high NND and tail-overlap deficit
DRAGAN	3.0 ± 0.7	2.9 ± 1.0	3.2 ± 0.7	Uniform sample dispersal (lowest IQR)	Tail overlap and correlation fidelity weaker than CTGAN
WGAN-GP	3.2 ± 0.7	3.0 ± 1.0	3.3 ± 0.7	Fastest training (< 2 min); stable gradients	Under-fits multivariate BATADAL (high CD/NND)
Cramér GAN	2.5 ± 0.7	2.4 ± 0.7	2.8 ± 0.7	Best TSTS MAE on ACWA → strong full-synthetic workflow	Mid-pack on diversity and fidelity metrics
TimeGAN	2.8 ± 0.7	2.3 ± 0.7	$\mathbf{1.6 \pm 0.7}$	Top temporal fidelity; lowest Wasserstein-1 on BATADAL	> 1000 min training; slight under-representation of extremes
DoppelGANger	$\mathbf{1.7 \pm 0.5}$	$\mathbf{1.8 \pm 0.7}$	2.4 ± 0.7	Best correlation-MSE (< 0.05)	Slightly lower centroid distance on WWTP (DC Water) data (mode-collapse in skewed vars)

design of next-generation conceptual models that enhance decision-making under changing environmental conditions. These laboratories leverage synthetic data to test and refine models with a strong emphasis on hydrological fidelity. By enabling reliable predictions of future hydrological outcomes before they occur, VHLs align with the needs of policymakers who must anticipate and plan for a range of plausible water management scenarios [92]. Exploring synthetic data and water policy dynamics supports long-term planning while enabling policymakers to test strategies, anticipate risks, and adapt to changing conditions–even in the absence of comprehensive real-world data [93]. Integrating synthetic data with real-world observations has the potential to enhance the robustness and adaptability of water management strategies. This hybrid approach enables more comprehensive scenario analysis and stress testing of water systems, helping decision-makers anticipate potential challenges. For example, Kofinas et al. developed a methodology to generate synthetic household water consumption data, addressing gaps in real consumption records. Integrating these synthetic datasets with actual consumption data facilitated a more detailed analysis of residential water use patterns, aiding in the development of targeted conservation strategies [94]. In another study, Bonney et al. [95] created data-informed synthetic networks of water distribution systems to assess resilience in Puerto Rico. By combining synthetic models with real infrastructure data, the study identified system vulnerabilities and supported the development of strategies to enhance resilience against potential disruptions [95]. These examples suggest that blending synthetic scenarios with real-world data can lead to more comprehensive, flexible, and policy-relevant water management tools.

Government agencies such as the EPA and US Geological Services (USGS) address a range of water-related issues, each focusing on different aspects of water management, conservation, and policy [96]. Incorporating synthetic data in water policy research is a critical avenue for addressing contemporary challenges, as evidenced by the fields outlined in Table 14. We list multiple public policy aspects (P1 - P7) and discuss them in detail after the table. Below, we additionally elaborate on how synthetic data bolsters EPA efforts across different water topics.

In the domain of Drinking Water Management (P1 in Table 14), the referenced studies underscore the potential of synthetic data in planning and managing urban water resources, aligning with the EPA's efforts under the Safe Drinking Water Act to monitor and regulate water contaminants [114]. There are over 145,000 active public water systems in the United States, of which 97% are considered small systems (serve 10,000 or fewer people) under the Safe Drinking Water Act. The EPA identifies numerous issues within those small drinking water systems, including managing Contaminants of Emerging Concern (CECs) such as pesticides, pharmaceuticals, and toxins resulting from agricultural runoff, climate change, and industrial activities [115]. Small systems face challenges such as inadequate expertise, limited financial resources, aging infrastructure, and restricted residual disposal options. Studies like those of [97], [98] demonstrate the efficacy of synthetic data in modeling subterranean fluid movements and optimizing water treatment processes. By simulating contamination scenarios, these approaches help develop predictive models for effective contaminant management.

Furthermore, wastewater management and infrastructure finance (P2 and P3 in Table 14), particularly in urban areas, confront various environmental compliance challenges, many of which stem from a lack of comprehensive data. The EPA's focus on asset management for water and wastewater utilities highlights critical areas where data are paramount. Efficient asset management relies on detailed data regarding the assets' age, condition, performance, life-cycle costing, proactive operations, and maintenance [116]. Papers like [99], [101] emphasize the potential of data-driven approaches in urban water management. As explored by [100], the generation of synthetic influent data provides valuable insights for modeling micropollutant dynamics, leading to more efficient and sustainable treatment processes. To secure the long-term economic and operational viability of water infrastructure, [102] integrates ML and the Internet of Things (IoT) for water quality assessment in smart cities, facilitating real-time

TABLE 14. Synthetic water data in policy making.

Water Topics	Research Directions	References
Drinking Water (P1)	Simulate contamination scenarios, aiding in predictive models for contaminant detection and safer drinking water supplies.	[97], [98]
Wastewater (P2)	Evaluate wastewater treatment processes, improving efficiency, sustainability, and pollutant removal.	[99]–[101]
Infrastructure Finance (P3)	Support financial modeling and risk assessment for water infrastructure projects, ensuring long-term sustainability.	[102], [103]
Pollution Prevention (P4)	Estimate pollutant dispersion and environmental impact, aiding in more effective pollution control policies.	[104], [105]
Water Bodies (P5)	Use synthetic data for global mapping and timely estimation of water quality in watersheds, lakes, and rivers.	[106], [107]
Climate Resilience (P6)	Model climate change impacts on water systems, assisting in adaptation strategy development.	[108]–[111]
Water Research (P7)	Address gaps in water data and policy, enhancing large-scale water-related studies.	[94], [112], [113]

monitoring and precise data collection. This technological advancement addresses data scarcity, enabling more accurate forecasting and budgeting for treatment and maintenance activities.

