



UvA-DARE (Digital Academic Repository)

DiTEC-WDN: A Large-Scale Dataset of Hydraulic Scenarios across Multiple Water Distribution Networks

Truong, Huy; Tello, Andrés; Lazovik, Alexander; Degeler, Victoria

DOI

[10.1038/s41597-025-06026-0](https://doi.org/10.1038/s41597-025-06026-0)

Publication date

2025

Document Version

Final published version

Published in

Scientific Data

License

CC BY

[Link to publication](#)

Citation for published version (APA):

Truong, H., Tello, A., Lazovik, A., & Degeler, V. (2025). DiTEC-WDN: A Large-Scale Dataset of Hydraulic Scenarios across Multiple Water Distribution Networks. *Scientific Data*, 12, Article 1733. <https://doi.org/10.1038/s41597-025-06026-0>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (<https://dare.uva.nl>)



OPEN

DATA DESCRIPTOR

DiTEC-WDN: A Large-Scale Dataset of Hydraulic Scenarios across Multiple Water Distribution Networks


Huy Truong^{1,3}  , Andrés Tello^{1,3}  , Alexander Lazovik¹ & Victoria Degeler² 

Privacy restrictions hinder the sharing of real-world Water Distribution Network (WDN) models, limiting the application of emerging data-driven machine learning, which typically requires extensive observations. To address this challenge, we propose the dataset DiTEC-WDN that comprises 36,000 unique scenarios simulated over either short-term (24 hours) or long-term (1 year) periods. We constructed this dataset using an automated pipeline that optimizes crucial parameters (e.g., pressure, flow rate, and demand patterns), facilitates large-scale simulations, and records discrete, synthetic but hydraulically realistic states under standard conditions via rule validation and post-hoc analysis. With a total of 228 million generated graph-based states, DiTEC-WDN can support a variety of machine-learning tasks, including graph-level, node-level, and link-level regression, as well as time-series forecasting. This contribution, released under a public license, encourages open scientific research in the critical water sector, eliminates the risk of exposing sensitive data, and fulfills the need for a large-scale water distribution network benchmark for study comparisons and scenario analysis.

Background & Summary

Water Distribution Networks (WDNs) are considered critical infrastructures as they provide clean and safe water to humans, which is one of the Sustainable Development Goals proposed by the United Nations. Water providers have to deal with critical challenges during the design, planning, and management phases of a WDN in order to fulfill this goal, such as adaptability and robustness to an ever changing environment. Climate, consumer behavior, aging infrastructure, failures, all can lead to drastic changes in the conditions under which WDNs must continue working adequately. Monitoring of WDN operations plays an important role in guaranteeing the water supply. The state of the network must be known at any given time to prevent unwanted situations, e.g., pipe leaks.

Hydraulic modeling has been the most straightforward approach for practitioners to simulate the WDN dynamics and aid design, planning and management. While pure physics-based hydraulic modeling is still being commonly used in the water domain, water engineering research and practice are experiencing a shift towards hybrid data-driven approaches. Such approaches combine the power of physics and mathematical simulation tools with data-driven deep learning models to solve water engineering problems. Paradoxically, while data are the key to such approaches, WDN operation data are scarce and seldom shared among practitioners and researchers due to privacy, safety and other domain related constraints^{1,2}. A notable example is nodal demand patterns. Demand is one of the most important inputs for solving the WDN hydraulics³. Surprisingly, it is one of the inputs that is rarely found in the WDN asset description files. It is common to find just a few demand patterns reused many times on several nodes in the network⁴, or no demand patterns at all⁵. The stochastic nature of water demand explains some of the uncertainties found in WDNs⁶, which should be properly modeled and considered in the simulations⁷. Hence, reusing the same demand patterns on multiple nodes assumes that several users/consumers have exactly the same water consumption behavior, which is unrealistic. This not only

¹Bernoulli Institute, University of Groningen, Groningen, The Netherlands. ²Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands. ³These authors contributed equally: Huy Truong, Andrés Tello. e-mail: h.c.truong@rug.nl; andres.tello@rug.nl

harms the robustness of the models to uncertainties, but also limits the variety of the data, which is especially important for deep learning data-driven approaches.

Benchmark datasets for WDN data analysis are, in fact, very limited^{8,9}. LeakDB⁸ is a dataset commonly used in research at the moment, but it only includes a single small WDN with limited variability of scenarios, as explained in Section Technical Validation. This limits the diversity of the data for training data-driven models. More commonly found in practice is a collection of static asset descriptions of water networks, and different algorithms and implementations for data generation from them. Researchers and practitioners working on data-driven approaches for water engineering lack data to train their models, and count on static asset descriptions of the WDN rather than operational data. Those asset descriptions are represented as configuration files, which serve as input to physics-based mathematical tools to simulate the data required for data-driven models' training. Although the simulation of WDN hydraulics from well-defined configuration files seems to be straightforward, it is a cumbersome process that involves expert knowledge, time-consuming models' calibration, uncertainties, and computational complexity, among other challenges. Moreover, such configuration files only allow practitioners to simulate the WDN states determined by the input parameters explicitly specified. Hence, if new data for a different WDN is needed, or a different condition in the input parameters needs to be evaluated, the whole process has to be repeated from scratch.

Contrary to previous approaches, we provide operational data, which are ready-to-use for model training purposes. The aim of this work is to support the shift towards data-driven approaches for WDN data analysis. We provide a multi-purpose dataset generated based on 36 publicly available WDNs, which includes 228 million network state snapshots, when operating under normal conditions. The synthetic nature of the data eliminates the privacy and safety concerns, facilitating data sharing among researchers, and within the commercial sector, without any risks. Even the models trained on our data can be shared because the models will learn generic patterns which are not attached to any particular clients or real water utility assets. Those pre-trained models can be adapted to solve use-case specific problems at a later stage. This approach can lower the institutional barriers and enhance collaboration between water utilities. While data and models can be unrestrictedly shared among practitioners and researchers, the way models are applied and sensitive use-case data will still be private and safe within each water company or research institution.

Previous work demonstrates the transferability of deep learning models trained on synthetic data to real-world applications¹⁰. It shows that a model trained on synthetic data performs well on real scenarios for pressure estimation with data provided by a water utility company. In addition, it shows that real operational conditions can be incorporated by using them to fine-tune a model pre-trained on synthetic data to adapt it to the real use-case data. The examples of tasks supported by our dataset are surrogate modeling, state estimation, and demand forecasting. The data provided includes all the inputs used for the simulations and their respective outputs, which allows researchers to work on surrogate modeling of physics-based mathematical simulation tools. The large number of provided snapshots allows practitioners to work on state estimation models. The data include unique demand patterns per node, facilitating demand forecasting.

Methods

To create operational data, we begin by collecting available WDNs represented as EPANET input files. The detailed procedure and file format are explained in the Data acquisition section. Based on these inputs, we implement a generation pipeline to produce synthetic scenarios. The pipeline first extracts available hydraulic parameters, using their fields to construct an optimization configuration, and their values to record in a global profiler, as detailed in the Preprocessing step. The configuration defines sampling boundary values for each parameter field, which serve as objectives of an optimization process. Throughout this process, the global profiler guides the selection and validation of potential sampling values, as described in the Hydraulic Sampling Parameters Optimization (HSPO) Section. The optimized configuration is used to sample parameter values that are then fed into EPANET to simulate scenarios. Finally, only error-free scenarios are retained and packaged into a compressed format, as outlined in the Simulation section. We now describe each step in detail.

Data acquisition. The described synthetic dataset was created based on publicly available WDNs. In order to achieve this, we collected data related to the topology and the physical properties of the networks' components. As mentioned before, such information is available as configuration files describing the assets of the WDNs. Initially, we collected the asset description data of 55 WDNs. In those initial files, we found duplicated data related to the same WDN but under different names. We also found configuration files with unreadable characters which did not allow a proper data reading. After a data depuration process, we included 36 WDNs in our final dataset. The full list of the WDNs included in our dataset, and their main components, is shown in the Supplementary Table S1.

Our data is generated using the EPANET¹¹ and WNTR¹² physics-based simulation tools, which allow us to run simulations of the hydraulic behavior of WDNs. These tools are widely used by researchers and practitioners in the water domain. All collected configuration files are represented in EPANET input file format (.inp).

The input file contains the metadata about the WDN and the description of the components of the network, the system's operation, water quality, and other options used at simulation time. The file is organized in sections, where each section begins with a keyword enclosed in brackets. For example, the sections related to the network components include: [TITLE], [JUNCTIONS], [RESERVOIRS], [TANKS], [PIPES], [PUMPS], [VALVES], and [EMITTERS]. The sections related to the system's operation include: [CURVES], [PATTERNS], [ENERGY], [STATUS], [CONTROLS], [RULES], and [DEMANDS]. The complete description of the input file format can be found in the EPANET 2.2 User Manual¹³.

