water

MDPI

Article

Data-Informed Synthetic Networks of Water Distribution Systems for Resilience Analysis in Puerto Rico

Kirk L. Bonney [1,*], Katherine A. Klise [1], Jason W. Poff [2], Samuel Rivera [2], Ian Searles [3] and Mikhail Chester [3]

[1] Energy Water Systems Integration Department, Sandia National Laboratories, Albuquerque, NM 87123, USA
[2] Department of Civil and Construction Engineering, Oregon State University, Corvallis, OR 97331, USA
[3] School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe, AZ 85281, USA
* Correspondence: klbonne@sandia.gov

Abstract: The increasing potential of infrastructure disruptions calls for high-quality infrastructure models to be used in resilience analysis and decision making. Unfortunately, many utilities and communities do not have access to accurate and detailed models due to a lack of data and resources. Furthermore, security restrictions on sharing infrastructure models present roadblocks to research, analysis, and decision making. Recent advances in the development of synthetic water distribution models provide a potential solution to this problem. There is an opportunity to improve these methods by leveraging incomplete pipe datasets to aid synthetic network generation. To address this gap, we developed a methodology for synthetic network generation that incorporates partial pipe data using a modification of the minimum cost flow algorithm for network generation and pipe sizing. This methodology demonstrates how partial pipe data can be leveraged to improve site-specific synthetic network generation. For the study area of Mayagüez, Puerto Rico, a synthetic model generated using 50% of real pipe data matches the pressure of the validation system with an average error of 23.5 m of head, which improves upon the average error of 31.6 m of head produced by a synthetic model generated using no data of the real pipes. Additionally, synthetic networks are shown to replicate the pressure response under a disruption scenario of the validation network, suggesting potential use in resilience analysis.

Keywords: water distribution systems; synthetic infrastructure; resilience analysis

Citation: Bonney, K.L.; Klise, K.A.; Poff, J.W.; Rivera, S.; Searles, I.; Chester, M. Data-Informed Synthetic Networks of Water Distribution Systems for Resilience Analysis in Puerto Rico. *Water* **2024**, *16*, 3356. https://doi.org/10.3390/w16233356

Academic Editor: Fernando António Leal Pacheco

Received: 30 September 2024
Revised: 13 November 2024
Accepted: 20 November 2024
Published: 22 November 2024

1. Introduction

Water distribution systems are a crucial aspect of urban infrastructure, delivering water to consumers through a complex network of water treatment plants, pipes, pumps, tanks, and valves. Numerical models of these systems are used in a wide range of applications, including network design [1,2], real-time control [3,4], resilience analysis [5–7], and risk assessment [8,9]. The development of accurate water distribution system models remains a challenge due to sparse and inaccurate utility asset data [10,11] and the lack of standard methods to build models from diverse data. Furthermore, while many large water utilities have invested in up-to-date and highly detailed water distribution system models, many small- to medium-sized water utilities do not build and maintain accurate models. Access to infrastructure models is further restricted due to data sharing limitations.

Advances in synthetic infrastructure model development provides a path toward developing accurate, site-specific infrastructure models from limited data. Applications include transmission and distribution power grids [12,13], drinking water distribution systems [10,14,15], wastewater distribution systems [16], and interdependent infrastructure [11]. Roadways are commonly used as a proxy network structure for synthetic infrastructure models [10,17,18]. When using road networks as a proxy for water systems, several key considerations need to be taken into account. For example, water utility right-of-ways

do not always follow roads, road intersections define connectivity that may not be present in water system, and water systems can contain multiple pipes that run along a single road. These important differences between road and water systems could have significant impact on the hydraulics of a synthetic model.

Given the challenges of building water distribution system models from imperfect data, the use of synthetic network generation to create pipe networks for water distribution models is an active area of research. DynaViBe used structural urban data and graph theory to generate synthetic case studies [14]. The software was extended to use additional geographic information system (GIS) data such as population and housing density [19]. Möderl et al. [20] built an ensemble of conceptual synthetic networks intended to represent a wide range of characteristics that could be found in real-life water distribution systems. Using water and sewer pipes from an alpine city, Mair et al. [21] used network correlation/similarity analysis to determine that 50% percent of the road network length correlated with 80–85% of the water and sewer pipes; however, a correlation between street type and pipe diameter was not found. Using the same alpine city, Mair et al. [22] illustrated that synthetic water distribution system models achieve high accuracy in pressure using only 30% of large-diameter pipes. Paez and Filion [23] created water distribution system models of five real cities (one in Europe, two in North America, and one in Asia). The real and synthetic networks were compared using a wide range of global topographic and reliability related metrics, including link density, average node degree, meshedness coefficient, the density of articulation points and bridges, pressure, and a modified resilience index. Ahmad et al. [10] created synthetic models of Tempe and the greater Phoenix area. Methods were validated using a benchmark model of North Marin, California (EPANET Net3), to compare the distribution of pipe diameters. Momeni et al. [24] used a benchmark model (Anytown) to generate synthetic networks and apply a multiobjective genetic algorithm to optimize network attributes and correct unusual network structures that can arise with synthetic methods. Approaches to synthetic network generation using linear programming were taken by Rehm et al. [25] in Cologne, Germany, and Zaucher et al. [18] in Piedmont, California.

The literature for synthetic network generation has focused on the design and planning of water distribution; however, the use of synthetic networks to build proxy models for decision making during disruption scenarios (i.e., resilience analysis) is an underdeveloped area of research. While Nikolopoulos et al. [26] quantified uncertainty in resilience by applying a stress testing framework to a synthetic network, the focus was on the demonstration of the framework rather than an evaluation of the suitability of synthetic networks as proxies for real systems. There are two research gaps that this paper seeks to address. The first gap is the underutilization of sparse datasets in synthetic network generation. In situations where only incomplete, low-quality pipe datasets are available, this would provide a way to build plausible water distribution system models despite data limitations. Leveraging an incomplete pipe dataset in conjunction with a synthetic network generation method could produce more realistic models. One notable exception is the work of Mair et al. [22], who explored the combination of a road network with subsets of real pipe data for synthetic network generation. However, the methodology was agnostic to whether or not network edges were associated with real pipe data. The second gap is the need for the additional verification of synthetic models for decision making during disruptive scenarios (e.g., pipe breaks, water treatment plant outage). Previous work typically presented high-level statistics on pressure and pipe diameters [19,20] and global network statistics [23]. However, in a water distribution system, even if the pipe diameter distribution, average pressure, and general network structure are similar, certain heterogeneous aspects of the real system (e.g., pressure distribution and response to disruption) may not necessarily be replicated. An example of synthetic model evaluation focused on spatial distributions of pressure can be found in Mair et al.'s study [22]; they used performance indices to compare nodal pressures of synthetic models to those of a real model. Evaluating the reproduction of model pressure is particularly important when using a synthetic model for resilience analysis, where it is necessary to accurately capture the spatial relationship between damage and

impact. Furthermore, it is valuable to understand how property-based (e.g., diameters, topology) measures of model similarity correlate with measures of model similarity based on hydraulic simulation and response to disruptive scenarios [27].

This work focuses on the use of synthetic network generation methods to build water distribution system models in Puerto Rico, where up-to-date models are in short supply and water systems are subject to a wide range of disruptive events. The city of Mayagüez was chosen as the case study, where a model was built from publicly available infrastructure data for Puerto Rico as part of this work [28]. This infrastructure dataset has also been used in other work for both water and power system analyses [29–31]. The Puerto Rico database includes infrastructure data for drinking water, wastewater, and power across the island. While these data are incomplete and include inaccuracies, the data quality in Mayagüez is high enough to build a water distribution system model for the purpose of validation (referred to as the "validation" model in this paper). In many of the other cities and rural areas of Puerto Rico, available water infrastructure data are sparse. Pipe data can be fragmented, and entire areas can be missing data entirely. Even when a pipe location is known, the diameter is not always recorded in the dataset. Given the incomplete nature of the data in these regions, it is difficult to create models of water distribution systems exclusively using data. While fully synthetic models could be used in place of models built from data, they may fail to capture important features of the system. Taking a combined approach to model development using both synthetic methods and incomplete pipe data to generate the pipe network could result in more accurate models.

The goal of this work was to use incomplete data to generate models that more closely represent the real system during normal operations and disruptive events. We accomplish this by pursing the following objectives: (1) develop a synthetic network generation technique that can incorporate incomplete pipe datasets for the development of more realistic models and (2) provide a detailed comparison between synthetic models generated with varied amounts of pipe data and the validation model focused on topological metrics, pipe diameters, spatial differences in pressure, and pressure response to a disruptive scenario. This work is organized as follows: Section 2 includes descriptions of the data and methods for the development of the validation model, the generation of synthetic models, and the suite of metrics and analyses used to compare the synthetic and validation models. Section 3 presents the results of applying the metrics and analyses to the generated synthetic models and the validation model. Section 4 includes a discussion of the results, the limitations of the experiment, and ideas for future work. Section 5 is a brief summary of this paper and key takeaways.