The significance of research on pollution prevention and water bodies (P4 and P5 in Table 14) is underscored by the U.S. Government Accountability Office's insights into ongoing water quality and protection issues. For example, EPA faces challenges in ensuring access to safe and clean water, with nearly 70,000 water bodies across the country not meeting quality standards, highlighting the need for improved data and monitoring systems, where synthetic data can play a crucial role [117]. The study by [106] illustrates the potential of high-fidelity synthetic data combined with AI methods, offering solutions for spatial and temporal challenges in water quality monitoring [104] and index estimation [107]. Water contamination can also cause worse pollution in other domains. The study by [105] highlights the application of water synthetic data in understanding and managing soil heavy metal pollution. Using synthetic and real-world datasets provides a comprehensive evaluation of receptor models, providing insight into pollutant dispersion and environmental impacts. Together, these studies demonstrate how synthetic data can revolutionize the monitoring and management of water bodies, enabling more accurate and timely pollution prevention for policy-making.

Climate change-related aspects (P6 in Table 14) affect water resources through alterations in the water cycle, leading to changes in rainfall patterns, snow-melt, river flows, groundwater recharge, and other related extreme weather events. It could exacerbate water scarcity and stress, affect agricultural and food production, influence natural ecosystems, and lead to extreme weather events such as floods and droughts [118]. The studies on climate model simulations [108], [111], sewer overflow modeling [109], and sea-level research [110] collectively highlight the essential role of synthetic data in understanding and addressing climate change impacts on water systems. Effective management and adaptation strategies regarding changing climate conditions are essential to mitigate these impacts and ensure water security. Besides those directions, synthetic data can support innovative data-driven approaches for monitoring and analyzing water use (P7 in Table 14), ranging from agricultural irrigation [113] to household consumption patterns [94], [112], highlighting the importance of precise data for effective water management and policy development.

In envisioning future advancements in water policy development, integrating advanced AI technologies, specifically deep learning and reinforcement learning, presents a promising frontier. Deep learning's proficiency in deciphering complex relationships within extensive datasets can significantly enhance our understanding of intricate water management systems. Reinforcement learning, in particular, has the potential to evaluate the impacts of various policy interventions. This method could enable policymakers to simulate scenarios and optimize outcomes before actual implementation. For instance, exploring the dynamics of water pollution control policies and optimizing pollution trading programs through scenario simulation using reinforcement learning could yield substantial insights, guiding effective and adaptive water policy strategies. Pursuing such research directions aligns with the ongoing efforts of governmental bodies, like the EPA and USGS, to leverage advanced technologies for sustainable water resource management.

VII. SUMMARY AND CONCLUSION

Consider a network of sensors in a lake measuring water pH and temperature; using these GAN models, we generate synthetic data that closely mimics the spatial distribution of water pH and temperatures. We then analyze this data using PCA and t-SNE to understand the spatial relationships and to predict how a temperature change in one node might affect nearby nodes, a preeminent aspect of environmental monitoring. Our study utilizes PCA and t-SNE to visualize the diversity in synthetic data, with CTGAN and DoppelGANger demonstrating promising results.

Our work also assesses the fidelity of synthetic data using a GRU classifier. For instance, TimeGAN demonstrates slower progression to high accuracy, indicating better mimicry and accurate temporal representation of the original data. This model can generate synthetic datasets that closely resemble pollution levels for water quality management, such as in a treatment plant, allowing for the development of more accurate predictive models to ensure water quality, especially when real-world pollution data are scarce. We find that synthetic data, like that from Cramer GAN and CTGAN, can replace original data in training predictive models. In an urban water distribution network context, these GAN models generate data representing various pressure and flow

scenarios. We use this synthetic data for emergency response simulations, such as predicting the effects of a main pipe burst or the need for its preventive maintenance, aiding in efficient crisis management and resource allocation.

The correlation analysis in our study highlights the ability of models like DoppelGANger and TimeGAN to preserve spatio-temporal dependencies. Applying this to environmental impact assessments near a river, these models simulate how a new industrial project might affect water quality and/or flow. Synthetic data can help predict environmental factors and assist in regulatory compliance and sustainable development planning. The nuanced capabilities of various GAN models identified in our study, such as capturing dataset diversity, fidelity, and usefulness for predictive modeling, directly apply to water resource management. For instance, in regions facing water scarcity, choosing the suitable GAN model based on these insights leads to effective modeling of water usage scenarios, assisting in strategic planning and conservation efforts.

Overall, the findings from our study on GAN models offer valuable insights into the selection and application of these models in water utilities. From temperature monitoring in lakes to predictive modeling in water treatment and distribution and even environmental impact estimation (such as for water-related public policies), choosing a GAN model plays a vital role. We can strategically leverage each model's strengths in fidelity, data mimicry, and spatio-temporal correlation preservation to address specific challenges in water resource management and environmental monitoring nationally and globally.

APPENDIX A
GAN MODEL PARAMETERS

A. ACWA DATASET

- **CTGAN**: batch_size = 150, epochs = 101, learning_rate = 5e-5, beta_1 = 0.5, beta_2 = 0.9
- **WGAN**: noise_dim = 32, dim = 64, batch_size = 64, epochs = 101, learning_rate = 5e-5, beta_1 = 0.5, beta_2 = 0.9
- **DRAGAN**: noise_dim = 64, dim = 64, batch_size = 150, epochs = 101, learning_rate = 2e-6, beta_1 = 0.5, beta_2 = 0.9
- **WGAN-GP**: noise_dim = 64, dim = 64, batch_size = 150, epochs = 101, learning_rate = [5e-5, 1e-3], beta_1 = 0.5, beta_2 = 0.9
- **Cramer GAN**: noise_dim = 32, dim = 64, batch_size = 64, epochs = 101, learning_rate = 1e-5, beta_1 = 0.5, beta_2 = 0.9
- **TimeGAN (Default)**: seq_len = 24, n_seq = 6, hidden_dim = 24, gamma = 1, noise_dim = 32, dim = 128, batch_size = 128, learning_rate = 5e-4
- **TimeGAN (After Tuning)**: seq_len = 24, n_seq = 8, hidden_dim = 24, gamma = 1, noise_dim = 32, dim = 128, batch_size = 32, log_step = 100, learning_rate = 5e-4, Train_steps = 10000
- **DoppelGANger (Default)**: batch_size = 100, lr = 0.001, betas = (0.2, 0.9), latent_dim = 20, gp_lambda = 2, pac = 1, epochs = 400, sequence_length = 56
- **DoppelGANger (After Tuning)**: batch_size = 32, lr = 0.001, betas = (0.2, 0.9), latent_dim = 24, gp_lambda = 2, pac = 1, epochs = 1000, sequence_length = 24, sample_length = 6, rounds = 1