In the context of this work, the EPANET input file represents the input to the data generation process. Accordingly, each section represents a collection of parameters that needs to be optimized in order to obtain a simulation outcome that is considered to be a valid state of the network. The [PATTERNS] section is used to

Components	Parameter	Type	Unit	Global Range/States
Head pump, Power pump, Pipe, PRV, PSV, FCV, TCV	Initial Status	Static (Category)	—	Closed/Opened/Active/CV
Head pump, Power pump	Base speed	Static (Float)	—	[0.9, 1.0]
Head pump, Power pump	Efficiency X	Curve	SIFU ^a	[0.0, 0.5]
Head pump, Power pump	Efficiency Y	Curve	%	[0.0, 77.0]
Head pump	Pump curve X	Curve	SIFU	[0.0, 0.88]
Head pump	Pump curve Y	Curve	m	[0.0, 211.02]
Head pump	Energy pattern	Pattern	kW-hours	[0.024093, 0.1234]
Power pump	Power	Static (Float)	kW	[372.85, 186424.97]
Pipe	Diameter	Static (Float)	m	[0.0010, 5.1816]
Pipe	Minor loss	Static (Float)	—	[0, 1000]
Pipe	Roughness	Static (Float)	mm (DW ^b) - (Otherwise)	[0.0015, 8333.3333]
Pipe	Length	Static (Float)	m	[0.01, 17003.20]
PRV	Initial Setting	Static (Float)	m	[0.0, 154.75]
PSV	Initial Setting	Static (Float)	m	[38.69, 49.23]
FCV	Initial Setting	Static (Float)	SIFU	[0.0, 0.9]
TCV	Initial Setting	Static (Float)	—	[0.0, 403101800000]
Tank	Elevation	Static (Float)	m	[2.00, 571.12]
Tank	Diameter	Static (Float)	m	[0.3048, 58.309]
Tank	Initial level	Static (Float)	m	[0.50, 548.64]
Tank	Minimum volume	Static (Float)	m ³	[0.000, 95965.597]
Junction	Input demand	Pattern	SIFU	[-1.388, 4.814]
Junction	Elevation	Static (Float)	m	[0., 154.75]
Reservoir	Base head	Static (Float)	m	[0, 500]
Reservoir	Head pattern	Pattern	m	[0.91, 70.42]

Table 1. List of available hydraulic parameters. ^aSIFU stands for SI Flow Units including LPS, LPM, MLD, CMH, and CMD. ^bDM refers to Darcy Weisbach headloss equation.

specify the water consumption patterns associated with each junction node. The pattern is represented as a list of values, where each element of the list represents the water consumption at time step t . Another important section is [TIMES], where we can specify the duration of the simulation and the time step, i.e., the sampling rate of the simulation's outputs.

At runtime, the simulation generates a set of outputs corresponding to time step t , which is the state of the network at such given time. In our work, each network state is called a *snapshot*. The collection of snapshots that span the entire duration of the simulation is called a *scenario*. This dataset includes 10 WDNs where each scenario spans 24 hours, and 26 WDNs where each scenario spans 1 year, with a 1-hour time step in both cases. The complete dataset comprises 1,000 scenarios per network, which represent 228 million snapshots of water networks' states.

Data generation. Following the network collection, we present a data generation scheme. Overall, the scheme involves three subsequent steps: Data Preprocessing, Hydraulic Parameter Optimization, and Simulation. The first step filters targeted simulation parameters and collects statistics across available networks. Both are then fed into an optimization algorithm to determine the sampling strategy and corresponding bounds for specific parameters. The last step plays a role in sampling concrete values, performing simulation, and encapsulating the data in a compressed format. The following sections explain each step in detail.

Preprocessing step. A static network description from an input file, described in Section Data Acquisition, contains useful simulation-oriented data and irrelevant information, such as titles, labels, and water quality parameters. Since this study focuses on hydraulic-related parameters, it is crucial to filter and refine only this specific data before proceeding to the next stage. Table 1 indicates selected parameters and their corresponding information. For some parameters, the original measurement units vary depending on the geographical region of each water network. For example, demand is measured in liters per second (LPS) in *hanoi* WDN, but in gallons per minute (GPM) in *ky8* WDN. For the sake of consistency, they are converted to the corresponding International System of Units (SI system) using the wrapper tool WNTR¹².

These selected parameters are then stored in a YAML configuration file. It is similar to the input file but contains essential metadata for both optimization and simulation phases, such as computed duration, time step, and names of skipped nodes. The configuration also records the sampling strategy and bounds for available parameters in a specific network. This metadata is included in the final delivery for reproducibility purposes.

Besides the configuration files, another type of information is computed by the profiler, calculating statistics of those 38 parameters collected from the original water networks. For each parameter, the profiler captures the *minimum*, *maximum*, *mean*, *standard deviation*, *first quartile*, *third quartile*, *parameter dimension*, and the *number of components* that can obtain this parameter. The statistics are computed for each baseline network and,

additionally, for the global network representing the overall perspective. We leveraged them to 1) determine the sampling range and size and 2) perform data imputation in case of missing values in the following step.

Hydraulic Sampling Parameters Optimization (HSPO). Consider a WDN with n nodes and m links, where each node and edge can obtain three types of parameters: static, pattern, and curve. The static parameter is a scalar or categorical value assigned per component, such as *elevation* or *status*. A pattern is a time-series that typically changes throughout the scenario (e.g., *junction input demand*, *head pattern*). A curve defines the relationship between two measurements, such as a *pump curve*, which reflects a pump's operating capacity based on flow rate and head. Assume a node has s_n static parameters, p_n patterns, and c_n curves, each with a maximum length l_n , with corresponding parameters for edges represented by s_e, p_e, c_e, l_e . Given a simulation duration d , a simulation candidate lies in a space of $s_n + p_n d + c_n l_n + s_e + p_e d + c_e l_e$ dimensions. Given this high dimensionality, we consider an alternative approach: identifying a sampling strategy to define appropriate values per parameter to generate a simulation candidate while reducing the search space, but preserving the data diversity. We call this the HSPO problem.

In particular, HSPO aims to identify stable pairs of sampling strategies and value bounds for each hydraulic parameter of all components. In other words, given a baseline WDN, the goal is to model numerous network variants and validate their parameters to ensure data stability within a specified time frame. There are two time frames, corresponding to the two dataset types: short-term and long-term. The short-term dataset includes scenarios observed over a 24-hour period, while the long-term dataset covers scenarios with a span of 1 year. Both use an hourly time step for sampling. Before diving into details, we outline the following potential sampling strategies to determine the value range of a specific parameter:

- **Keep.** Following the principle “Doing nothing is better than doing anything”, this strategy preserves the parameter's state as in the baseline network. This approach significantly reduces the search space and, therefore, mitigates the optimization complexity¹⁴.
- **Series.** This strategy applies an existing series of a particular parameter across all components. For instance, *pump curve pattern* can be retrieved in the pump manual supplied by the manufacturer and applied to every pump curve within the networks. The value is then shared across all scenarios.
- **Sampling.** Given a predefined range $[min, max]$ of a particular parameter, we uniformly sample a new value for a hydraulic parameter per component. This approach ensures that every component has its own distinct value. For patterns and curves, this strategy leverages statistics from the profiler to sample series accordingly.
- **Perturbation.** For a parameter, we gather the mean and standard deviation from the baseline WDN and sample from a Gaussian distribution. This strategy is beneficial when the parameter's value is unavailable in the target network, allowing us to use values from the global perspective.
- **Factor.** We sample a scale and bias to apply a linear transformation to existing values gathered from the baseline network. This approach ensures consistency, which is essential for certain parameters. For example, three adjacent pipes should have similar diameters. In such a case, the **Factor** serves as a potential strategy, while **Sampling** and **Perturbation** cause a pipe bottleneck as a modeling anomaly in practice.
- **Substitute.** It randomly selects an existing value of the target parameter and shares it with all components. This approach also injects minor noise into the values to maintain diversity. Similar to the Factor method, it respects consistency in modeling.
- **Terrain.** This is a special strategy applicable to junctions' elevation. In particular, we employ the Diamond Square algorithm with proper noise to generate a 2D height map¹⁵. Given the nodal coordinates from the input file, we project the network onto the map to obtain new elevations.
- **Automatic Demand Generator (ADG).** This sampling strategy is specially tailored for junctions' input demands, the most crucial but scarce parameter. Due to its importance, Section Automatic Demand Generator is dedicated to describing this approach.

For each target WDN, a default blueprint configuration is set up as follows: **ADG** for junction input demand, **terrain** for junction elevation, **factor (substitute)** for pipe diameter, and **keep** for all remaining parameters. Following this, the configuration is fed into a HSPO algorithm to iteratively refine the sampling strategy and sampling values for all parameters until convergence.

Particle Swarm Optimization (PSO). Assume that **sampling** is the default generation strategy, each parameter needs a lower bound lb and upper bound ub to construct the sample space. This yields a total of $2D$ sampling parameters in the HSPO problem. To address this challenge, we chose Particle Swarm Optimization (PSO)¹⁶, a simple yet robust approach extensively validated in water-related parameter optimization tasks^{17–19}.