2. Methods

The following section includes details on the development of a validation model for the Mayagüez water service area, the methodology used to generate data-informed synthetic models, and the suite of metrics that were used to compare the validation model to the synthetic models. The analysis was carried out using several open source Python packages, including the Water Network Tool for Resilience (WNTR) [5] to build and run hydraulic simulations, GeoPandas [32] to analyze geospatial data, and OSMNX [33] to obtain the road network from OpenStreetMaps [34].

Hydraulic analyses were ran using the EPANET simulator through the Python interface provided by WNTR. Steady-state simulations were used to compare pressure across the validation and synthetic models as described in Section 2.3.3. Extended-period steady-state simulations were used to simulate pressure for the disruption scenario described in Section 3.4. In both situations, pressure-dependent demand was used to ensure that demand at each junction is a function of pressure. Minimum pressure was set to 0 m, and required pressure was set to 20 m. Pressure in meters assumes a fluid density of 1000 kg/m^3 to convert pressure in Pa to meters of water pressure.

2.1. Study Area

The water distribution system in Mayagüez, Puerto Rico, was the central object of the analysis. The focus on Puerto Rico was motivated by the lack of water infrastructure models across the island, the complexity of the water systems, and the need for resilience analyses that can consider vulnerabilities like hurricanes, power outages, and component failure from aging infrastructure. In addition to advancing the available techniques for synthetic water infrastructure, focusing on a case study in Puerto Rico supports the modeling needs of the island.

Validation Model Development

The government of Puerto Rico hosts a publicly available database that includes infrastructure data that can be used to build models. For much of the island, the completeness and quality of these datasets are insufficient for the development of detailed water distribution system models. However, the data for Mayagüez are of high enough quality for detailed model development. It is important to keep in mind that the goal for this work was not to create a fully calibrated water utility model from the data. Rather, it was to create a model that includes site-specific heterogeneous details of the real system based on available data and that can be used to quantify how well synthetic network generation methods capture model behavior. The data and assumptions that were used to build the model are described in Appendix A and summarized in the following paragraphs.

The Puerto Rico infrastructure database is available at https://gis.pr.gov/ (accessed on 1 October 2022) [28]. The data include water, road, and power infrastructure assets across the island and have been used in several publications [29–31]. Data on water assets include the location, geometry, and attributes of water treatment plants, pipes, pumps, tanks, and valves stored in GIS-compatible files. While the data do not include operational data (e.g., valve and pump controls), the raw data can be used to build fairly complete models of select water service areas. Simple data quality control analysis was performed to identify regions that have adequate data for model development. This included the identification of missing assets (e.g., populated areas that have no pipes) and missing attributes (e.g., pipe data that had no diameter or roughness values). The quality control analysis identified the Mayagüez Water Service Area as a high-quality candidate for validation water distribution system model development. This water service area contained both highly populated areas with a dense grid pipe layout and less populated areas with branching pipelines.

The water distribution network model was built in WNTR using a combination of raw data and assumptions regarding connectivity and other model attributes. Pipe diameter, roughness, and length were obtained from the database. Pipe connectivity was determined based on the proximity between pipe end points. A junction was added at each pipe end point, and elevation was assigned using United States Geological Survey Digital Elevation Model (DEM) data [35]. Junction demands were estimated using the total water treatment capacity and data on nearby building footprints. The elevation of each reservoir (water treatment plants) was assigned using the DEM data. Water pressure leaving the reservoir was set to 70 m. Tank height and diameter were obtained from the database. Some tank diameters were corrected using estimates from aerial photography. Tank elevations were assigned using the DEM data. Pumps were oriented such that they increase pressure as water travels away from the nearest reservoir. The flow capacity of the pumps was determined from the data, and headgain was set to 75 m. Since the model was intended to be used in steady-state simulations for this analysis, no demand patterns were added to the model. Given that the tank levels were unknown, the model was simulated for 48 h in an extended period steady-state simulation, and steady-state tank levels were assigned based on the time when the model system average pressure stabilized. While valve data existed, they were not added to the model due to uncertainty in their operations. The water distribution system in Mayagüez was known to have numerous pressure-reducing valves. Modeled pressures were generally higher than expected because those valves were not added to the model. Pertinent model attributes and the percentage of data that were

available for each asset are also listed in Appendix A. The resulting water distribution system model is shown in Figure 1. The Mayagüez model includes 7496 pipes, 6287 junctions, 124 tanks, 89 pumps, and 6 reservoirs. The average system pressure from a steady-state simulation was 75 m.

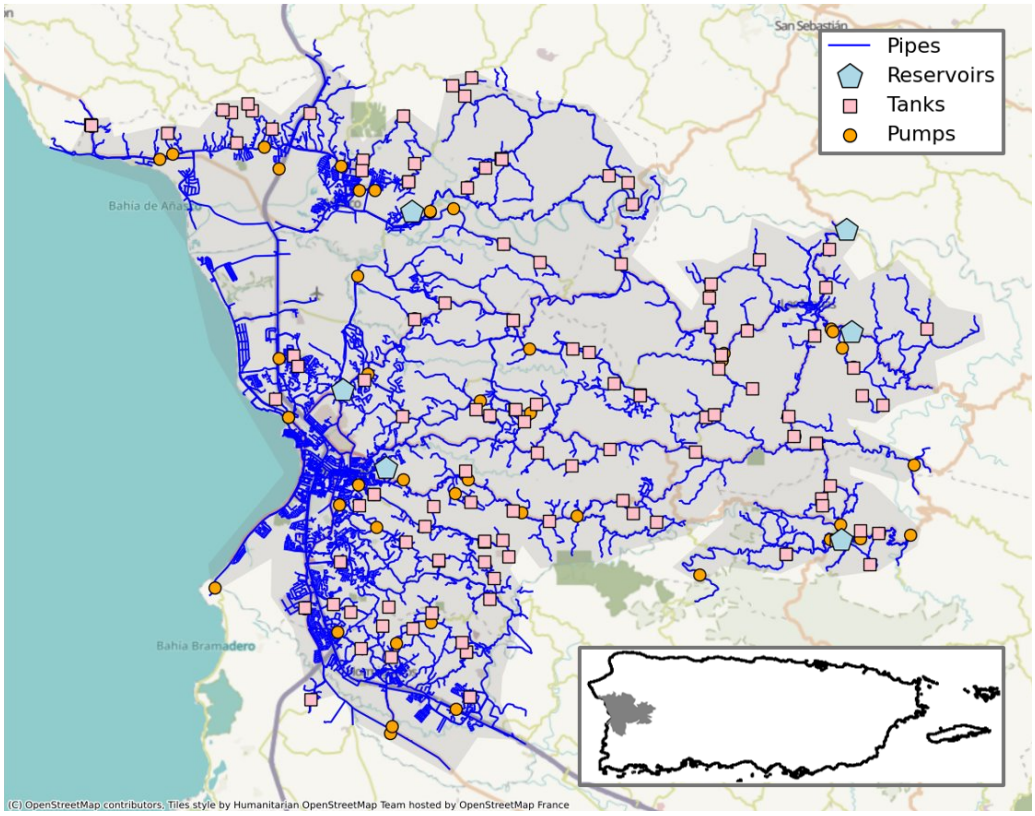

Figure 1. Pipe, reservoir, tank, and pump datasets in Mayagüez. Service area shown in transparent gray.

2.2. Synthetic Network Generation

The method for generating synthetic water distribution systems was built on previous work. In particular, the use of the minimum cost flow algorithm for pipe diameters came from Ahmad et al. [10], and the method of using a connectivity index for adding loops to a network came from Mair et al. [17]. Similar to the approaches of both papers, the road network covering the water service area was used as a candidate network structure, a subset of which was selected to form the synthetic water network. However, the method presented here differs from these methods through the use of a minimum cost flow to determine both a minimal spanning tree as well as the pipe diameters of the edges. Additionally, a modification of the minimum cost flow optimization problem (MMCF) allowed for the incorporation of real pipe data (location and diameters) to drive network generation toward the real system. Once the minimal water network was generated using the MMCF, network redundancy was added using the connectivity index from Mair et al. [17], and the rest of the water model was built by adding real reservoirs, tanks, and pumps from the utility dataset to the synthetic network via geospatial operations.

The following subsections further detail the methodology, including the data requirements and sources, steps involved in synthetic network generation, and a description of the synthetic networks generated to produce comparisons with the validation model.

2.2.1. Data Requirements and Sources

The minimum data requirements to produce a water distribution system model using this method were the service area boundary, road network, the location and magnitude of demand points, the location and capacity supplied from each reservoir, and elevation data

within the service area. In situations where the location and magnitude of demand were unknown (i.e., when there were no utility data or validation model), demands could be approximated using the building footprint.