B. WWTP (DC WATER) DATASET

- **CTGAN**: batch_size = 250, epochs = 101, learning_rate = 5e-5, beta_1 = 0.5, beta_2 = 0.9
- **WGAN**: noise_dim = 32, dim = 128, batch_size = 128, epochs = 101, learning_rate = 5e-5, beta_1 = 0.5, beta_2 = 0.9
- **DRAGAN**: noise_dim = 128, dim = 128, batch_size = 250, epochs = 101, learning_rate = 2e-6, beta_1 = 0.5, beta_2 = 0.9
- **WGAN-GP**: noise_dim = 128, dim = 128, batch_size = 250, epochs = 101, learning_rate = [5e-5, 1e-3], beta_1 = 0.5, beta_2 = 0.9
- **Cramer GAN**: noise_dim = 32, dim = 128, batch_size = 128, epochs = 101, learning_rate = 1e-5, beta_1 = 0.5, beta_2 = 0.9
- **TimeGAN (Default)**: seq_len = 24, n_seq = 6, hidden_dim = 24, gamma = 1, noise_dim = 32, dim = 128, batch_size = 128, learning_rate = 5e-4
- **TimeGAN (After Tuning)**: seq_len = 24, n_seq = 13, hidden_dim = 24, gamma = 1, noise_dim = 32, dim = 128, batch_size = 200, log_step = 100, learning_rate = 5e-4, Train_steps = 10000
- **DoppelGANger (Default)**: batch_size = 100, lr = 0.001, betas = (0.2, 0.9), latent_dim = 20, gp_lambda = 2, pac = 1, epochs = 400, sequence_length = 56
- **DoppelGANger (After Tuning)**: batch_size = 200, lr = 0.001, betas = (0.2, 0.9), latent_dim = 24, gp_lambda = 2, pac = 1, epochs = 1000, sequence_length = 24, sample_length = 6, rounds = 1

C. BATADAL DATASET

- **CTGAN (Default)**: batch_size = 500, epochs = 501, learning_rate = 2e-4, beta_1 = 0.5, beta_2 = 0.9, critic_loss and generator_loss observations
- **CTGAN (Tuned)**: batch_size = 250, epochs = 101, learning_rate = 5e-5, beta_1 = 0.5, beta_2 = 0.9
- **WGAN (Default)**: noise_dim = 32, dim = 128, batch_size = 128, log_step = 100, epochs = 501, learning_rate = 5e-4, beta_1 = 0.5, beta_2 = 0.9, generator and discriminator loss observations
- **WGAN (Tuned)**: noise_dim = 32, dim = 128, batch_size = 128, epochs = 101, learning_rate = 5e-5, beta_1 = 0.5, beta_2 = 0.9
- **DRAGAN (Default)**: noise_dim = 128, dim = 128, batch_size = 500, epochs = 501, learning_rate = 1e-5, beta_1 = 0.5, beta_2 = 0.9, loss observations
- **DRAGAN (Tuned)**: noise_dim = 128, dim = 128, batch_size = 250, epochs = 101, learning_rate = 2e-6, beta_1 = 0.5, beta_2 = 0.9

- **TimeGAN (Default)**: seq_len = 24, n_seq = 6, hidden_dim = 24, gamma = 1, noise_dim = 32, dim = 128, batch_size = 128, learning_rate = 5e-4
- **TimeGAN (Tuned)**: seq_len = 24, n_seq = 26, hidden_dim = 24, gamma = 1, noise_dim = 32, dim = 128, batch_size = 200, log_step = 100, learning_rate = 5e-4, train_steps = 10000
- **DoppelGANger (Default)**: batch_size = 100, lr = 0.001, betas = (0.2, 0.9), latent_dim = 20, gp_lambda = 2, pac = 1, epochs = 400, sequence_length = 56
- **DoppelGANger (Tuned)**: batch_size = 200, lr = 0.001, betas = (0.2, 0.9), latent_dim = 24, gp_lambda = 2, pac = 1, epochs = 1000, sequence_length = 24, sample_length = 6, rounds = 1

ACKNOWLEDGMENT

The authors acknowledge Virginia Tech's A3 Research Laboratory (https://ai.bse.vt.edu/) for their feedback and support in the generation of ACWA data. We would also like to acknowledge the Commonwealth Cyber Initiative (CCI) for their overall provision. We acknowledge Halah Shehadah for her initial support in reviewing relevant literature during the early stages of the article. We also acknowledge the partial use of ChatGPT-4 (https://chat.openai.com/) for assistance in refining parts of the article for Sections III-B and IV, and only during the final editing stage. The views expressed in this article by the authors are solely those of the authors and do not necessarily reflect the views of DC Water or other treatment facilities, including author Wang's employer, the U.S. Food and Drug Administration (FDA). The authors declare no conflicts of interest. Data are available upon request; the ACWA dataset is publicly accessible at (https://github.com/AI-VTRC/ACWA-Data). *(Md Nazmul Kabir Sikder and Yingjie Wang are as co-first authors.)*

REFERENCES

[1] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intell. Syst.*, vol. 24, no. 2, pp. 8–12, Mar. 2009.

[2] F. A. Batarseh and A. Kulkarni, "AI for water," *Computer*, vol. 56, no. 3, pp. 109–113, Mar. 2023.