Mathematically, PSO opts to construct a solution $\mathbf{X} \in \mathbb{R}^{D \times 2}$, representing a sampling configuration. This configuration drives a function *gensim*: $\mathbb{R}^{D \times 2} \rightarrow \mathbb{R}^{out \times d}$ to yield a scenario corresponding to *out* measurements, each as a d -length time-series. Internally, *gensim* implicitly solves a system of equations¹² and typically produces a large batch of diverse scenarios in practice. From this perspective, only N_{cases} created scenarios are considered to evaluate a sampling solution. Nevertheless, their measurements could exhibit anomalies, such as negative pressure or time inconsistency. As the dataset is expected to be clean, this violates our assumption. To alleviate this, we form a set of rules $R = \{r_1, r_2, \dots\}$ in which each rule judges whether simulated outcomes are valid.

Formally, a binary function *validate*: $\mathbb{R}^{D \times 2} \rightarrow \{1, 0\}$ is defined as follows:

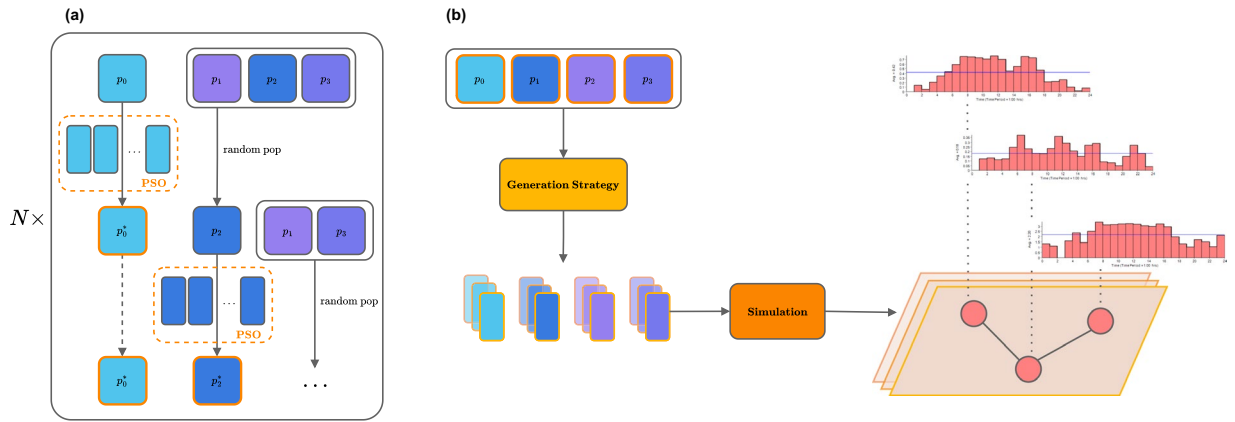


Fig. 1 Illustration of the dataset generation. The left figure (a) shows a divide-and-conquer PSO optimizing a strategy’s configuration. The right figure (b) depicts the usage of the optimized configuration to sample parameter sets and simulate diverse scenarios with unique characteristics (e.g., per-node demand patterns).

$$\text{validate}(X) = \begin{cases} 1 & \text{if } \forall r \in R, r(\text{gensim}(X)) = \text{true} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Nevertheless, empirical trials indicate that simultaneously optimizing all sampling parameters struggles to converge, more frequently for long-term cases ($t = 8,760$). This can be attributed to the large search space. To mitigate this, we facilitate PSO with a divide-and-conquer approach. As shown in Fig. 1, PSO considers a sampling set of a particular parameter while maintaining the fixed state of other sets at every timestep. This isolation reduces the complexity and makes PSO more manageable than addressing all parameters simultaneously. After an iteration, the updated optimal value for the selected parameter is fixed for the remainder of the epoch. A new PSO is then executed to optimize a random candidate from the parameter list, iterating until the list is empty. This process repeats across multiple epochs until the maximum number of epochs is reached or the intermediate solution is desired.

At each iteration, a sampling solution could be formed as a concatenation of the latest optimized and other sets. We evaluated the “goodness” of this solution by defining a fitness function $f_{\text{success}} : \mathbb{R}^{D \times 2} \rightarrow \mathbb{R}$ computing the average success rate over N_{cases} generation cases:

$$f_{\text{success}}(X) = \frac{\sum_{i=1}^{N_{\text{cases}}} \text{validate}(X)}{N_{\text{cases}}} \quad (2)$$

Considering the stochastic nature, we set N_{cases} to 100 to estimate the goodness of each sampling solution. However, merely relying on f_{success} leads to a collapse of the solution since particles tend to shrink in a local optimum, which is unrealistic and results in poor generalization. For instance, in one case of *junction elevation*, PSO proposed a narrow sampling range of [0.12, 0.12], resulting in flat terrain. To restrict such cases, a customized fitness function was designed.

Assume a particle i has its position represented as a solution $x_i \in \mathbb{R}^{D \times 2}$, we designed a fitness function $f_{\text{pso}} : \mathbb{R}^{D \times 2} \rightarrow \mathbb{R}$ as follows:

$$f_{\text{pso}}(x_i) = f_{\text{success}}(x_i)(\alpha f_{\text{ubiqr}}(x_i)) + (1-\alpha)f_{\text{range}}(x_i) \quad (3)$$

where α is a hyper-parameter balancing the two auxiliary criteria: diversity indicator f_{ubiqr} , and range expansion measurement f_{range} . While the success ratio f_{success} still plays a crucial role in assessing goodness, we encourage PSO to find optimal solutions beyond the baseline scenario. The f_{ubiqr} computes the Upper Bound of the Inter-Quartile Range (UBIQR), a statistical measure of the spread of populations²⁰. In this study, we compare the UBIQR of *junction output demand* between a generated case and the baseline, denoted as y_i and y_{bl} . For the sake of brevity, we implied a simulation executed before computing this fitness (i.e., $y_i = \text{sim}(x_i)$). Mathematically, the comparison can be written as:

$$f_{\text{ubiqr}}(y_i) = \frac{\text{UBIQR}(y_i)}{\text{UBIQR}(y_{bl})} \quad (4)$$

The last fitness f_{range} encourages the expansion of the sampling range. For **Sampling** strategy with two normalized value bounds ($v_{\text{min}}, v_{\text{max}}$), the calculation is expressed as:

$$f_{\text{range}} = |v_{\text{max}} - v_{\text{min}}| \quad (5)$$

Using the combination loss given in Equation (3), the modified PSO algorithm iteratively evaluates and “exploits” values of the sampling set of a specific hydraulic parameter while holding the latest states of other parameters constant over an extended timeframe. In addition, as shown in Fig. 1, parameter permutation introduces uncertainty, allowing PSO to explore solutions within a dynamic landscape. This strategy enables us to retrieve optimal sampling sets of hydraulic parameters for all available networks. These sets are stored in corresponding networks’ configurations and, therefore, leveraged by a simulation to produce data points on a large scale.

Simulation. Subfigure (b) of Fig. 1 illustrates the simulation workflow. Overall, the entire workflow leverages multi-core processing powered by a high-performance computing cluster. From the previous stage, an optimal sampling set associated with its strategy was transferred to the Generator, where we sampled actual simulation parameters. Thanks to the prior optimization process, the sampling yields statistically sound parameter values aligned with plausible scenarios. These parameters were batched and passed through a Simulator, where EPANET was used to simulate outcome scenarios. Importantly, only successfully executed scenarios indicated by Error Code 0 from EPANET feedback were accepted. This ensures their hydraulic feasibility. Additionally, scenario validation was further assessed using expert-defined rules (e.g., in-range pressure and time consistency). Finally, the input and output parameters of the validated scenarios were encapsulated in a compressed file.

Automatic Demand Generator. The ADG algorithm aims to generate the junctions’ demand patterns for each node in a WDN. The demand pattern is defined following an additive model of three components: a daily pattern, a yearly seasonal pattern, and noise, as expressed in Equation (6).

$$D = \text{daily}(x_t) + \text{yearly}(x_t) + \varepsilon_t : t \in T \quad (6)$$

where D is the demand pattern of each node in the network, $\text{daily}(x_t)$ is the daily pattern, $\text{yearly}(x_t)$ is the yearly seasonal pattern, and ε_t is white noise. The demand patterns generated are a multiplier time-series, i.e., a factor that is multiplied by the *base demand* of each node specified in the configuration file of each WDN. The generated time-series are normalized in the range [0, 1].

Daily Pattern. The daily pattern defines the water consumption per day based on consumption profiles: household, commercial, extreme-demand, and zero-demand. The consumption profiles are determined by splitting the 24-hour of a day into four 6-hour segments. Thus, starting at midnight, these segments represent the water consumption from 00:00 to 06:00, 06:00 to 12:00, 12:00 to 18:00, and 18:00 to 00:00. Each segment is assigned either a low, medium or high consumption. The range for low, medium or high consumption is defined by lower and upper bounds determined at random. Thus, from N random numbers in the range [0.00, 1.00], we compute the quantiles Q1 and Q3. Then, the low consumption goes from [0.00, Q1), the medium consumption goes from [Q1, Q3), and high consumption is in the range [Q3, 1]. For example, the household profile is represented as (low, high, medium, low). It is assumed a low consumption between midnight and six in the morning, with a peak consumption in the morning when people are preparing for work. Then, after noon, the demand gradually decreases during the day because people are at work, and finally the demand is low again at the end of the day when people are going to bed. In a similar way, the commercial profile is defined as (high, high, high, medium). In this case, assuming a high consumption most of the time with a small decrease at the end of the day.