Data sources used in this experiment include OpenStreetMaps [34] for road data, the Humanitarian Data Exchange [36] for building data, USGS [35] for elevation data, and the Puerto Rico infrastructure database [28] for data related to the Mayagüez water distribution system (i.e., service area, demands, and reservoirs).

2.2.2. Initialization

The service area boundary was approximated by creating a polygon around the water distribution assets within the Mayagüez Water Service Area. Road data within the service area were obtained and converted to a topological network using the Python package OSMNX [33].

Next, the water demand and supply were attached to the nodes of the road network. In this case, demand locations and magnitude were approximated from building footprints, using the same methods as those used to develop the validation model. The demand data were attached to the road network through a basic geospatial data processing technique known as "snapping", which determines the nearest road node for each demand. If multiple demands attach to the same road node, their summed demand was assigned to the node.

Water treatment plants were added as new nodes and connected to the nearest node in the road network. To ensure adequate connectivity between reservoirs and surrounding roadways, additional roads were added to the road network in each cardinal direction (north, south, east, west) within a 1500 m radius surrounding the facility. These additional roadways were potential pathways that would only be assigned flow if needed by the minimum cost flow problem. This approach generalized previous reservoir attachment methods, which assume that reservoirs have a single connection to the network [10,17]. This assumption does not hold true in Mayagüez, and negatively impacts synthetic model performance. A supply value for the MMCF was assigned to each reservoir using the total capacity of the reservoir.

2.2.3. Incorporation of Pipe Data

In order to leverage the capabilities of the MMCF, there must be a source of pipe diameter assignments to road edges. There were multiple options for where the diameters can come from. For example, known pipe diameters and locations could come from a subject matter expert (e.g., utility engineer), numerical analysis, machine learning, or sparse datasets.

In this analysis, a sparse dataset of pipe diameters was sourced from the utility dataset. The known pipe diameters were transferred from the pipe network to the road segments via a geospatial data processing technique called spatial association. The objective of spatial association was to accurately assign pipe diameters from a dataset to road segments based on nearby pipes. If no pipes were associated with a road, the assigned diameter was 0, and if there was at least one pipe associated to a road the assigned diameter was equal to the diameter of the largest pipe. These values were input into the MMCF to inform how the network layout and diameters were chosen. The spatial association method expands on the infrastructure collocation work by Mair et al. [21] and includes additional steps to account for potential road–pipe misalignment.

2.2.4. Modified Minimum Cost Flow

Minimum cost flow is an optimization problem designed to distribute a resource from sources to demands across a network. A candidate solution is one that finds flow values for each edge that balances the resource across source and demand nodes. In the standard formulation of minimum cost flow, a minimal spanning tree is a candidate solution that minimizes the sum of flows multiplied by corresponding weights.

Through solving the minimum cost flow on a road network, using sources and demands representing those of a water distribution system, a reasonable network for a water distribution system model can be obtained. Flow assignments to the roads can be converted to pipe diameters using the following formula:

$$d = \sqrt{(4Pf)/(\pi v)} \tag{1}$$

where d is diameter (m), P is the peak demand multiplier (unitless), f is flow (m^3/s), and v is velocity (m/s), which was set to 1.52 m/s for all pipes following the assumptions made in [10]. The peak demand multiplier was used to convert the average demand to peak demand to ensure the diameters were adjusted appropriately for high-flow periods. The value can be estimated empirically; however, this analysis used a value for P that was calibrated to produce a network with am average pipe diameter similar to that of the validation network. A value of $P = 6.76$ was chosen so that the average length-weighted diameter of the synthetic network matches that of the validation network, which was approximately 0.152 m. After applying this conversion, diameters equal to zero can be dropped from the network to obtain a synthetic network with diameters assigned to each edge.

One of the central contributions of this work is a modification of the standard minimum cost flow that allows for the input of target diameters. Target diameters are diameter values for a subset of road edges representing desired diameters for the final pipe network. These diameters could reflect known diameters of the system (e.g., a system operator knows that there is a 0.2 m pipe running alongside a certain road) or could come from another source such as machine learning predictions of pipe diameters for the system. In this analysis, known pipe diameters were assigned to nearby roads using spatial association to provide target diameters. The assignments from spatial association provide probable predictions of what the pipe diameter should be on each edge.

The modification was a change to the optimization function to favor solutions with flows close to the provided target flow, which were obtained by applying the inverse of Equation (1) to target diameters. For edges with a target flow, the objective function minimizes the difference between the target flow and the flow assigned to the edge. For edges without a target flow, the objective function minimizes flow along the edge, which was encoded by setting the target flow to zero.

Given a network with node set $\mathcal{V}$ and edge set $\mathcal{E}$, the MMCF optimization problem is formally defined as follows:

$$\text{minimize} \quad \sum_{(u,v)\in\mathcal{E}} W(u,v) \cdot |(F(u,v) - T(u,v))| \tag{2}$$

$$\text{subject to} \quad \sum_{v:(u,v)\in\mathcal{E}} F(u,v) - \sum_{v:(v,u)\in\mathcal{E}} F(u,v) = B(u) \quad \forall u \in \mathcal{V} \tag{3}$$

$$L(u,v) \leq F(u,v) \leq U(u,v) \quad \forall (u,v) \in \mathcal{E} \tag{4}$$

$$F(u,v) \geq 0 \quad \forall (u,v) \in \mathcal{E} \tag{5}$$

where

- $\mathcal{E}$ is the set of edges in the network;
- $\mathcal{V}$ is the set of nodes in the network;
- $W(u,v)$ is the weight on edge $(u,v) \in \mathcal{E}$;
- $F(u,v)$ is the flow on edge $(u,v) \in \mathcal{E}$;
- $T(u,v)$ is the target flow on edge $(u,v) \in \mathcal{E}$;
- $B(u)$ is the net supply (positive) or demand (negative) at node $u \in \mathcal{V}$;
- $U(u,v)$ and $L(u,v)$ are the lower and upper bounds on the flow for $(u,v) \in \mathcal{E}$.

Equation (2) defines the objective of the optimization problem, which minimized the difference between output flow and target flow if a target is provided and minimized the flow to zero if no target is provided. Equation (3) conserves flow throughout the nodes of

the network based on supply and demand values. Equation (4) enforces lower and upper bounds for flow assignments on the edges. The lower bound was set to zero and the upper bound was set to the flow value corresponding to a 1.829 m pipe diameter to ensure that pipe diameters were not too large. Equation (5) ensures that all assigned flows are positive.

The results produced flow values for each edge on a continuous scale that corresponded to continuous diameter values, as opposed to discrete values as is typically expected by engineering standards. In this analysis, the continuous values were not modified to match preset discrete values. Through dropping all edges that were assigned a flow value of zero, a minimal spanning tree that connects all demand nodes and reservoirs was obtained. The flow values were then converted to diameters using Equation (1). The continuous nature of the MMCF means that diameters can be arbitrarily small, so a minimum diameter of 0.051 m was applied (after dropping flow values that were zero). The resulting network and diameter assignments were used as the foundation for the synthetic water distribution system model, which was completed by following the steps outlined in the next subsection.

2.2.5. Post-Processing

The solution from the MMCF was designed to be minimal. There were no loops and no redundancy in the connectivity of the network. Since water distribution systems typically have some level of redundancy, a technique using the connectivity index was used to increase the connectivity of the network [22]. This process added additional paths along roadways if the existing path was much longer than the existing shortest path between its start and end node. This analysis added additional connections if the new path was at least 50% shorter.

Other major assets and data were added to the network to complete the model. Note that while model calibration was an important aspect of model development, calibration was not included in this analysis. The location and attributes of tanks and pumps were extracted from the validation models. Elevations were extracted from DEM data. The reservoir head was extracted from the validation model. Although the Puerto Rico datasets include valves, they were not used in this analysis due to a lack of certainty in their operation. Reservoirs and tanks were added by connecting them to the nearest junction in the network. Pumps were attached to the nearest pipe and oriented so that the pump faces away from the nearest reservoir. For more details, refer to Appendix A, where similar methods were used to generate the validation model.

2.2.6. Generated Models

Three synthetic models for Mayagüez were generated using the framework established in the previous sections. Each synthetic network used a different amount of randomly selected pipe layout and diameter data provided by spatial association, which assigned a diameter to each road based on real pipe data, with a diameter of zero indicating no pipe. A network built using 0% of the pipe data was generated to represent a typical synthetic network that had no external knowledge of the real pipe layout. Two networks were built using 50% and 100% of the pipe data to represent cases with partial and full access to the real pipe data, respectively. It is important to note that while the network built with 100% utilizes all pipe data, the data had been transferred from the real water pipe network to the road network and therefore lost accuracy. These synthetic networks are referred to as "Synthetic X%", where X is the percentage of pipe layout and diameter data that the network was built with. The validation network is referred to simply as "Validation".