[3] D. Baltrunas, A. Elmokashfi, and A. Kvalbein, "Measuring the reliability of mobile broadband networks," in *Proc. Conf. Internet Meas. Conf.*, Nov. 2014, pp. 45–58.

[4] Z. S. Bischof, F. E. Bustamante, and N. Feamster, "Characterizing and improving the reliability of broadband internet access," 2017, *arXiv:1709.09349*.

[5] L. Chen, J. Lingys, K. Chen, and F. Liu, "AuTO: Scaling deep reinforcement learning for datacenter-scale automatic traffic optimization," in *Proc. Conf. ACM Special Interest Group Data Commun.*, Aug. 2018, pp. 191–205.

[6] R. Grandl, G. Ananthanarayanan, S. Kandula, S. Rao, and A. Akella, "Multi-resource packing for cluster schedulers," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 4, pp. 455–466, Feb. 2015.

[7] J. Jiang, R. Das, G. Ananthanarayanan, P. A. Chou, V. N. Padmanabhan, V. Sekar, E. Dominique, M. Goliszewski, D. Kukoleca, R. Vafin, and H. Zhang, "Via: Improving internet telephony call quality using predictive relay selection," in *Proc. ACM SIGCOMM Conf.*, Aug. 2016, pp. 286–299.

[8] N. Liu, Z. Li, J. Xu, Z. Xu, S. Lin, Q. Qiu, J. Tang, and Y. Wang, "A hierarchical framework of cloud resource allocation and power management using deep reinforcement learning," in *Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Jun. 2017, pp. 372–382.

[9] H. Mao, M. Alizadeh, I. Menache, and S. Kandula, "Resource management with deep reinforcement learning," in *Proc. 15th ACM Workshop Hot Topics Netw.*, Nov. 2016, pp. 50–56.

[10] B. Montazeri, Y. Li, M. Alizadeh, and J. Ousterhout, "Homa: A receiver-driven low-latency transport protocol using network priorities," in *Proc. Conf. ACM Special Interest Group Data Commun.*, Aug. 2018, pp. 221–235.

[11] S. Sundaresan, X. Deng, Y. Feng, D. Lee, and A. Dhamdhere, "Challenges in inferring internet congestion using throughput measurements," in *Proc. Internet Meas. Conf.*, Nov. 2017, pp. 43–56.

[12] N. Tuptuk, P. Hazell, J. Watson, and S. Hailes, "A systematic review of the state of cyber-security in water systems," *Water*, vol. 13, no. 1, p. 81, Jan. 2021.

[13] T. McGregor, S. Alcock, and D. Karrenberg, "The RIPE NCC internet measurement data repository," in *Proc. Int. Conf. Passive Act. Netw. Meas.*, Jan. 2010, pp. 111–120.

[14] S. Antonatos, K. G. Anagnostakis, and E. P. Markatos, "Generating realistic workloads for network intrusion detection systems," in *Proc. 4th Int. Workshop Softw. Perform.*, Jan. 2004, pp. 207–215.

[15] Y. Denneulin, E. Romagnoli, and D. Trystram, "A synthetic workload generator for cluster computing," in *Proc. 18th Int. Parallel Distrib. Process. Symp.*, Apr. 2004, pp. 243–250.

[16] S. Di, D. Kondo, and F. Cappello, "Characterizing and modeling cloud applications/jobs on a Google data center," *J. Supercomput.*, vol. 69, no. 1, pp. 139–160, Jul. 2014.

[17] A. Ganapathi, Y. Chen, A. Fox, R. Katz, and D. Patterson, "Statistics-driven workload modeling for the cloud," in *Proc. IEEE 26th Int. Conf. Data Eng. Workshops (ICDEW)*, Mar. 2010, pp. 87–92.

[18] D.-C. Juan, L. Li, H.-K. Peng, D. Marculescu, and C. Faloutsos, "Beyond Poisson: Modeling inter-arrival time of requests in a datacenter," in *Proc. Adv. Knowl. Discovery Data Mining: 18th Pacific-Asia Conf.*, Tainan, Taiwan, Jan. 2014, pp. 198–209.

[19] T. Li and J. Liu, "Cluster-based spatiotemporal background traffic generation for network simulation," *ACM Trans. Model. Comput. Simul.*, vol. 25, no. 1, pp. 1–25, Jan. 2015.

[20] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using GAN for improved liver lesion classification," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 289–293.

[21] C. Zhang, S. R. Kuppannagari, R. Kannan, and V. K. Prasanna, "Generative adversarial network for synthetic time series data generation in smart grids," in *Proc. IEEE Int. Conf. Commun., Control, Comput. Technol. Smart Grids (SmartGridComm)*, Oct. 2018, pp. 1–6.

[22] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2019, pp. 7333–7343.

[23] C. Bowles, L. Chen, R. Guerrero, P. Bentley, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernández, J. Wardlaw, and D. Rueckert, "GAN augmentation: Augmenting training data using generative adversarial networks," 2018, *arXiv:1810.10863*.

[24] S. Assefa, "Generating synthetic data in finance: Opportunities, challenges and pitfalls," in *Proc. 1st ACM Int. Conf. AI Finance*, Jan. 2020, pp. 1–8.

[25] R. Labaca-Castro, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2023, pp. 73–76.

[26] Z. Lin, A. Jain, C. Wang, G. Fanti, and V. Sekar, "Using GANs for sharing networked time series data: Challenges, initial promise, and open questions," in *Proc. ACM Internet Meas. Conf.*, Oct. 2020, pp. 464–483.

[27] *U.S. Department of Homeland Security–Industrial Control Systems-Cyber Emergency Response Team, Year in Review*, ICS-CERT, New Delhi, India, 2015.

[28] B. Walton, "Water sector prepares for cyberattacks," *Circle Blue*, vol. 9, pp. 1–12, Oct. 2016.

[29] M. Cava, "Uber to pay $148 million over undisclosed data breach that ex-CEO paid hackers to keep quiet," USA Today, New York, NY, USA, Tech. Rep., 2018.