Using the consumption ranges described before, we generate random samples for each of the 6-hour segments. The number of *samples_per_hour* is determined based on the sampling frequency (*time_step*) defined in the configuration file. Those 6-hour segments are then concatenated to compose the 24-hour corresponding to one day. Then, these 24-hour samples are repeated to span the entire *duration* of the demand pattern. The daily demand pattern is generated using the periodic function described in Equation (7).

$$\text{daily}(x_t) = \cos(x_t) + \sin(x_t) + z_t : t \in T \quad (7)$$

where the $\cos(\cdot)$ and $\sin(\cdot)$ terms introduce the daily periodicity in the time-series, x_t represents the previously generated random sample at time t , and the z_t term represents white noise. The noise component guarantees that each repetition of the 24-hour pattern along the time-series is not a fidelity copy of the previous one. Finally, we use the Savitzky-Golay filter^{21,22} to smooth the generated time-series.

After the consumption profiles are defined, they have to be assigned to each node in the network. Hence, we need to determine which nodes belong to household profile and which ones to commercial. Domain knowledge indicates that commercial nodes are grouped in certain regions of the WDN. In order to resemble this characteristic, we propose to cluster the nodes into two main groups: household and commercial. The clusters are computed using the Louvain community detection algorithm, a heuristic approach that maximizes the modularity of the network²³. This algorithm works in two phases. In the first phase, each node i is isolated and belongs to a community C . Then, the modularity gain is computed after each node is moved to its neighbor communities. If there is no positive gain in modularity, the node remains in its original community. This phase is repeated until no individual move can improve the modularity. For directed graphs, the modularity gain is computed as follows^{23–25}:

$$\Delta Q = \frac{k_{i,\text{in}}}{m} - \gamma \frac{k_i^{\text{out}} \cdot \sum_{\text{tot}}^{\text{in}} + k_i^{\text{in}} \cdot \sum_{\text{tot}}^{\text{out}}}{m^2} \quad (8)$$

where m is the size of the network, γ is the resolution parameter which controls the size of the communities²⁶, $k_i^{\text{out}}, k_i^{\text{in}}$ are the outer and inner weighted degrees of node i , and $\Sigma_{\text{tot}}^{\text{in}}, \Sigma_{\text{tot}}^{\text{out}}$ are the sum of in-going and out-going links incident to nodes in community C .

In the second phase, the communities found in the previous step become nodes in the network, and the sum of the weights in the corresponding communities becomes the link weights in the new graph. Then the whole algorithm is applied again. The algorithm stops when no modularity gain is achieved or when the modularity is lower than a certain *threshold*.

At this stage, we have coherent communities within each WDN. Now, we need to define the number of nodes from those communities that will be assigned to either commercial or household profiles. According to the statistics provided by the association of water companies in the Netherlands, about 28% of the users belong to the commercial sector^{27–29}. We randomly choose the *percentage_commercial* from the range (0.25, 0.35). This allows to resemble commercial consumption profiles in other countries around the world. We set the number of nodes that will be assigned to the commercial consumption profile as $\text{num_nodes_commercial} = \text{floor}(\text{percentage_commercial} \times \text{total_number_of_nodes})$. After that, we iterate the communities found in the previous stage and sequentially assign the nodes in each community to the commercial consumption profile until we reach the *num_nodes_commercial*. Finally, the remaining nodes will be assigned to household profile at this stage. While household and commercial profiles are self-explanatory, extreme and zero-demand are a special type of consumption profiles.

The extreme-demand is a special case for some nodes with a very high water consumption. Thus, the extreme-demand profile is represented as (high, high, high, high). Usually, an extreme node represents a group of nodes, commonly external to the water network, but also connected to it. We set the *extreme_dem_rate* = 0.02, i.e., 2% of the scenarios will have nodes whose demand is always high. In addition, we limited the number of nodes per scenario that can have extreme demand values, specifically we set *max_extreme_dem_junctions* = 2. Domain knowledge can help to determine this parameter if the number of extreme nodes is known beforehand. The nodes to be assigned an extreme-demand profile are chosen at random and excluded from the nodes in the household or commercial profile. Then, for these nodes, the demand is randomly generated in the range [Q3, 1], as described before.

The zero-demand is another special case that represents nodes that do not consume water, but which are part of the network. Thus, these nodes always have a zero-demand value. These nodes are used for monitoring and control of the network operation, or they are modeled due to a planned expansion of the network. We set the *zero_dem_rate* = 0.05, i.e., 5% of the scenarios will have nodes whose demand is zero. Likewise, 5% of the total number of nodes in the WDN will be assigned the zero-demand profile. Alternatively, the *zero_dem_rate* can be set to the ratio between the number of nodes in the baseline network whose *base demand* is zero with respect to the total number of nodes, and accordingly, the number of nodes belonging to this profile. The zero-demand nodes are chosen at random and excluded from the remaining household or commercial profiles.

The presence and use of both, extreme-nodes and zero-demand nodes, at modeling WDNs are seen in the baselines and also confirmed by experts in the water management domain. Including these two profiles in the generated data enables to cover a wider range of pressures and demands compared to the baselines. Otherwise, if the baselines have those types of nodes but those are not included in our generation algorithm, there is a mismatch between baseline and our data. Our goal is to extend the range of the generated data but still cover and resemble the WDNs baselines.

Yearly Pattern. The yearly pattern defines the trend of water consumption in the entire year, considering a seasonal component with a peak consumption in summer. The default configuration assumes the European summer season starting in June with a 3-month span. In addition, to introduce variability in the data, beneficial for training deep learning models, we randomly move the summer period along the entire year for approximately 20% of simulated scenarios. This approach introduces the seasonal patterns in other regions across the globe. The yearly pattern is composed of a yearly component, a seasonal component, and noise, as described in Equation (9).

$$\text{yearly}(x_t) = y(x_t) + s(x_t) + z_t : t \in T \quad (9)$$

where $y(x_t)$ is the yearly component generated using a Fourier time-series as described by Equation (10), $s(x_t)$ is the seasonal pattern generated using a periodic cosine function as described by Equation (11), and z_t is white noise.

$$y(x_t) = A_0 + \sum_{n=1}^H \left(A_n \cos \left(2\pi \frac{nx_t}{\text{num_samples}} \right) + B_n \sin \left(2\pi \frac{nx_t}{\text{num_samples}} \right) \right) : t \in T \quad (10)$$

where the Fourier coefficients A_n and B_n determine the amplitude of the signal, and they are randomly sampled from a uniform distribution in the range [0, 1), the value of H represents the number of harmonics used for the time-series, and the periodicity of the signal is 24-hour for the short-term dataset and 1-year for the long-term. The periodicity is given by the number of samples parameter *num_samples*.

$$s(x_t) = C \left(\cos \left(2\pi \frac{x_t - s_{\text{peak}}}{\text{num_samples}} \right) \right) : t \in T \quad (11)$$