2.3. Performance Metrics

To understand how the synthetic models compare to the validation model, several performance criteria were selected. Global topological metrics were included to get a sense of the network structure and compare the results to the work performed by Paez et al. [23]. The distribution of diameters and node pressure was included to identify

differences in network attributes and hydraulics. Finally, a disruption scenario was used to track how synthetic networks mimic the behavior of the validation model under stress. The disruption scenario selected for this analysis approximates a failure scenario that has frequently occurred at one of the primary water treatment plants in Mayagüez [37,38]. In addition to evaluating the similarity of synthetic networks to the validation network, the goal of this diverse suite of metrics was to understand how model-related quantities changed through the incorporation of various amounts of known pipe data to the MMCF.

2.3.1. Topology

A collection of global topological metrics were used to compare the overall structure of the networks. While not an exhaustive list, the chosen metrics aimed to cover a breadth of network qualities. Metric choices were informed by the work of Paez et al. [23]. To define these metrics, let $\mathcal{V}$ and $\mathcal{E}$ represent the nodes and edges of a network, and let the size of $\mathcal{V}$ be n and the size of $\mathcal{E}$ be m. Four topological metrics are given as follows [39]:

$$q = \frac{2m}{n(n-1)} \qquad < k > = \frac{2m}{n}$$

$$R_m = \frac{m-n+1}{2n-5} \qquad \Delta\lambda = \lambda_1(A) - \lambda_2(A)$$

Edge density q is the fraction between the number of edges in a network and maximum amount of edges possible and measures the overall connectivity of the network. Larger values indicate higher connectivity. An average node degree $< k >$ also captures connectivity but uses data from the nodes rather than edges. The meshedness coefficient R_m is the fraction between the number of independent loops in a network and the maximum number possible and captures the redundancy of the network. Larger values indicate higher redundancy. Spectral gap $\Delta\lambda$ is the difference between the first and second eigenvalues of the adjacency matrix for the network and measures the robustness of the network by determining whether a network has "good expansion" [39]. Good-expansion networks are those that have a high minimum number of components that must be removed before creating a disconnected network. While spectral gap is an abstract, unitless measurement, it can be used to compare networks where a larger value indicates a more robust network. Authors Paez et al. [23] called out spectral gap as uniquely useful for identifying differences in networks where other metrics indicate similarity. Together, these metrics summarize fundamental topological dimensions of water networks, facilitating a broad topological comparison across the synthetic and validation network structures.

2.3.2. Diameter

To evaluate the impact of including known pipe diameters in the MMCF, the diameter assignments and distributions of the synthetic networks were compared to those of the validation networks. A direct comparison of diameter assignments cannot be made between a synthetic network and the validation network since they exist on different topologies. Instead, diameter assignments of synthetic networks were evaluated against the spatial association results, which do share the same topological structure as the synthetic networks. The chosen metric was the length-weighted mean absolute diameter error (D-MAE) and is computed as the mean over the absolute value of each diameter error, which is the difference between the synthetically assigned diameter and the spatially associated diameter, and weighted by the edges' proportion of the total network length. Because the discretization of pipe segments differs between the synthetic networks and the validation network, weighting by length was used to normalize the error rather than treating all errors with equal weight. Let $\mathcal{E}$ be the set of edges in the road network data for the service area, let $l(e)$ be the fraction of total network length for edge e, let $d_{\mathrm{SA}}(e)$ be the spatially associated diameter assigned to edge e, and let $d_{\mathrm{synth}}(e)$ be the synthetic diameter assigned to edge e, and then the D-MAE is computed via

$$\text{D-MAE} = l(e) \times \sum_{e \in \mathcal{E}} \left| d_{\text{SA}}(e) - d_{\text{synth}}(e) \right|$$

The D-MAE retained the units of the input variables and was useful for its straightforward interpretation; for example, a D-MAE of 1 indicates that, on average, a pipe in the synthetic network deviates 1 m from the spatially associated diameter. High values for the D-MAE indicate a large mismatch between the diameter assignments of the given synthetic network and those captured by spatial association, whereas low values indicate similarity between synthetic diameters and spatial association diameters.

The following set of diameter bin centroids were used to generate diameter distributions: 0.051, 0.102, 0.152, 0.203, 0.254, 0.305, 0.356, 0.406, and 0.457 m. This binning scheme was designed to separate the major diameter sizes seen in Mayagüez while also remaining compatible with the continuous nature of the MMCF outputs. Instead of reporting on total edge counts for each bin, the percentage of network length was computed to account for the varying discretization of pipes across networks.

2.3.3. Pressure

Steady-state pressures were simulated and compared to form an understanding of the hydraulic differences of the models during normal conditions. Initial tank levels were highly influential in these simulations; however, these levels were uncalibrated in the validation model for Mayagüez since it was built from basic data. To obtain reasonable initial values, a 96 h simulation starting with all tanks being 90% full was run for the validation model, and tank levels were extracted from a time point where the system pressure had stabilized. These individual tank levels were then applied to the validation and synthetic models.

The synthetic models have distinct network structures from that of the validation model, so there cannot be a direct comparison at the node level. Given this restraint, a comparison of pressure was generated by averaging nodal pressures within each census block group. While some resolution in the pressure results was lost when aggregating to census block groups, general spatial patterns of pressure were still able to be captured. It is important to note that pressure-dependent demand simulations using EPANET can still result in junctions with negative pressure. Since negative pressure has no physical meaning beyond the absence of pressure, the negative pressures were converted to zero before averaging the nodal pressures within each census block group.

To quantitatively compare pressure between models, the mean absolute error (MAE) was used to summarize differences in the census block group pressure. MAE is defined as follows, where i indexes over the census block groups, $p_{\text{synth}}(i)$ is the i^{th} census block group pressure in one of the synthetic networks, and $p_{\text{valid}}(i)$ is the i^{th} census block group pressure for the validation network:

$$\text{P-MAE} = \sum_{i} \left| p_{\text{synth}}(i) - p_{\text{valid}}(i) \right|.$$

2.4. Response to Disruption

A disruption scenario was developed based on recent events where water service from the Miradero water treatment plant was temporarily disabled. The Miradero water treatment plant in Mayagüez, Puerto Rico, has experienced multiple breakdowns in its raw water supply line, causing significant water shortages in the area. In April 2023, raw water intake from the Río Grande de Añasco to the Miradero plant required emergency replacement [37], and similarly, in late December 2023, the pumping system for the same plant experienced a breakdown over multiple days, preventing the plant from distributing drinking water to many residents and critical locations such as hospitals [38]. After repairs began on the pumping system, additional leaks on a 0.1524 m pipe were identified, which

extended the period of repair. The event was declared a state of emergency by the mayor to supply water to residents via trucks.

This disruptive event was modeled by setting the base head of the Miradero reservoir to zero meters in the model and simulating for 48 h using the same initial conditions as the steady-state simulations in Section 2.3.3. While the models were not fully calibrated for long-term simulation (lacking demand patterns, valves, and controls), these details were consistent across the validation and synthetic models. Nodal pressure drops between the baseline and failure scenarios are plotted at the last time step (48th hour) to reveal spatial patterns that do or do not emerge across the synthetic and validation models.

3. Results

Throughout the following sections, the synthetic networks are compared to the validation network across multiple performance metrics. These results seek to answer two questions: (1) how are the synthetic models different or similar to the validation model and (2) how do the performance criteria change as different amounts of pipe data are provided to the MMCF algorithm. In Section 3.1, four global topological metrics are evaluated across the networks. In Section 3.2, diameters assigned to pipes are compared. In Section 3.3, steady-state nodal pressures are compared. In Section 3.4, the pressure response to the same disruptive event is compared across the networks.

3.1. Topology Comparison

The four topological metrics described in Section 2.3.1 were evaluated for the three synthetic networks and the validation network. The results are presented in Table 1. The metrics include edge density, average node degree, meshedness coefficient, and spectral gap. While there are countless metrics to choose from, the chosen metrics cover a range of network qualities, as laid out by Yazdani et al. [39]. In addition to the computed topological metrics, the number of nodes and edges, as well as total network length for each network, are included.

Table 1. Topological metrics.

	Validation	**Synthetic-0%**	**Synthetic-50%**	**Synthetic-100%**
Total edges	7585	8570	9014	9459
Total nodes	6417	7460	7657	7802
Network length (m)	1,302,762	1,190,159	1,233,406	1,287,466
Edge density, q	0.00037	0.00031	0.00031	0.00031
Average node degree, $< k >$	2.36	2.30	2.35	2.42
Meshedness coefficent, R_m	0.091	0.075	0.089	0.11
Spectral gap, $\Delta\lambda$	1.35	7.12	7.29	7.34

The three synthetic networks generally have one to two thousand more edges and nodes than the validation network; however, the validation network has a longer total network length than the synthetic networks. The difference in edge and node count can be attributed to differences in the discretization of the water pipe data (longer line segments) versus the road data (shorter line segments). When pipe diameter data are included in synthetic network generation in the case of the Synthetic-50% and Synthetic-100% networks, additional target flows are added to the MMCF, which results in more edges being included in the network diameter since edges with a target flow are encouraged to have positive values, instead of being minimized to zero.