[30] G. Rubin, "Many company hacks go undisclosed to sec despite regulator efforts," *Wall Street J.*, vol. 1, pp. 1–12, Oct. 2019.

[31] M. N. K. Sikder, M. B. T. Nguyen, E. D. Elliott, and F. A. Batarseh, "Deep H_2O: Cyber attacks detection in water distribution systems using deep learning," *J. Water Process Eng.*, vol. 52, Apr. 2023, Art. no. 103568.
[32] M. N. K. Sikder and F. A. Batarseh, "Outlier detection using AI: A survey," *AI Assurance*, vol. 2023, pp. 231–291, Jan. 2023.
[33] F. A. Batarseh and L. Freeman, *AI Assurance: Towards Trustworthy, Explainable, Safe, And Ethical AI*. Alpharetta, GA, USA: Elsevier, 2022.
[34] A. Erba, R. Taormina, S. Galelli, M. Pogliani, M. Carminati, S. Zanero, and N. O. Tippenhauer, "Constrained concealment attacks against reconstruction-based anomaly detectors in industrial control systems," in *Proc. Annu. Comput. Secur. Appl. Conf.*, Dec. 2020, pp. 480–495.
[35] P. B. Cheung, J. E. Van Zyl, and L. F. R. Reis, "Extension of EPANET for pressure driven demand modeling in water distribution system," *Comput. Control Water Ind.*, vol. 1, no. 1, pp. 16–311, 2005.
[36] R. Taormina et al., "Battle of the attack detection algorithms: Disclosing cyber attacks on water distribution networks," *J. Water Resour. Planning Manage.*, vol. 144, no. 8, Aug. 2018, Art. no. 04018048.
[37] K. E. Emam, L. Mosquera, and R. Hoptroff, *Practical Synthetic Data Generation: Balancing Privacy And The Broad Availability Of Data*. Sebastopol, CA, USA: O'Reilly Media, 2020.
[38] J. Lin, C. Sreng, E. Oare, and F. A. Batarseh, "NeuralFlood: An AI-driven flood susceptibility index," in *Proc. AAAI Symp. Ser.*, pp. 94–101, vol. 2, no. 1, Jan. 2024.
[39] M. Sophocleous, "Global and regional water availability and demand: Prospects for the future," *Natural Resour. Res.*, vol. 13, no. 2, pp. 61–75, Jun. 2004.
[40] G. Wang, Q.-S. Jia, M. Zhou, J. Bi, J. Qiao, and A. Abusorrah, "Artificial neural networks for water quality soft-sensing in wastewater treatment: A review," *Artif. Intell. Rev.*, vol. 55, no. 1, pp. 565–587, Jan. 2022.
[41] M. S. Rahim, K. A. Nguyen, R. A. Stewart, D. Giurco, and M. Blumenstein, "Machine learning and data analytic techniques in digital water metering: A review," *Water*, vol. 12, no. 1, p. 294, Jan. 2020.
[42] A. Kulkarni, M. Yardimci, M. N. K. Sikder, and F. A. Batarseh, "P_2O: AI-driven framework for managing and securing wastewater treatment plants," *J. Environ. Eng.*, vol. 149, no. 9, Sep. 2023, Art. no. 04023045.
[43] F. A. Batarseh, M. O. Yardimci, R. Suzuki, M. N. K. Sikder, Z. Wang, and W. Mao, "Realtime management of wastewater treatment plants using AI," Water Res. Found., Denver, CO, USA, Tech. Rep., 2022.
[44] F. A. Batarseh, A. Kulkarni, C. Sreng, J. Lin, and S. Maksud, "ACWA: An AI-driven cyber-physical testbed for intelligent water systems," *Water Pract. Technol.*, vol. 18, no. 12, pp. 3399–3418, Dec. 2023.
[45] F. Batarseh, A. Kulkarni, C. Sreng, J. Lin, and S. Maksud. (2023). *ACWA Data*. [Online]. Available: https://github.com/AI-VTRC/ACWA-Data
[46] A. P. Mathur and N. O. Tippenhauer, "SWaT: A water treatment testbed for research and training on ICS security," in *Proc. Int. Workshop Cyber-Phys. Syst. Smart Water Netw. (CySWater)*, Apr. 2016, pp. 31–36.
[47] C. M. Ahmed, V. R. Palleti, and A. P. Mathur, "WADI: A water distribution testbed for research in the design of secure cyber physical systems," in *Proc. 3rd Int. Workshop Cyber-Phys. Syst. Smart Water Netw.*, Apr. 2017, pp. 25–28.
[48] M. A. Zolfagharipoor and A. Ahmadi, "A decision-making framework for river water quality management under uncertainty: Application of social choice rules," *J. Environ. Manage.*, vol. 183, pp. 152–163, Dec. 2016.
[49] K. Jenkins, "Synthetic data and public policy: Supporting real-world policymakers with algorithmically generated data," *Policy Quart.*, vol. 19, no. 2, pp. 29–39, May 2023.
[50] C. White. (2021). *Open Data: An Overview of Current Policies, Benefits, and Challenges of Requiring Water Data Integration*. Accessed: Nov. 15, 2023. [Online]. Available: https://internetofwater.org/blog/open-data-policies/
[51] EPANET. (1998). *Office of Water Data Integration Efforts. Audit Report No. E1nwg6-15-0001-8100177*. Accessed: Nov. 15, 2023. [Online]. Available: https://www.epa.gov/sites/default/files/2015-09/documents/8100177_0.pdf
[52] N. Ramirez. (2019). *The Great Big List of Data Privacy Laws by State*. Accessed: Nov. 15, 2023. [Online]. Available: https://www.osano.com/articles/data-privacy-laws-by-state
[53] E. Salomons, L. Sela, and M. Housh, "Hedging for privacy in smart water meters," *Water Resour. Res.*, vol. 56, no. 9, p. 2020, Sep. 2020.
[54] M. Wilchek and Y. Wang, "Synthetic differential privacy data generation for revealing bias modelling risks," in *Proc. IEEE Int. Conf. Parallel Distrib. Process. Appl., Big Data Cloud Comput., Sustain. Comput. Commun., Social Comput. Netw. (ISPA/BDCloud/SocialCom/SustainCom)*, Sep. 2021, pp. 1574–1580.
[55] M. N. K. Sikder, "AI methods for anomaly detection in cyber-physical systems: With application to water and agriculture," Ph.D. dissertation, Virginia Tech, Arlington, VA, USA, Tech. Rep., 2025. [Online]. Available: https://hdl.handle.net/10919/124470
[56] J. Yoon, D. Jarrett, and M. van der Schaar, "Time-series generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Sep. 2019, pp. 5508–5518.
[57] M. Ring, D. Schlör, D. Landes, and A. Hotho, "Flow-based network traffic generation using generative adversarial networks," *Comput. Secur.*, vol. 82, pp. 156–172, May 2019.
[58] A. Desai, C. Freeman, Z. Wang, and I. Beaver, "TimeVAE: A variational auto-encoder for multivariate time series generation," 2021, *arXiv:2111.08095*.
[59] N. Kodali, J. Abernethy, J. Hays, and Z. Kira, "On convergence and stability of GANs," 2017, *arXiv:1705.07215*.
[60] M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos, "The Cramer distance as a solution to biased Wasserstein gradients," 2017, *arXiv:1705.10743*.
[61] A. C. Belkina, C. O. Ciccolella, R. Anno, R. Halpert, J. Spidlen, and J. E. Snyder-Cappione, "Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets," *Nature Commun.*, vol. 10, no. 1, p. 5415, Nov. 2019.
[62] S. Wold, K. H. Esbensen, and P. Geladi, "Principal component analysis," *Anal. methods*, vol. 2, nos. 1–3, pp. 37–52, Aug. 1987.
[63] L. van der Maaten and G. E. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, Jan. 2008.
[64] C. Esteban, S. L. Hyland, and G. Rätsch, "Real-valued (Medical) time series generation with recurrent conditional GANs," 2017, *arXiv:1706.02633*.
[65] A. Hassanzadeh, A. Rasekh, S. Galelli, M. Aghashahi, R. Taormina, A. Ostfeld, and M. K. Banks, "A review of cybersecurity incidents in the water sector," *J. Environ. Eng.*, vol. 146, no. 5, May 2020, Art. no. 03120003.
[66] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, nos. 3–4, pp. 211–407, 2014.
[67] N. Patki, R. Wedge, and K. Veeramachaneni, "The synthetic data vault," in *Proc. IEEE Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2016, pp. 399–410.
[68] T. Sarkar, "Synthetic data generation-a must-have skill for new data scientists," *Towards Data Sci.*, vol. 19, pp. 1–12, Dec. 2018.
[69] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1153–1176, 2nd Quart., 2016.
[70] A. Gautam, M. Sit, and I. Demir, "Realistic river image synthesis using deep generative adversarial networks," *Frontiers Water*, vol. 4, Feb. 2022, Art. no. 784441.
[71] R. K. Mazumder, G. Modanwal, and Y. Li, "Synthetic data generation using generative adversarial network for burst failure risk analysis of oil and gas pipelines," *ASCE-ASME J. Risk Uncertainty Eng. Syst., B, Mech. Eng.*, vol. 9, no. 3, pp. 1–21, Sep. 2023.
[72] A. E. Bakhshipour, A. Koochali, U. Dittmer, A. Haghighi, S. Ahmad, and A. Dengel, "A Bayesian generative adversarial network (GAN) to generate synthetic time-series data, application in combined sewer flow prediction," 2023, *arXiv:2301.13733*.
[73] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2017, pp. 214–223.
[74] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Dec. 2017, pp. 5769–5779.
[75] R. Sauber-Cole and T. M. Khoshgoftaar, "The use of generative adversarial networks to alleviate class imbalance in tabular data: A survey," *J. Big Data*, vol. 9, no. 1, p. 98, Aug. 2022.
[76] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.
[77] L. Aviñó, M. Ruffini, and R. Gavaldà, "Generating synthetic but plausible healthcare record datasets," 2018, *arXiv:1807.01514*.
[78] G. Cormode, C. Procopiuc, D. Srivastava, E. Shen, and T. Yu, "Differentially private spatial decompositions," in *Proc. IEEE 28th Int. Conf. Data Eng.*, Apr. 2012, pp. 20–31.