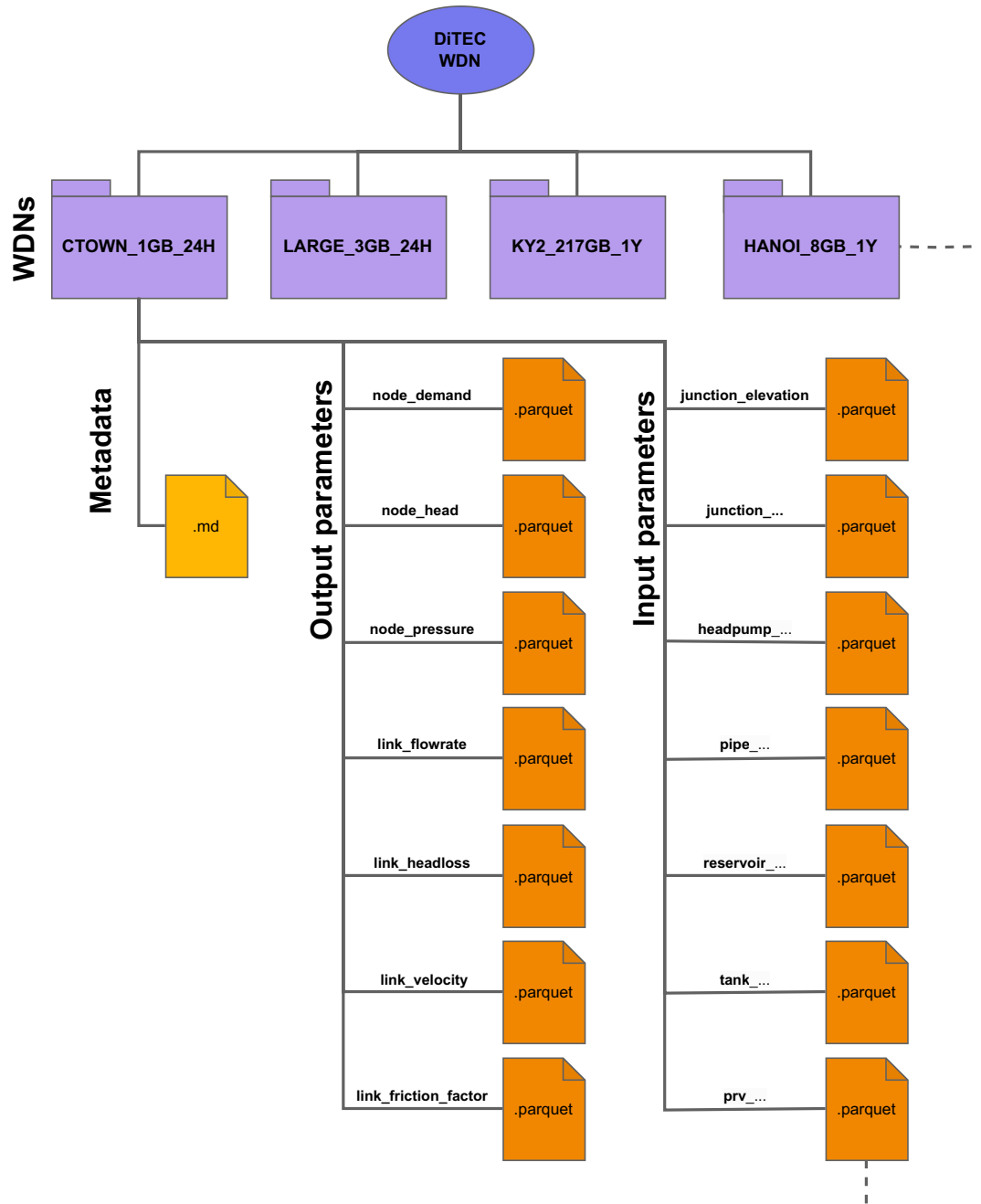


Fig. 2 The folder organization structure. The DiTEC-WDN collection includes 36 WDNs represented as folders. Every folder contains metadata and seven output parameters while the number of input parameters varies based on the available components per network. The dataset metadata is fed into a Markdown (.md) file structured as Dataset Card⁵⁶. In addition, parameter values are stored in one or more .parquet file(s), depending on their size. A .parquet file stores indices and node (link) values as distinct columns.

where C is a constant that represents the amplitude of the signal, reaching its maximum value in the summer peak s_{peak} , $num_samples$ defines the periodicity of the signal. Finally, the yearly time-series are normalized in the range $[0, 1]$.

Noise. The noise component ε_p , from Equation (6), is used to model the high and unexpected fluctuations in water consumption. Such variations can be caused by unpredictable changes in consumer behavior, network maintenance, transients or other unforeseen circumstances⁸. The noise component was sampled from a Gaussian normal distribution centered at zero and a standard deviation randomly sampled from a uniform distribution in the range $[min_noise, max_noise]$.

Metadata	Data Type	Description
adj_list	List	Adjacency list in which each element has a format of (source node name, dest node name, pipe name).
backup_times	Float	Backup time.
batch_size	Integer	Batch size defines how many samples the simulation takes per time.
duration	Integer	Simulation time in hours.
extreme_dem_rate	Float	Extreme demand rate indicates the rate appearing extreme demand in some nodes.
fcv_tune	Dict	FCV's optimized configuration.
fractional_cpu_usage	Float	Settings the CPU usage per worker as a part of the optimization process.
gen_batch_size	Integer	Batch size of random matrix generation, a part of the simulation process.
gpv_tune	Dict	GPV's optimized configuration.
head_pump_tune	Dict	Head pump's optimized configuration.
index_tracers	List	Selected scenario indices for recovering an interrupted simulation.
inp_paths	List	Paths to .INP file containing metadata of a WDN.
junction_tune	Dict	Junction's optimized configuration.
max_extreme_dem_junctions	Integer	The maximum allowed amount of extreme nodes in a scenario.
mem_per_worker	Float	Allocated memory in GB for each worker in simulation process.
mem_per_worker	Float	Allocated memory in GB for each worker in simulation process.
noise_range	Tuple	Lower and upper bounds of the addition noise in generating demand patterns.
num_cpus	Integer	Number of CPUs dedicated for the simulation process.
num_samples	Integer	Number of expected scenarios.
odims	Ordered Dict	parameter dimension associated with available components
okeys	Ordered Dict	parameter names associated with available components
onames	Ordered Dict	instance names associated with available components
output_path	String	path storing the simulation outcome.
p_commercial	Tuple	Lower and upper bounds of the demand of commercial nodes in generating demand patterns.
pbv_tune	Dict	PBV's optimized configuration.
pipe_tune	Dict	Pipe's optimized configuration.
power_pump_tune	Dict	Power pump's optimized configuration.
pressure_range	Tuple	Lower and upper bounds of a valid pressure.
profile_commercial	Tuple	Demand level of four quarters of the day in a commercial node.
profile_extreme	Tuple	Demand level of four quarters of the day in an extreme node.
profile_household	Tuple	Demand level of four quarters of the day in a household node.
prv_tune	Dict	PRV's optimized configuration.
psv_tune	Dict	PSV's optimized configuration.
ray_temp_path	String	Temporarily path for Ray.
reservoir_tune	Dict	Reservoir's optimized configuration.
save_success_inp	Boolean	Flag indicates whether saving a valid scenario in INP file for debugging only.
sim_outputs	List	Simulation outputs to be recorded
skip_names	List	Some abnormal nodes should be skipped in validation stage.
summer_amplitude_range	Tuple	Amplitude range of demand increase during summer period
summer_rolling_rate	Float	Probability of rolling summer period to mimic opposite seasons between two hemispheres.
summer_start	Float	normalized time remarking the beginning of summer.
tank_tune	Dict	Tank's optimized configuration.
tcv_tune	Dict	TCV's optimized configuration.
temp_path	String	Path to a folder storing temporary files.
time_consistency	Boolean	Flag indicates whether the input and output time-series must be equal in length.
time_step	Float	Simulated time sampling rate in hours.
verbose	Boolean	Flag indicates whether to print debug information during the simulation process.
yearly_pattern_num_harmonics	Integer	The number of terms in a Fourier series for a yearly pattern.
yield_worker_generator	Boolean	Flag indicates a generator to yield simulation outputs for saving memory.
zero_dem_rate	Float	Probability of appearing zero-demand nodes serving as water flow transitions and connections in the water system.

Table 2. List of metadata recorded in the **.md** file.

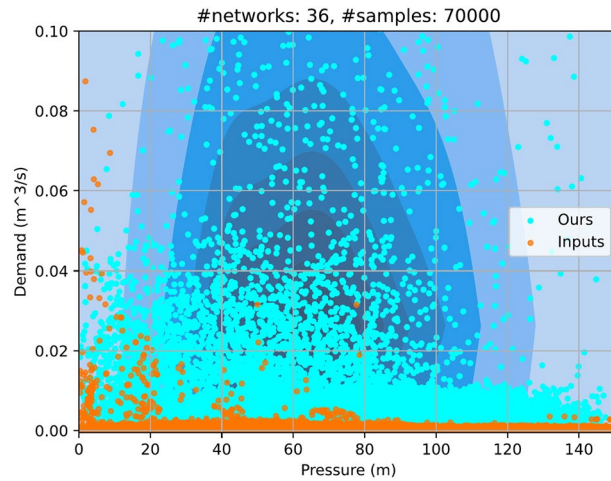


Fig. 3 Density distribution of *pressure* and *demand* across WDNs in DiTEC-WDN (cyan) and original ones from Input files (orange). The contours denote the data point density of the DiTEC-WDN dataset³⁰, with darker blue indicating higher concentration at the center and lighter blue showing lower density when going outward. In baseline networks, data points whose pressure is outside the range of (0, 151] in meters, are excluded due to the impractical operation conditions¹⁰.

Data Records

The DiTEC-WDN dataset³⁰ is available at <https://doi.org/10.57967/hf/6341> under CC BY 4.0 license. This dataset comprises 36,000 synthesized scenarios devised from publicly available WDNs that served as structural backbones. Specifically, the network's topology, node names, and link names remain unchanged, while other parameter values are machine-generated automatically. The repository where the raw dataset is located supports several data interface options to read and process *.parquet* files, allowing practitioners to select a concrete parameter or a subset of networks. Before use, the downloaded dataset requires an additional preprocessing step. Specifically, we removed columns corresponding to nodes along with their adjacent links, listed in *skip_names* in the metadata. Network metadata is accessible in any *.parquet* file in the corresponding folder. Additionally, to analyze graph topology, the metadata contains *adj_list* formatted as a list of tuples (source node, adjacent link, destination node).

The DiTEC-WDN dataset³⁰ comprises 36 WDNs, which includes ky1³¹, ky2³¹, ky3³¹, ky4³¹, ky5³¹, ky6³¹, ky7³¹, ky8³¹, ky10³¹, ky13³¹, ky14³¹, ky16³¹, ky18³¹, ky24_v³¹, 19 Pipe System³², Anytown³³, new_york³⁴, Jilin³⁵, hanoi³⁶, fossolo³⁷ (foss_poly_1), FOWM³⁸, EPANET Net 3³⁹, FFCL-1⁴⁰, Zhi Jiang⁴¹ (ZJ), WA1⁴², OBCL-1⁴², modena³⁷, NPCL-1⁴³, Marchi Rural⁴⁴ (RuralNetwork), CTOWN⁴, d-town⁴⁵, balerma⁵, L-TOWN⁴⁶, KL⁴⁷, Exnet⁴⁸ (EXN), and Large⁴⁹.