By design, the MMCF will minimize flow on edges where no pipe data are present, so the outputs with no or partial data (Synthetic 0% and Synthetic 50%) tend to have a shorter total length than the networks built with all of the data (Validation and Synthetic 100%).

Edge density measures the overall connectivity of a network, and its value can be directly interpreted as the percentage of edges in the network out of the total possible number of edges given the node set. The validation network has a very low edge density, which is matched closely by the synthetic networks. There is little difference among the synthetic networks, indicating that the inclusion of pipe data does not affect the overall connectivity of the networks. Average node degree gives an alternate view on the overall connectivity of the networks. All four networks have an average node degree in the range 2.3 to 2.4, and as the percentage of pipe data increases, there is a slight increase in average node degree. The meshedness coefficient measures network redundancy and is similar to edge density, but instead of measuring the percentage of total possible edges, it measures the percentage of total possible loops. The synthetic networks have similar meshedness coefficient percentages to the validation network; Synthetic 0% has a slightly lower value, Synthetic 50% closely matches the validation network, and Synthetic 100% has a slightly higher value. The overall similarity to the validation network indicated that the choice of 50% for the connectivity index threshold when adding loops to the networks was appropriate. Spectral gap is an abstract metric extracted from the eigendecomposition of the adjacency matrix and is known as a measurement of network robustness. The spectral gaps of the synthetic networks were all significantly higher than that of the validation network, indicating that the synthetic networks were more robust. One interpretation of spectral gap is that low values indicate the presence of edges or nodes that would disrupt network flow if removed, while higher values indicate that many nodes and edges could be removed before network flow is majorly effected [39].

3.2. Diameter Comparison

All four models share a similar overall network layout, indicating the effectiveness of using the road network for synthetic pipe network generation (Figure 2). Despite these large-scale similarities, there are details that indicate significant nuances. When focusing on large-diameter pipes (those plotted in red), there are two areas in particular that vary across the synthetic networks, which are captured in Figure 2 with black rectangles. The large-diameter pipes in the western region and the northeast region are laid out very differently between the validation network and the 0% network. However, the 50% and 100% networks appear to replicate the structure of the large-diameter pipes found in the validation network much better.

The distribution of diameters in the synthetic models was distinct from that of the validation model; however, the addition of more pipe data influenced the synthetic pipe diameter distribution toward that of the validation model (Figure 3). In the synthetic network, the majority of network length is covered by small pipes (between 0.0127 and 0.0635 m), and the network length covered by a diameter bin decreases as the diameter size increases. This pattern mostly holds true for the validation network; however, it is the second smallest diameter bin (between 0.0635 and 0.1143 m) that contains the majority of network length. The synthetic networks built with more pipe data have distributions that more closely match that of the validation network, particularly in the first three bins. Synthetic 0% clearly favored small diameters, which is the natural conclusion of an algorithm driven exclusively by minimizing flows. On the other hand, the Synthetic 50% and Synthetic 100% networks favor the smallest diameter bin less and assign larger diameters more often, reflecting the incorporation of information about the real network diameters. There is less of an improvement in the larger-diameter bins, which reported much noisier results as data were added.

The length-weighted mean absolute error (D-MAE) is computed between the diameter assignments in each synthetic network and the diameter assignments from spatial association. When pipe layout information was provided to the MMCF, there was a notable decrease in the D-MAE of the synthetic model diameters on roads against those assigned by spatial association, which confirmed that the MMCF was incorporating information

about the network correctly. The D-MAE for Synthetic 0% was 0.0983 m, the D-MAE for Synthetic 50% was 0.0711 m, and the D-MAE for Synthetic 100% was 0.0575 m.

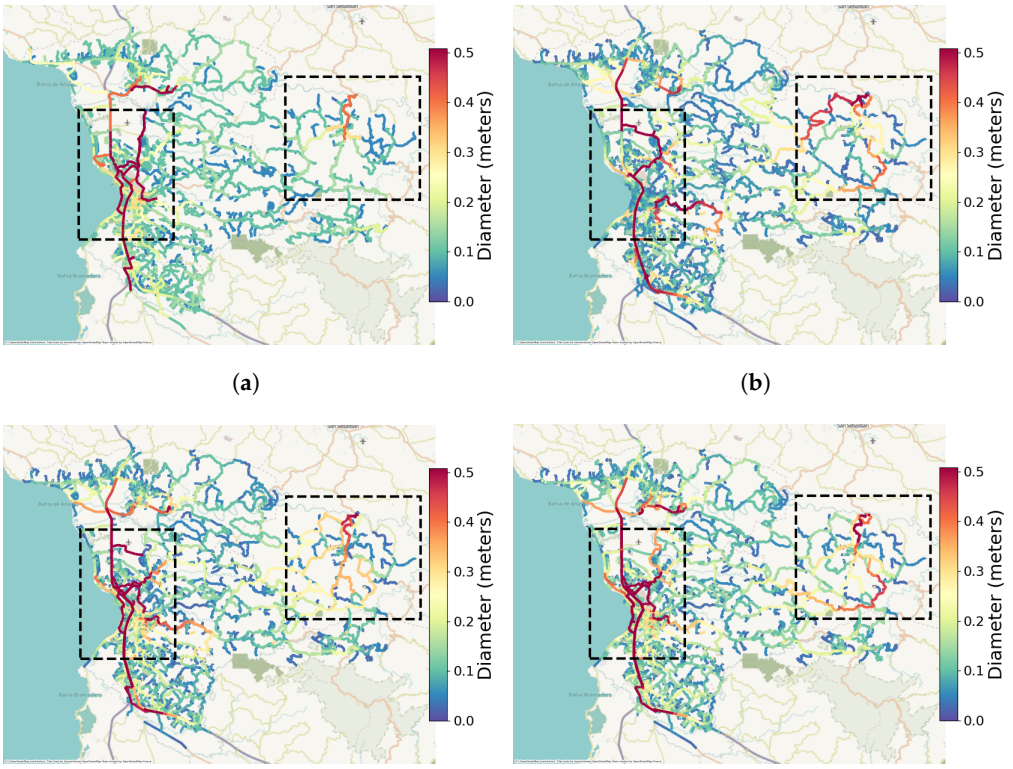

Figure 2. Validation and synthetic models with pipes colored by diameter (meters). Two regions, one western and one eastern, with significant structural differences are highlighted by boxes. (**a**) Validation; (**b**) Synthetic 0%; (**c**) Synthetic 50%; (**d**) Synthetic 100%.

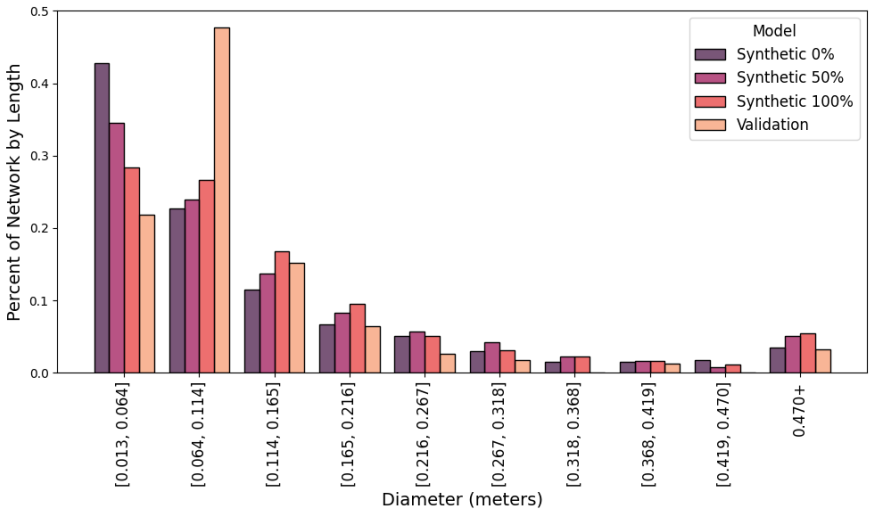

Figure 3. Distribution of network length across diameter bins.

3.3. Pressure Comparison

Nodal pressures from steady-state simulation are shown in Figure 4. In general, pressures are very high across all models, resulting from the lack of pressure-reducing valves. The models display a large amount of heterogeneity both internally and when compared to each other. In the validation network, the city center along the west coast

tends to have large pressures with pressure decreasing in all directions. However, the synthetic networks maintain high pressure where the validation network does not to the north and south of the city center.