[79] Y. Sun, A. Cuesta-Infante, and K. Veeramachaneni, "Learning vine copula models for synthetic data generation," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, vol. 33, no. 1, pp. 5049–5057.

[80] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data synthesis based on generative adversarial networks," 2018, *arXiv:1806.03384*.

[81] S. T. K. Jan, Q. Hao, T. Hu, J. Pu, S. Oswal, G. Wang, and B. Viswanath, "Throwing darts in the dark? Detecting bots with limited data using neural data augmentation," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2020, pp. 1190–1206.

[82] W. Zhou, X.-M. Kong, K. Li, X.-M. Li, L.-L. Ren, Y. Yan, Y. Sha, X. Cao, and X. Liu, "Attack sample generation algorithm based on data association group by GAN in industrial control dataset," *Comput. Commun.*, vol. 173, pp. 206–213, May 2021.

[83] T. Fearn, "Probabilistic principal component analysis," *NIR news*, vol. 25, no. 3, p. 23, May 2014.

[84] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete algorithms*, Jan. 2007, pp. 1027–1035.

[85] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[86] R. Cahuantzi, X. Chen, and S. Güttel, "A comparison of LSTM and GRU networks for learning symbolic sequences," in *Proc. Sci. Inf. Conf.*, Jan. 2021, pp. 771–785.

[87] S. Das, A. Tariq, T. Santos, S. S. Kantareddy, and I. Banerjee, "Recurrent neural networks (RNNs): Architectures, training tricks, and introduction to influential research," in *Neuromethods*, 2023, pp. 117–138.