Each WDN contains 1,000 distinct scenarios. A scenario is a sequence of snapshots, capturing key measurements sampled hourly from all components. Each snapshot describes a directed graph in which nodes involve *reservoir*, *tank*, and *junction*, and links represent *pipe*, *head pump*, *power pump*, *PRV*, *PSV*, *FCV*, and *TCV* valves. Note that some valve types are omitted, as they are unavailable in the dataset. In particular, we recorded input parameters of all components (as described in Table 1) and seven simulation outputs: *pressure*, *demand*, *head*, *flow rate*, *velocity*, *head loss*, and *friction factor* with units defined per network and in standard units.

Each WDN is located in a folder named as `<network>_<capacity>_<duration>`. The `<network>` name corresponds to the baseline network, the `<capacity>` indicates the physical size (varying from 1 GB to 232 GB), and the `<duration>` specifies the simulation period which can be 24 hours (24H) or 1 year (1Y). As shown in Fig. 2, each directory physically contains a metadata Markdown (*.md*) file, seven output parameters, and several input parameters stored in *.parquet* files. The metadata includes network topology, node, edge names, and auxiliary information served for optimization, generation, and simulation phases as detailed in Table 2. For *.parquet* files, their naming follows the syntax `<component>_<parameter>_<index>_<type>_<io>`. The `<component>` and `<parameter>` define which component the parameter belongs to. The `<index>` represents the shard index of the *.parquet* file. The `<type>` specifies the parameter category—*curve*, *static*, or *dynamic*—while `<io>` indicates whether the parameter is *input* or *output*.

Each parameter is associated with a table whose values are arranged based on the parameter type as follows:

- *Static* tables have dimensions ($num_scenarios \times num_components$).
- *Pattern* tables have dimensions ($(num_scenarios * num_snapshots) \times num_components$).
- *Curve*-related tables have dimensions ($(num_scenarios * num_curve_points) \times num_components$).

where $num_scenarios$ stands for the number of scenarios, $num_snapshots$ represents the number of snapshots, num_curve_points refers to the number of curve points, and $num_components$ indicates the number of nodes or links.

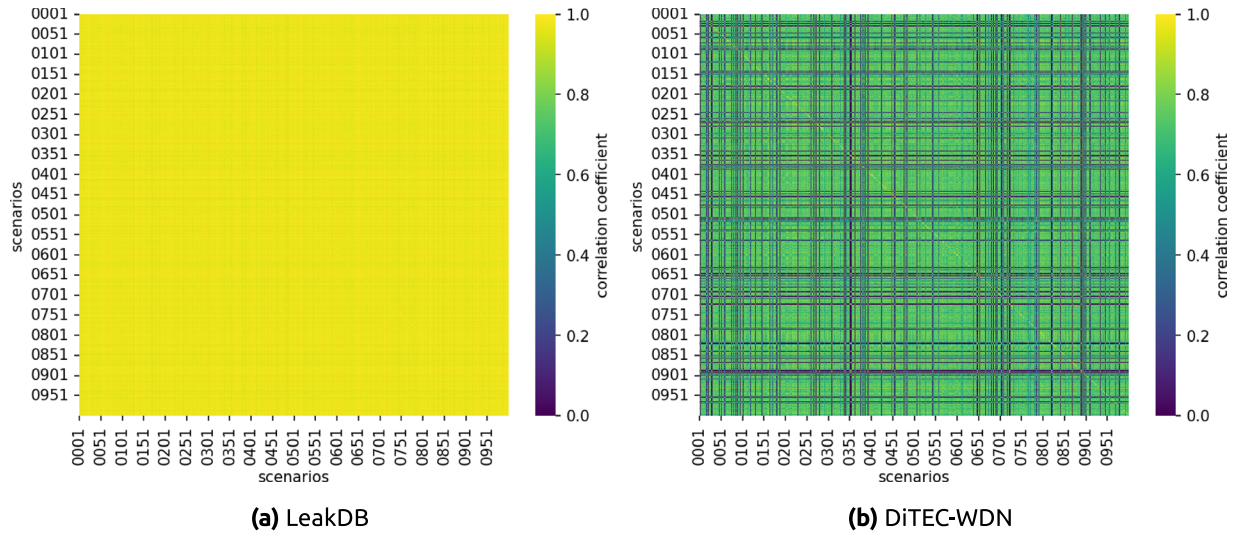


Fig. 4 Correlation matrices of generated demands between all scenarios in Hanoi WDN. The left figure (a) shows the correlation between scenarios in the data generated in LeakDB⁸. The right figure (b) shows the correlation between the scenarios in our dataset. Both matrices include all 1,000 scenarios, each containing 1-year of demand data. The low correlation between scenarios in our dataset shows the diversity of the data, contrary to the similarity observed across LeakDB scenarios.

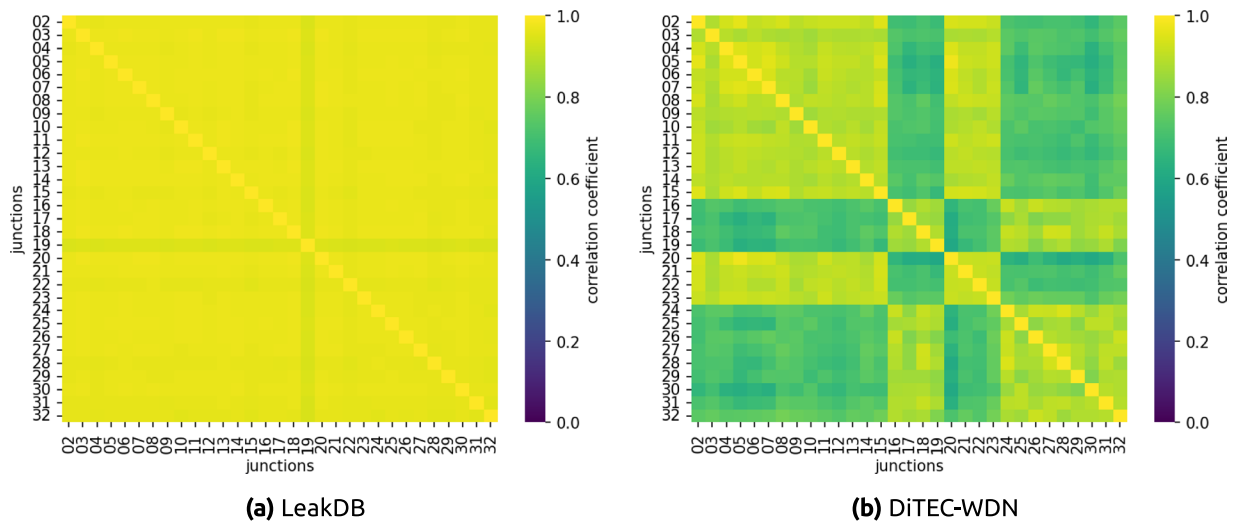


Fig. 5 Correlation matrices of generated demands between junction nodes in a randomly chosen scenario from Hanoi WDN. The left figure (a) shows the correlation between junction demands in the data generated in LeakDB⁸. The right figure (b) shows the correlation between the junction demands in our dataset. The high correlation in LeakDB shows the overuse of demand patterns for several nodes, contrary to what is observed in our dataset. The blocks in the correlation matrix of our dataset highlight the difference between household and commercial demand patterns.

Technical Validation

To assess the dataset quality, we compared DiTEC-WDN against (1) baseline networks and (2) LeakDB⁸, a well-known dataset. We visualized data distribution in the former and examined *demand patterns* in the latter, highlighting their scarcity and the risks of overuse in the existing dataset.