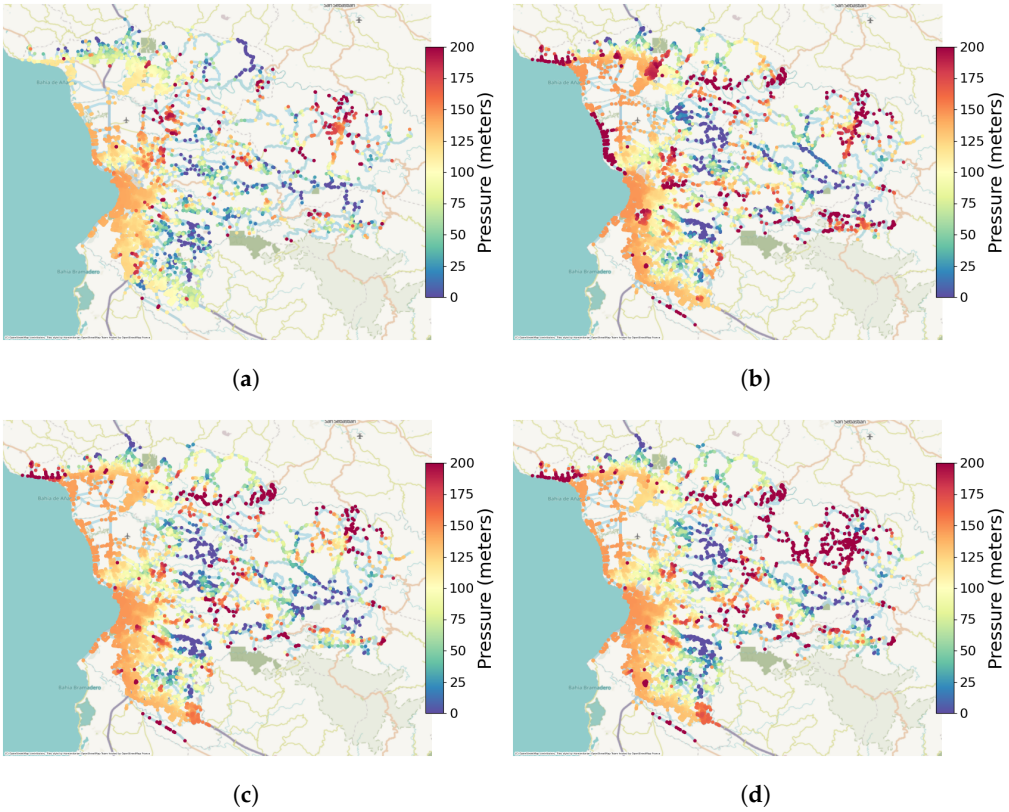

Figure 4. Nodal pressures from steady-state simulation. (**a**) Validation; (**b**) Synthetic 0%; (**c**) Synthetic 50%; (**d**) Synthetic 100%.

To describe pressure more quantitatively, nodal pressures in each network are averaged within census block groups (CBGs), and the CBG pressures of the synthetic networks are plotted against the CBG pressures of the validation network in Figure 5. The CBG pressures for each network roughly follow a linear relationship with the CBG pressures of the validation network, indicating that CBGs with low pressure in the validation generally have a low pressure in the synthetic networks and likewise for high pressures. There is still a considerable amount of deviation from the validation pressures present in the synthetic models, which can be summarized by computing the MAE between average CBG pressures for the validation network and each synthetic network. The CBG pressure MAE for Synthetic 0% is 31.6 m, the MAE for Synthetic 50% is 23.5 m, and the MAE for Synthetic 100% is 23.2 m.

3.4. Response to Disruption Comparison

The pressure response to the disruptive scenario of each synthetic network matched the spatial patterns of the pressure response to the same scenario in the validation network, Figure 6. The average nodal pressure drops across each model are as follows: 32.5 m for validation, 64.0 m for Synthetic 0%, 21.0 m for Synthetic 50%, and 15.5 m for Synthetic 100%. The pressure drop for the Synthetic 0% model is much more severe than the other models, and the synthetic models with additional pipe data experience pressure drops smaller than the pressure drop in the validation model. The validation network shows that the area affected by the disruption is primarily in the downtown area surrounding the water treatment plant and the region of the network directly south (Figure 6a). This spatial pattern is broadly captured by the synthetic networks, with some notable differences occurring in the northern region of the network, where the synthetic networks model a drop

in pressure that is not present in the validation network. In the affected region, Synthetic 0% (Figure 6b) has extreme pressure drops of up to 100 m, while Synthetic 50% (Figure 6c) and Synthetic 100% (Figure 6d) have moderate pressure drops of around 40 m, matching what is seen in the validation network more closely.

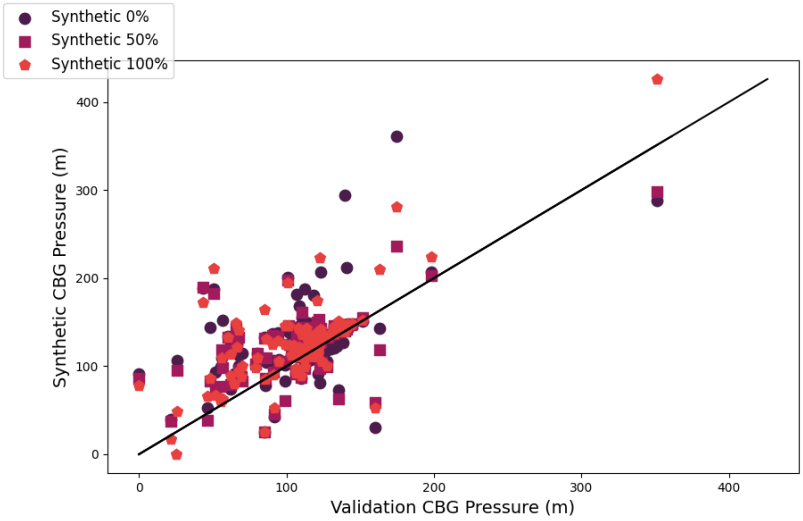

Figure 5. Scatterplot of synthetic pressures against validation pressures, averaged within CBGs. Each CBG is represented three times for each synthetic model. CBG boundaries for Mayagüez are shown in the inset map.

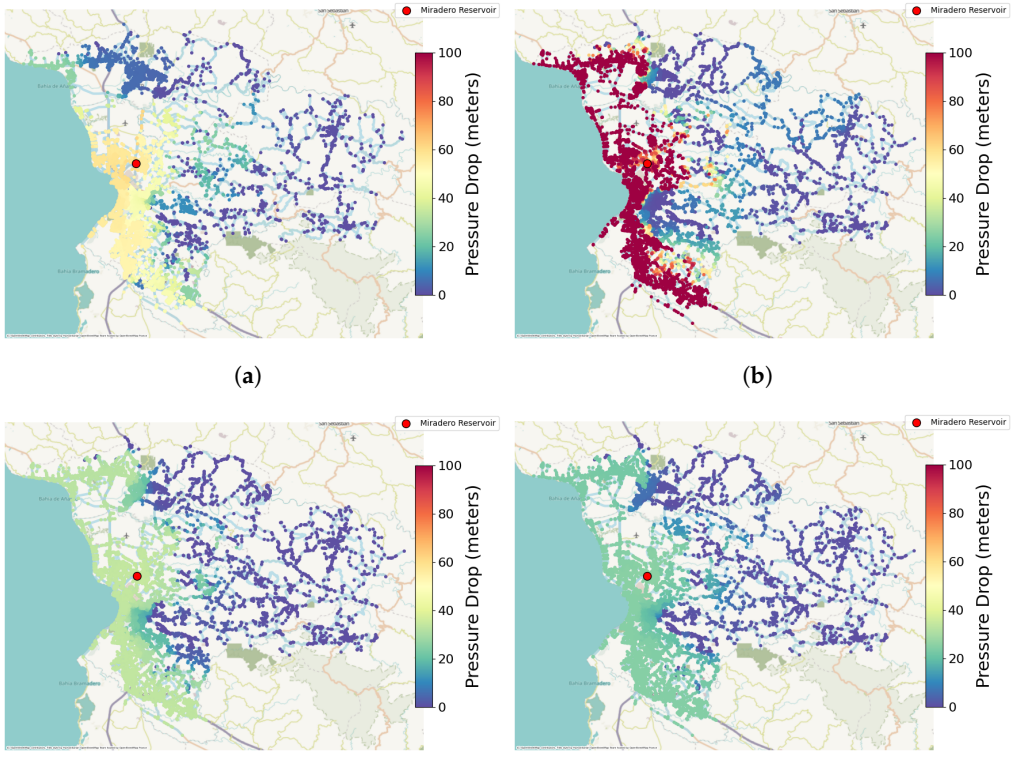

Figure 6. Nodal pressures drop 48 h after failure at the Miradero plant; pressure drop is computed as the difference between the baseline and failure scenario. (**a**) Validation; (**b**) Synthetic 0%; (**c**) Synthetic 50%; (**d**) Synthetic 100%.