[88] S. R. Krishnan, M. K. Nallakaruppan, R. Chengoden, S. Koppu, M. Iyapparaja, J. Sadhasivam, and S. Sethuraman, "Smart water resource management using artificial intelligence—A review," *Sustainability*, vol. 14, no. 20, p. 13384, Oct. 2022.

[89] A. Cantor, M. Kiparsky, R. Kennedy, S. Hubbard, R. Bales, L. C. Pecharroman, K. Guivetchi, C. McCready, and G. Darling, "Data for water decision making: Informing the implementation of California's open and transparent water data act through research and engagement," Tech. Rep., 2018.

[90] R. Altenburger et al., "Future water quality monitoring: Improving the balance between exposure and toxicity assessments of real-world pollutant mixtures," *Environ. Sci. Eur.*, vol. 31, no. 1, pp. 1–17, Dec. 2019, Art. no. 12.

[91] R. Muñoz, S. A. Vaghefi, F. Drenkhan, M. J. Santos, D. Viviroli, V. Muccione, and C. Huggel, "Assessing water management strategies in data-scarce mountain regions under uncertain climate and socio-economic changes," *Water Resour. Manage.*, pp. 4083–4100, Apr. 2024.

[92] M. Thyer, H. Gupta, S. Westra, D. McInerney, H. R. Maier, D. Kavetski, A. Jakeman, B. Croke, C. Simmons, D. Partington, M. Shanafield, and C. Tague, "Virtual hydrological laboratories: Developing the next generation of conceptual models to support decision making under change," *Water Resour. Res.*, vol. 60, no. 4, p. 2022, Apr. 2024.

[93] A. Cantor, M. Kiparsky, S. S. Hubbard, R. Kennedy, L. C. Pecharroman, K. Guivetchi, G. Darling, C. McCready, and R. Bales, "Making a water data system responsive to information needs of decision makers," *Frontiers Climate*, vol. 3, Nov. 2021, Art. no. 761444.

[94] D. T. Kofinas, A. Spyropoulou, and C. S. Laspidou, "A methodology for synthetic household water consumption data generation," *Environ. Model. Softw.*, vol. 100, pp. 48–66, Feb. 2018.

[95] K. L. Bonney, K. A. Klise, J. W. Poff, S. Rivera, I. Searles, and M. Chester, "Data-informed synthetic networks of water distribution systems for resilience analysis in Puerto Rico," *Water*, vol. 16, no. 23, p. 3356, Nov. 2024.

[96] U.S. Environmental Protection Agency. (2023). *Water Topics*. Accessed: Nov. 21, 2023. [Online]. Available: https://www.epa.gov/environmental-topics/water-topics

[97] X. Yang, T. A. Buscheck, K. Mansoor, Z. Wang, K. Gao, L. Huang, D. Appriou, and S. A. Carroll, "Assessment of geophysical monitoring methods for detection of brine and CO_2 leakage in drinking water aquifers," *Int. J. Greenhouse Gas Control*, vol. 90, Nov. 2019, Art. no. 102803.

[98] L. Godo-Pla, P. Emiliano, F. Valero, M. Poch, G. Sin, and H. Monclús, "Predicting the oxidant demand in full-scale drinking water treatment using an artificial neural network: Uncertainty and sensitivity analysis," *Process Saf. Environ. Protection*, vol. 125, pp. 317–327, May 2019.

[99] S. Eggimann, L. Mutzner, O. Wani, M. Y. Schneider, D. Spuhler, M. M. de Vitry, P. Beutler, and M. Maurer, "The potential of knowing more: A review of data-driven urban water management," *Environ. Sci. Technol.*, vol. 51, no. 5, pp. 2538–2553, Mar. 2017.

[100] L. J. P. Snip, X. Flores-Alsina, I. Aymerich, S. Rodríguez-Mozaz, D. Barceló, B. G. Plósz, L. Corominas, I. Rodriguez-Roda, U. Jeppsson, and K. V. Gernaey, "Generation of synthetic influent data to perform (micro)pollutant wastewater treatment modelling studies," *Sci. Total Environ.*, vols. 569–570, pp. 278–290, Nov. 2016.

[101] G. Fu, Y. Jin, S. Sun, Z. Yuan, and D. Butler, "The role of deep learning in urban water management: A critical review," *Water Res.*, vol. 223, Sep. 2022, Art. no. 118973.

[102] S. Ahmed, M. Mahzabin, S. Shahpar, S. I. Tonni, and M. S. Rahman, "Assessment of water quality in smart city environment leveraging ML-IoT," in *Proc. Int. Conf. 4th Ind. Revolution Beyond*, Jan. 2022, pp. 215–227.

[103] N. Ahmad, M. Chester, E. Bondank, M. Arabi, N. Johnson, and B. L. Ruddell, "A synthetic water distribution network model for urban resilience," *Sustain. Resilient Infrastructure*, vol. 7, no. 5, pp. 333–347, Sep. 2022.

[104] J. Liu, "A synthetic data-driven solution for urban drinking water source management," Master's thesis, Norwegian Univ. Sci. Technol. (NTNU), Trondheim, Norway, 2023.

[105] Y. Hu, S. Yang, H. Cheng, and S. Tao, "Systematic evaluation of two classical receptor models in source apportionment of soil heavy metal (loid) pollution using synthetic and real-world datasets," *Environ. Sci. Technol.*, vol. 56, no. 24, pp. 17604–17614, 2022.

[106] J. Kravitz, M. Matthews, L. Lain, S. Fawcett, and S. Bernard, "Potential for high fidelity global mapping of common inland water quality products at high spatial and temporal resolutions based on a synthetic data and machine learning approach," *Frontiers Environ. Sci.*, vol. 9, Mar. 2021, Art. no. 587660.

[107] M. Y. Chia, C. H. Koo, Y. F. Huang, W. Di Chan, and J. Y. Pang, "Artificial intelligence generated synthetic datasets as the remedy for data scarcity in water quality index estimation," *Water Resour. Manage.*, vol. 37, no. 15, pp. 1–16, Dec. 2023.