Comparative Analysis. *DiTEC-WDN vs. Baseline networks.* Figure 3 highlights the contrast in data distribution between baseline networks (orange) and DiTEC-WDN dataset³⁰ (cyan) along the *demand* and *pressure* axes. On the left, baseline data points correspond to high demand and low pressure, indicating that only a few nodes receive sufficient supply while most experience pressure drops. Similarly, on the right side, the pressure of baseline points is stable only when their corresponding demand approaches near zero. This reflects the demand scarcity and suboptimal simulation parameters. An alternative approach is leveraging these networks to build a synthetic dataset, where parameters are drawn from a random distribution^{9,50,51}. However, this could violate

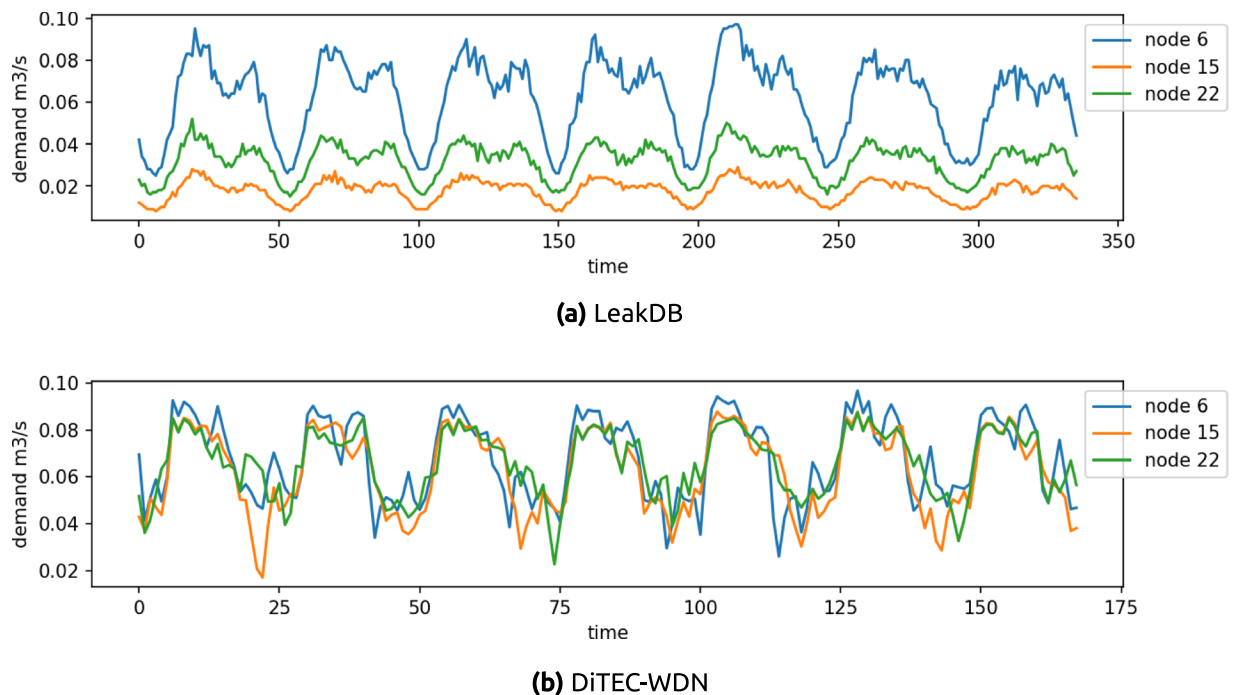


Fig. 6 Time-series of the generated demands of three randomly chosen junction nodes from Hanoi WDN. Figure (a) shows one week of the demands generated in LeakDB⁸, sampled every 30 minutes. The reuse of a single pattern for different nodes is clearly observed in LeakDB. Figure (b) shows one week of demands from our dataset, sampled every 60 minutes. In this case, the fluctuations observed in the time-series show a different consumption pattern per node.

realism and consistency. For instance, arbitrarily sampling nodal elevation or pipe diameter may result in unrealistic scenarios, such as spiky terrain or pipe bottlenecks in the WDN.

In contrast, we specifically designed the parameter spaces and enforced strict rule validation to ensure hydraulic stability across scenarios while expanding into a larger space. As a result, the DiTEC-WDN dataset³⁰ provides a broader, more realistic receptive field within the typically operational pressure range. This enables the robustness of training data-driven machine-learning models. Moreover, DiTEC-WDN's variability allows water researchers to analyze diverse scenarios without repetitive simulations, thereby preventing inconsistent results among studies and ensuring more sustainable research practices.

DiTEC-WDN vs. LeakDB. Another important analysis is how our generated data differ from the commonly used existing benchmark dataset, LeakDB⁸. Figure 4 shows the demand correlation matrices between the 1,000 scenarios generated in LeakDB and our generated data. As can be seen in Fig. 4(a), the scenarios generated in LeakDB are highly correlated. The correlation matrix shows only slight variations between some scenarios, which implies data redundancy. This limits the capacity of deep learning models to learn from such data. On the contrary, in our dataset, the correlation between scenarios is much lower, as can be seen in Fig. 4(b). This confirms the diversity of the generated data, allowing the models to see a larger space of solutions during the training process.

Similar conclusions can be drawn from the correlation matrices between the junction demands in an arbitrary scenario (see Fig. 5). The high correlation shown in Fig. 5(a) exposes the negative effect of demand patterns overuse in existing approaches. In contrast, Fig. 5(b) shows a moderate correlation between junction demands in our data, implying there is a pattern in consumption demand, but this is not identical for every node in the WDN. In addition, the block patterns shown in Fig. 5(b) display the difference between households and commercial consumption profiles described in Section Automatic Demand Generator.

Finally, the time-series of one week demand for three nodes from a random scenario in LeakDB and our dataset are shown in Fig. 6. The time-series depicted in Fig. 6(a) show how a single demand pattern is reused for the three nodes in LeakDB. While the noise shows some subtle variations, each time-series looks like a translated and scaled version of the other. Contrary, our data exhibit consumption patterns, but the fluctuations in each time-series are independent, as shown in Fig. 6(b).

Usage Notes

Limitations. Despite the diversity of simulation parameters recorded in the dataset, there are three limitations: (1) incorporating realism into our dataset, (2) replicating unexpected situations, and (3) storing the auto-generated hyper-parameters.

First, regarding realism, traditional hydraulic models (e.g., INP files) were typically assessed through calibration, which involves aligning the model output with measurements recorded by sensors⁵². However, due to the synthetic generation of all parameters and the unavailability of sensors, such calibration was infeasible. Instead, we ensured that inputs, including parameters and operational conditions, were within the statistical range observed in the original models. We also verified the generated scenarios through rule-based checks using constraints defined by water domain experts, along with confirmation of an error-free status from the EPANET simulation feedback.

We assume all scenarios are under normal conditions and components are functioning correctly. Accordingly, anomalies, such as negative pressure, leakage, fire-fighting, or pipe break, are excluded from the dataset. Nonetheless, the dataset's normal conditions still provide a useful foundation for training models in anomaly-related applications (e.g., leak detection⁵³).

For the last limitation, some hyper-parameters generated during the simulation process, such as the locations of extreme-demand and zero-demand nodes, and nodal demand profiles, cannot be recorded. To address this, these extreme-demand and zero-demand nodes can be identified using high-pass and low-pass filters, respectively, while demand profilers can be classified by an unsupervised machine-learning algorithm, such as K-Means⁵⁴ or DBSCAN⁵⁵.

Data availability

The DiTEC-WDN dataset³⁰ in *.parquet* format is freely available under a CC BY 4.0 license at <https://doi.org/10.57967/hf/6341>. Users can download complete networks or selected parameter subsets and access the *.parquet* files via *pyarrow*, *Hugging Face datasets*, or our custom data interface at [GitHub repository](#).

Code availability

The optimization algorithm and generation tool are available on Zenodo (<https://doi.org/10.5281/zenodo.15649072>) under the MIT License. The repository includes a detailed tutorial and wiki to guide scenario generation for a customized WDN. The outcome dataset is stored in *Zarr*, an efficient compressed format. Conversion to *.parquet* can be performed using the *zarr2parquet.py* script.

Received: 19 March 2025; Accepted: 23 September 2025;