4. Discussion

The MMCF algorithm successfully demonstrated the ability to improve synthetic pipe layout and diameter generation by incorporating information provided by a pipe dataset (Figures 2 and 3). The topological assessment of the models indicated that while the incorporation of pipe data affected various topological metrics, it did not uniformly increase similarity to the validation network's topological properties (Table 1). Spectral gap in particular suggested a deep structural difference between the road network and pipe network that could not be overcome with the addition of data.

The steady-state pressure of the synthetic models reflected the pressures of the validation model more accurately as additional data were included in model generation (Figure 4). However, even when the complete pipe dataset for the real system was used for synthetic model generation, there was a sizeable gap between synthetic and validation pressure. This suggested that constraining the synthetic models to the road network may come with inherent limitations for reproducing the conditions of the real system. Additionally, the pressure response of the synthetic models to a reservoir failure was compared to the pressure response of the validation model. All synthetic models showed the capability of capturing the same general spatial patterns of pressure drop caused by the reservoir failure as those found in the validation model.

Limitations and Future Work

While Mayagüez is a complex and diverse water service area to study, containing many reservoirs, tanks, and pumps and covering both urban and rural areas, a singular case study is not enough to make generalizations about the methods. It would be informative to evaluate additional locations for the suitability of synthetic networks for resilience analysis. Additionally, it would be valuable to apply this comparison on models with valves and controls, which was not implemented in this work. The lack of utility right-of-ways in the road network could preclude certain critical network connections from being made in the synthetic networks (e.g., large pipes that do not follow roads), so methodological improvements could include the incorporation of the pipe network edges into the road network, as was performed in the work by Mair et al. [22], or through the inclusion of utility right-of-way datasets.

The MMCF algorithm provides an opportunity to customize synthetic network generation so that the layout and pipe diameters more closely reflect provided inputs. In this work, pipe datasets are used in conjunction with spatial association to provide this input to the MMCF; however, this water infrastructure information could come from data-driven machine learning models [40,41] or operator expertise. While a single synthetic model may never perfectly capture the dynamics of the real system, ensemble approaches that generate many synthetic models may be able to provide a more accurate picture of system resilience [41], which could be produced by incorporating uncertainty into the synthetic network generation method [42]. Other types of failure (e.g., pipe breaks and leaks, pump failures, water quality) and response actions could be simulated and compared along with a response action for a deeper resilience assessment of synthetic models [27].

5. Conclusions

This work focused on the development of synthetic water distribution models from sparse data with a focus on generating and evaluating site-specific synthetic models for use in resilience analysis. The objectives of this work were to (1) design a synthetic model generation technique that can leverage incomplete pipe data to improve model performance and (2) provide a thorough comparison of multiple synthetic models generated with various amounts of pipe data to the validation model in Mayagüez. The spatial association and modified minimum cost flow (MMCF) methods presented in this paper provide a way to incorporate pipe network data into the generation of a synthetic water network. Three synthetic models were generated with this technique, each corresponding to models generated using 0%, 50%, and 100% of real pipe data. These models were compared to a

validation model in terms of topological metrics, pipe diameters, steady-state pressure, and pressure response to a disruptive scenario.

This research highlights opportunities for synthetic network generation for site-specific analysis. The augmentation of synthetic network generation with partial pipe data allows for the creation of synthetic models that more accurately reflect the real system than those generated without any pipe data. The ability of the synthetically generated models to reproduce the system response to a disruptive event that was seen in the validation model suggests that synthetic models could be used in resilience analysis for systems without access to a high-quality model. However, despite the promising similarity between synthetic models and the validation model, even small differences could result in drastically different simulation results due to the complexities of these systems. For this reason, additional research must be performed on synthetic networks before they can be used to make critical decisions for the system they represent.

Author Contributions: K.L.B.: methodology, writing, software, formal analysis, conceptualization, visualization, and data curation; K.A.K.: methodology, software, writing, conceptualization, data curation, project management, supervision, and funding acquisition; J.W.P.: writing—review and editing, and validation; S.R.: methodology, conceptualization, and writing—review and editing; I.S.: conceptualization and writing—review and editing; M.C.: conceptualization, writing—review, and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded through the Laboratory Directed Research and Development (LDRD) for the Resilient Energy Systems (RES) Mission Campaign at Sandia National Laboratories under project number 23-0069. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

Data Availability Statement: Open source geospatial data of Puerto Rico infrastructure along with a digital elevation map was used to create the ground truth, or validation, water distribution system model for the Mayagüez water service area. The Puerto Rico infrastructure database is publicly available at https://gis.pr.gov/ (accessed on 1 October 2022) [28]. The United States Geological Survey Digital Elevation Model (DEM) data is available at https://apps.nationalmap.gov/downloader/ (accessed on 1 February 2023) [35]. The roads of the defined service area are sourced from OpenStreetMaps [34] and converted to a topological network using the Python package OSMNX [33]. Building footprint data is available from the Humanitarian Data Exchange project at https://data.humdata.org/dataset/hotosm_pri_buildings (accessed on 1 November 2022) [36]. Data outputs such as generated models are subject to Sandia copyright and information review and will be made available upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CBG	Census Block Groups=;
DEM	Digital Elevation Map;
D-MAE	Length-Weighted Mean Absolute Error;
GIS	Geographic Information System;
P-MAE	Pressure Mean Absolute Error;
MMCF	Modified Minimum Cost Flow.

Appendix A. Mayagüez Water Distribution Model

Open source geospatial data of Puerto Rico infrastructure along with a digital elevation map were used to create the ground truth, or validation, water distribution system model for the Mayagüez water service area. The Puerto Rico infrastructure database are publicly

available at https://gis.pr.gov/ (accessed on 1 October 2022) [28]. The United States Geological Survey (USGS) Digital Elevation Model (DEM) data is available at https://apps.nationalmap.gov/downloader/ (accessed on 1 February 2023) [35] and building footprint data are available from the Humanitarian Data Exchange project at https://data.humdata.org/dataset/hotosm_pri_buildings (accessed on 1 November 2022) [36]. The steps and assumptions that were used to build the model are described below. Model attributes and the percent of data that are available for each attribute are listed in Table A1. This table includes a description of how missing data were handled. The Water Network Tool for Resilience (WNTR) [5] was used to create a steady-state water distribution system model and process geospatial data.

Note that while valve data exist for the water service area, valves were not added to the model at this time due to uncertainty in their operations. Furthermore, the model does not include pump controls or demand time series. These updates could be added in future research.

Step 1: Extract data within the water service area: The database, which includes all pipes across the island, is first filtered according to an attribute of water pipes that defines the water service area. Data associated with the Mayagüez water service area are used to create a polygon defined as the convex hull around this set of pipes. This polygon is used to isolate additional data (i.e., water treatment plants, tanks, pumps) that should be included in the water service area.

Step 2: Connect pipes and create junctions: Each pipe is defined with a line geometry, which can include a single straight line segment or multiple straight line segments. However, the data do not include junctions that define how pipes are linked together. In order to connect pipes, a small threshold was used to identify pipe endpoints that are within close proximity. In this analysis, this threshold was set to 10 m. Endpoints that are within 10 m are grouped together to form a single junction. NetworkX [43] is then used to identify regions of the pipe layout (network structure) that are not fully connected. For the Mayagüez water service area, the largest connected component makes up 93.3% of the pipes. The other unconnected regions are along the outer boundary of the system and are removed.

Step 3: Assign elevation to each junction: Elevation is assigned to each junction using DEM data from the USGS [35].

Step 4: Assign demand to each junction: Demand is assigned to each junction using the assumption that demand is proportional to a nearby building footprint. Building polygons were obtained from OpenStreetMaps [36]. WNTR was used to identify building centers that are within 250 m of each pipe. The building footprint (measured in square meters) was then assigned to the nearest pipe endpoint. The base demand for each junction is defined as a weighted fraction of the total water treatment capacity for the water service area. The water treatment capacity for each water treatment plant within the service area was obtained from the database.

Step 5: Add junctions and pipes to the model: The junctions and pipes identified above are then added to an empty water distribution system model. Junction attributes include the junction coordinate, elevation, and base demand (described above). Pipe attributes include the start and end junction, length, diameter, and roughness. Pipe length was computed from the pipe line geometry. Pipe diameters and roughness are obtained from the database. Pipes with a missing diameter were assigned a diameter of 6 inches. Pipes with missing roughness were assigned a roughness of 100.