[108] N. Y. Krakauer and B. M. Fekete, "Are climate model simulations useful for forecasting precipitation trends? Hindcast and synthetic-data experiments," *Environ. Res. Lett.*, vol. 9, no. 2, Jan. 2014, Art. no. 024009.

[109] D. Bendel, F. Beck, and U. Dittmer, "Modeling climate change impacts on combined sewer overflow using synthetic precipitation time series," *Water Sci. Technol.*, vol. 68, no. 1, pp. 160–166, Jul. 2013.

[110] P. J. Watson, "Development of a unique synthetic data set to improve sea-level research and understanding," *J. Coastal Res.*, vol. 313, pp. 758–770, May 2015.

[111] Y. Wang, J. Chandrasekaran, F. Haberkorn, Y. Dong, M. Gopinath, and F. A. Batarseh, "DeepFarm: AI-driven management of farm production using explainable causality," in *Proc. IEEE 29th Annu. Softw. Technol. Conf. (STC)*, Oct. 2022, pp. 27–36.

[112] M. C. Santos, A. I. Borges, D. R. Carneiro, and F. J. Ferreira, "Synthetic dataset to study breaks in the consumer's water consumption patterns," in *Proc. 4th Int. Conf. Math. Statist.*, Jun. 2021, pp. 59–65.

[113] T. Foster, T. Mieno, and N. Brozovıı, "Satellite-based monitoring of irrigation water use: Assessing measurement errors and their implications for agricultural water management policy," *Water Resour. Res.*, vol. 56, no. 11, p. 2020, Nov. 2020.

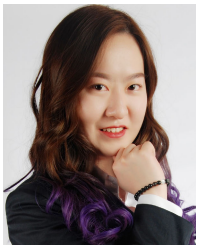
[114] U.S. Environmental Protection Agency. (2023). *Drinking Water Regulations*. Accessed: Nov. 21, 2023. [Online]. Available: https://www.epa.gov/dwreginfo/drinking-water-regulations

[115] U.S. EPA Office Research and Development. (2016). *Small Drinking Water Systems Research and Development*. Accessed: Nov. 21, 2023. [Online]. Available: https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=ORD&direntryid=326870

[116] U.S. Environmental Protection Agency. (2018). *Asset Management Programs for Stormwater and Wastewater Systems*. Accessed: Nov. 21, 2023. [Online]. Available: https://www.epa.gov/sustainable-water-infrastructure/asset-management-programs-stormwater-and-wastewater-systems
[117] Government Accountability Office. (2023). *Water Quality and Protection*. Accessed: Nov. 21, 2023. [Online]. Available: https://www.gao.gov/water-quality-and-protection
[118] United Nations. (2023). *Water—At the Center of the Climate Crisis*. Accessed: Nov. 27, 2023. [Online]. Available: https://www.un.org/en/climatechange/science/climate-issues/water

MD NAZMUL KABIR SIKDER (Member, IEEE) received the B.Sc. degree in electrical and electronics engineering from Bangladesh University of Engineering and Technology, in 2015, and the M.S. and Ph.D. degrees in computer engineering from Virginia Tech, in 2022 and 2024, respectively.

He is currently a Presidential Postdoctoral Fellow at Virginia Tech, working with the Commonwealth Cyber Initiative (CCI) and the A3 Laboratory. He previously worked at BEM Controls LLC., where he contributed to AI-driven smart grid solutions, and at Grameenphone Ltd., where he worked on LTE network deployment and automation tools. His research interests include AI assurance, cybersecurity, and machine learning applications for critical infrastructure, particularly cyber-physical security in water distribution systems. His work includes AI-driven threat detection, anomaly detection in SCADA systems, and explainable AI frameworks. His research has been published in IEEE and other high-impact venues and he won the 2022 Intelligent Water Systems Challenge for AI applications in water systems. He has research interests in trustworthy AI, cyber-physical security, and AI-driven decision making.



YINGJIE WANG received the dual B.S. degree in mathematics and statistics from Pennsylvania State University, and the M.S. degree in data science and analytics from Georgetown University. She is currently pursuing the Ph.D. degree in electrical and computer engineering with Virginia Tech.

She has been a Federal Data Scientist with U.S. Food and Drug Administration (FDA), for the past five years, where she leads AI-driven initiatives in pharmaceutical quality surveillance. Her contributions include developing signal detection models, deploying natural language processing (NLP) solutions, and designing data visualization dashboards for regulatory decision-making. She has co-authored multiple publications on AI applications in public policy, cybersecurity, and pharmaceutical quality, featuring in high-impact journals and conferences. Her research interests include explainable AI (XAI) and causal AI into decision support systems to uncover cause-and-effect relationships in complex, and dynamic environments. By enhancing transparency in AI-driven recommendations, her work aims to improve decision-making under uncertainty and promote responsible AI governance.

FERAS A. BATARSEH (Senior Member, IEEE) is currently an Associate Professor with the Department of Biological Systems Engineering (BSE), Virginia Tech. He is affiliated with the Center for Advanced Innovation in Agriculture (CAIA) and the Commonwealth Cyber Initiative (CCI), Virginia Tech, and the School of Systems Biology, George Mason University (GMU). He is the Director of the AI Assurance and Applications (A3) Laboratory. He has taught AI courses at GMU, the University of Maryland Baltimore County (UMBC), Georgetown University, and George Washington University (GWU). His research has been published in leading journals and international conferences, and he has authored multiple book chapters and books. His three recent books, *Federal Data Science*, *Data Democracy*, and *AI Assurance*, were published by Elsevier Academic Press. His research interests include AI assurance, cyberbiosecurity, intelligent water systems, and AI for agricultural policies.

Dr. Batarseh is an Active Member of the Agricultural and Applied Economics Association (AAEA) and the Association for the Advancement of Artificial Intelligence (AAAI).