Published online: 03 November 2025

References

- Brumbelow, K., Torres, J., Guikema, S., Bristow, E. & Kanta, L. Virtual cities for water distribution and infrastructure system research. In *World environmental and water resources congress 2007: Restoring our natural habitat*, 1–7, [https://doi.org/10.1061/40927\(243\)469](https://doi.org/10.1061/40927(243)469) (2007).
- Sitzenfrei, R., Möderl, M. & Rauch, W. Automatic generation of water distribution systems based on gis data. *Environmental modelling & software* **47**, 138–147 (2013).
- Giustolisi, O. & Walski, T. M. Demand components in water distribution network analysis. *Journal of Water Resources Planning and Management* **138**, 356–367 (2012).
- Ostfeld, A. *et al.* Battle of the water calibration networks. *Journal of water resources planning and management* **138**, 523–532 (2012).
- Reca, J. & Martínez, J. Genetic algorithms for the design of looped irrigation water distribution networks. *Water resources research* **42** (2006).
- Zanfei, A., Menapace, A., Brentan, B. M., Sitzenfrei, R. & Herrera, M. Shall we always use hydraulic models? a graph neural network metamodel for water system calibration and uncertainty assessment. *Water Research* **242**, 120264 (2023).
- Cassiolato, G. H., Ruiz-Femenia, J. R., Salcedo-Diaz, R. & Ravagnani, M. A. Water distribution networks optimization considering uncertainties in the demand nodes. *Water Resources Management* **38**, 1479–1495 (2024).
- Vrachimis, S. G., Kyriakou, M. S. *et al.* Leakdb: a benchmark dataset for leakage diagnosis in water distribution networks:(146). In *WDSA/CCWI joint conference proceedings*, vol. 1 (2018).
- Tello, A., Truong, H., Lazovik, A. & Degeler, V. Large-scale multipurpose benchmark datasets for assessing data-driven deep learning approaches for water distribution networks. *Engineering Proceedings* **69**, <https://doi.org/10.3390/engproc2024069050> (2024).
- Truong, H., Tello, A., Lazovik, A. & Degeler, V. Graph neural networks for pressure estimation in water distribution systems. *Water Resources Research* **60**, e2023WR036741, <https://doi.org/10.1029/2023WR036741> (2024).
- Rossman, L. A. The epanet programmer's toolkit for analysis of water distribution systems. In *WRPMD'99: Preparing for the 21st Century*, 1–10 (1999).
- Klise, K. A., Murray, R. & Haxton, T. An overview of the water network tool for resilience (wntr). In *Proceedings of the 1st International WDSA/CCWI Joint Conference, Kingston, Ontario, Canada* (Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2018).
- Rossman, L. A. *et al.* Epanet 2 users manual (2000).
- Donninger, C. Null move and deep search. *ICGA Journal* **16**, 137–143, <https://doi.org/10.3233/ICG-1993-16304> (1993).
- Fournier, A., Fussell, D. & Carpenter, L. Computer rendering of stochastic models. *Commun. ACM* **25**, 371–384, <https://doi.org/10.1145/358523.358553> (1982).
- Kennedy, J. & Eberhart, R. Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks*, 4, 1942–1948, <https://doi.org/10.1109/ICNN.1995.488968> (1995).
- Suribabu, C. R. & and, T. R. N. Design of water distribution networks using particle swarm optimization. *Urban Water Journal* **3**, 111–120, <https://doi.org/10.1080/15730620600855928> (2006).
- Moghaddam, A., Mokhtari, M., Afsharnia, M. & Minaee, R. P. Simultaneous hydraulic and quality model calibration of a real-world water distribution network. *Journal of Water Resources Planning and Management* **146**, 06020007, [https://doi.org/10.1061/\(ASCE\)WR.1943-5452.0001209](https://doi.org/10.1061/(ASCE)WR.1943-5452.0001209) (2020).
- Meirelles, G., Manzi, D., Brentan, B., Goulart, T. & Luvizotto, E. Calibration model for water distribution network using pressures estimated by artificial neural networks. *Water Resources Management* **31**, 4339–4351 (2017).
- Vinutha, H., Poornima, B. & Sagar, B. Detection of outliers using interquartile range technique from intrusion dataset. In *Information and decision sciences: Proceedings of the 6th international conference on ficta*, 511–518 (Springer, 2018).
- Savitzky, A. & Golay, M. J. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry* **36**, 1627–1639 (1964).

22. Luo, J., Ying, K. & Bai, J. Savitzky–golay smoothing and differentiation filter for even number data. *Signal processing* **85**, 1429–1434 (2005).
23. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* **2008**, P10008 (2008).
24. Traag, V., Waltman, L. & Van Eck, N. From louvain to leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
25. Dugué, N. & Perez, A. *Directed Louvain: maximizing modularity in directed networks*. Ph.D. thesis, Université d'Orléans (2015).
26. Newman, M. E. Equivalence between modularity optimization and maximum likelihood methods for community detection. *Physical Review E* **94**, 052315 (2016).
27. Vewin. Dutch drinking water statistics 2015. (2015).
28. Vewin. Dutch Drinking Water Statistics 2017 - From source to tap. <https://www.vewin.nl/wp-content/uploads/2024/08/Drinkwaterstatistieken-2017-EN.pdf> Online; accessed 05-March-2025 (2017).
29. Vewin. Dutch Drinking Water Statistics 2022 - From source to tap. <https://www.vewin.nl/wp-content/uploads/2024/06/vewin-dutch-drinking-water-statistics-2022-eng-web.pdf> Online; accessed 05-March-2025 (2022).
30. Truong, H., Tello, A., Lazovik, A. & Degeler, V. ditec-wdn, <https://doi.org/10.57967/hf/6341> (2025).
31. Jolly, M. D., Lothes, A. D., Sebastian Bryson, L. & Ormsbee, L. Research database of water distribution system models. *Journal of Water Resources Planning and Management* **140**, 410–416 (2014).
32. Wood, D. J. & Charles, C. O. Hydraulic network analysis using linear theory. *Journal of the Hydraulics division* **98**, 1157–1170 (1972).
33. Walski, T. M. *et al.* Battle of the network models: Epilogue. *Journal of water resources planning and management* **113**, 191–203 (1987).
34. Schaake, J. C. Jr & Lai, D. Linear programming and dynamic programming application to water distribution network design (1969).
35. Bi, W. & Dandy, G. C. Optimization of water distribution systems using online retrained metamodels. *Journal of Water Resources Planning and Management* **140**, 04014032 (2014).
36. Fujiwara, O. & Khang, D. B. A two-phase decomposition method for optimal design of looped water distribution networks. *Water resources research* **26**, 539–549 (1990).
37. Bragalli, C., Ambrosio, C., Lee, J., Lodi, A. & Toth, P. Ibm research report water network design by minlp water network design by minlp (2008).
38. Walski, T. M. 05 federally owned water main (1984).
39. Clark, R. M., Rossman, L. A. & Wymer, L. J. Modeling distribution system water quality: Regulatory implications. *Journal of water resources planning and management* **121**, 423–428 (1995).
40. Rossman, L. A. & Boulos, P. F. Numerical methods for modeling water quality in distribution systems: A comparison. *Journal of Water Resources planning and management* **122**, 137–146 (1996).
41. Zheng, F., Simpson, A. R. & Zecchin, A. C. A combined nlp-differential evolution algorithm approach for the optimization of looped water distribution systems. *Water Resources Research* **47** (2011).
42. Vasconcelos, J. J., Rossman, L. A., Grayman, W. M., Boulos, P. F. & Clark, R. M. Kinetics of chlorine decay. *Journal-American Water Works Association* **89**, 54–65 (1997).
43. Clark, R. M. Applying water quality models. In *Computer modeling of free-surface and pressurized flows*, 581–612 (Springer, 1994).
44. Marchi, A., Dandy, G., Wilkins, A. & Rohrlach, H. Methodology for comparing evolutionary algorithms for optimization of water distribution systems. *Journal of Water Resources Planning and Management* **140**, 22–31 (2014).
45. Marchi, A. *et al.* Battle of the water networks ii. *Journal of water resources planning and management* **140**, 04014009 (2014).
46. Vrachimis, S. G. *et al.* Dataset of battledim: Battle of the leakage detection and isolation methods. In *Proc., 2nd Int CCWI/WDSA Joint Conf. Kingston, ON, Canada: Queen's Univ* (2020).
47. Kang, D. & Lansey, K. Revisiting optimal water-distribution system design: Issues and a heuristic hierarchical approach. *Journal of Water resources planning and management* **138**, 208–217 (2012).
48. Farmani, R., Savic, D. A. & Walters, G. A. Exnet benchmark problem for multi-objective optimization of large water systems. *Modelling and control for participatory planning and managing water systems* (2004).
49. Sitzenfrey, R., Hajibabaei, M., Hesarkazzazi, S. & Diaio, K. Dual graph characteristics of water distribution networks—how optimal are design solutions? *Complex & Intelligent Systems* **9**, 147–160 (2023).
50. Ashraf, I., Strotherm, J., Hermes, L. & Hammer, B. Physics-informed graph neural networks for water distribution systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 21905–21913, <https://doi.org/10.1609/aaai.v38i20.30192> (2024).
51. Kerimov, B., Taormina, R. & Tscheikner-Gratl, F. Towards transferable metamodels for water distribution systems with edge-based graph neural networks. *Water Research* **261**, 121933, <https://doi.org/10.1016/j.watres.2024.121933> (2024).
52. Trifunović, N. *Introduction to Urban Water Distribution, Second Edition: Theory*. IHE Delft lecture note series (Taylor & Francis Group, 2020).
53. Örn Gardarsson, G., Boem, F. & Toni, L. Graph-based learning for leak detection and localisation in water distribution networks*. *IFAC-PapersOnLine* **55**, 661–666, <https://doi.org/10.1016/j.ifacol.2022.07.203> (2022).
54. Lloyd, S. Least squares quantization in pcm. *IEEE Transactions on Information Theory* **28**, 129–137, <https://doi.org/10.1109/TIT.1982.1056489> (1982).
55. Ester, M., Kriegel, H.-P., Sander, J., Xu, X. *et al.* A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, vol. 96, 226–231 (1996).
56. Mitchell, M. *et al.* Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19*, 220–229, <https://doi.org/10.1145/3287560.3287596> (ACM, 2019).

Acknowledgements

This work is funded by the project DiTEC: Digital Twin for Evolutionary Changes in Water Networks (NWO 19454). We thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Hábrók high performance computing cluster. Also, we appreciate Hugging Face for hosting the dataset repository.

Author contributions

All authors conceptualized the idea. H.T., A.T. and V.D. involved in writing and reviewing this manuscript. H.T. contributed to the methodology, developed the optimization and generation tool, investigation, visualization, and data curation. A.T. contributed in methodology, validation, visualization, and developed the demand generation. A.L. and V.D. provided resources, supervised the project, and contributed to project administration and funding acquisition.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-025-06026-0>.

Correspondence and requests for materials should be addressed to H.T. or A.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025