Step 6: Add reservoirs to the model: Water treatment plants are modeled as reservoirs. The location and elevation of each reservoir is determined based on the nearest junction. Water pressure from each reservoir is assumed to be 70 m. This sets the reservoir head at 70 m (100 psi) above the reservoir elevation.

Step 7: Add tanks to the model: The location and elevation of each tank is determined based on the nearest junction. Tank height and diameter are obtained from the database. Some tank diameters were corrected using aerial photography. The tank height is used to

define the maximum fill level. The minimum fill level was set to 0, and the initial water level was each tank was set to 90% of the maximum fill level.

Step 8: Add pumps to the model: The location of each pump is determined based on the nearest pipe. Two junctions are added to the nearest pipe, and the pump is inserted as a line geometry between them. The pump is oriented to face away from the nearest reservoir. Distance to the nearest reservoir is computed using the NetworkX multi-source Dijkstra path length algorithm [43]. The pump curve is defined using a flow capacity from the database. The headgain for each pump is set to 75 m (106 psi). The pumps are not assigned controls, the pumps are assumed to be on all the time.

Step 9: Refine the model: Minor model refinements were identified based on results from hydraulic simulations. This identified potential missing pipes that would connect critical pathways, and missing boundary conditions to represent water demands outside the water service area.

Table A1. Water distribution system data for the Mayagüez water service area. The fraction of data available is listed in Column 3. Notes on default values, additional data sources, and estimations are listed in Column 4.

	Attribute	Data Availability	Notes
Pipes	Diameter	1.00	If missing, default value is 6 in
	Roughness	0.97	If missing, default value is 100
	Length	1.00	Computed from line geometry
	Connectivity	None	Estimated from endpoint proximity
Junctions	Elevation	1.00	Assigned using USGS DEM
	Demand	None	Estimated from building footprint
Reservoirs	Elevation	1.00	Assigned using USGS DEM
	Head	None	Computed, elevation + 70 m pressure
Tanks	Height	0.86	If missing, default value is average tank height
	Diameter	0.27	Override based on visual inspection where applicable or use default value of average tank diameter
	Elevation	1.00	Assigned using USGS DEM
Pumps	Direction	None	Estimated, pumps point away from nearest reservoir
	Flow capacity	0.64	If missing, default value is average pump flow capacity
	Headgain capacity	None	Estimated, based on elevation difference between pump and highest system elevation (with a maximum of 500 m)

Appendix B. Spatial Association

Spatial association is used to assign properties from the utility dataset to their nearest road segment. This facilitates the transfer of information from one network to another. Spatial association assigns pipe properties to road segments that are in close proximity. Proximity is captured using geometric buffers around the pipes and find which roads intersect with a particular pipe. These intersections are used as a basis for associating a pipe and pipe attributes (e.g., diameter, roughness) to a road. This approach works in certain ideal scenarios but fails to capture the nuance and variety present across water distribution system models. Three key issues have been identified with the stated approach—segmentation misalignment, directional misalignment, and positional misalignment—as shown in Figure A1. In what follows, each of these issues and the measures taken to resolve them are described.

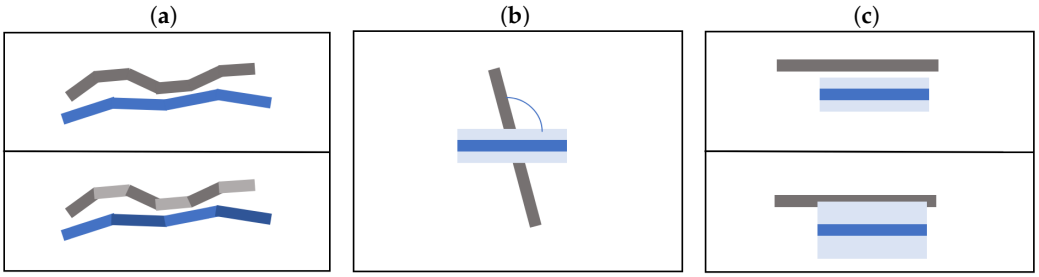

Figure A1. Diagrams illustrating issues with spatial association: (**a**) segmentation misalignment, (**b**) directional misalignment, and (**c**) positional misalignment. In each figure, the gray line is a road segment, the dark blue line is a water pipe segment, and the light blue rectangle is a buffer around the water pipe.

Segmentation misalignment occurs because individual pipes and roads do not necessarily have the same breakpoints. For example a long segment of road may be represented as a single linestring, but a parallel pipe that follows the road could be represented by multiple smaller linestrings arranged in parallel.

These kinds of inconsistencies can misinform the statistics that are assigned for diameter and roughness. To address this, a procedure known as "exploding" is applied to the geometries before performing intersection. Exploding splits linestrings into multiple straight line segments. For example, applying the procedure to a single pipe asset that has a linestring geometry made of three connected straight line segments produces three new pipe assets arranged in parallel. By exploding pipe and road geometries, the intersection of the two infrastructures may be performed at a higher resolution, resulting in more accurate statistics for diameter and roughness.

Directional misalignment occurs because some road geometries run perpendicular to some pipe geometries. If direction is not considered, a pipe can be associated with a perpendicular road, which is inaccurate. The computation of the relative bearing between pipe and road intersections can be used to assess whether or not a road is aligned with a pipe. The metric used for relative bearing here is cosine similarity, which is defined as follows:

$$\text{cosine similarity} := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{A1}$$

where **A** represents the geometry of a road, and **B** represents the geometry of a pipe. A cosine similarity of 0 indicates orthogonality, and a cosine similarity closer to -1 or 1 indicates that the geometries are parallel. In this analysis, intersections where $|\text{cosine similarity}| < 0.7$ are removed to prevent the spatial association of pipes with perpendicular roads.

Positional misalignment occurs due to variations in how closely pipes and roads are situated. Certain areas of the service region may have very tight proximity between roads and pipes, whereas in other areas, the proximity may be much looser. A small pipe buffer size can appropriately capture the intersections of the former scenario, but not the latter. A larger pipe buffer size may capture the intersections in the latter scenario as well, but it also captures undesirable intersections in the more tightly arranged areas.

This misalignment is addressed through the implementation of an iterative procedure that makes use of variable buffer sizes, rather than choosing a single fixed buffer size. Intersections are first identified with the smallest buffer, and for each road geometry, there is a list of water pipes whose buffers overlap with the road. For each intersecting water pipe, a copy of the road geometry is created and inherits the associated water pipe attributes. These water pipes are then dropped out of the dataset, and the procedure is run again with a larger buffer size.

Below is a summary of the general spatial association procedure including the adjustments made for the three types of misalignment:

- Explode pipes into straight line segments.
- Expand each pipe line segment into a rectangle by adding a small buffer around the geometry.
- Intersect each pipe rectangle with the road linestrings using the iterative approach described above.
- Filter intersections based on cosine similarity between the pipe and road. Intersections that are too close to perpendicular are removed.
- Assign attributes of interest (e.g., diameter and roughness) to each road based on the intersecting pipe geometries. In the case of multiple intersections, a statistic such as the maximum or average should be taken.

The results from spatial association can be used to build synthetic networks with pipe diameter data associated with road segments, as described in Section 2.2. Other applications include machine learning, where the spatial associated pipe property is used to predict the presence of pipes along roadways.

References

1. Cunha, M.; Marques, J.; Creaco, E.; Savić, D. A dynamic adaptive approach for water distribution network design. *J. Water Resour. Plan. Manag.* **2019**, *145*, 04019026. [CrossRef]
2. Anchieta, T.; Meirelles, G.; Carpitella, S.; Brentan, B.; Izquierdo, J. Water distribution network expansion: An evaluation from the perspective of complex networks and hydraulic criteria. *J. Hydroinform.* **2023**, *25*, 628–644. [CrossRef]
3. Hatchett, S.; Uber, J.; Boccelli, D.; Haxton, T.; Janke, R.; Kramer, A.; Matracia, A.; Panguluri, S. Real-time distribution system modeling: Development, application, and insights. In Proceedings of the Eleventh International Conference on Computing and Control for the Water Industry, Exeter, UK, 5 September–7 September 2011; Centre for Water Systems, University of Exeter: Exeter, UK, 2011.
4. Jun, S.; Lansey, K.E. Comparison of AMI and SCADA systems for leak detection and localization in water distribution networks. *J. Water Resour. Plan. Manag.* **2023**, *149*, 04023061. [CrossRef]
5. Klise, K.A.; Bynum, M.; Moriarty, D.; Murray, R. A software framework for assessing the resilience of drinking water systems to disasters with an example earthquake case study. *Environ. Model. Softw.* **2017**, *95*, 420–431. [CrossRef]
6. Nikolopoulos, D.; Ostfeld, A.; Salomons, E.; Makropoulos, C. Resilience assessment of water quality sensor designs under cyber-physical attacks. *Water* **2021**, *13*, 647. [CrossRef]
7. Meng, F.; Fu, G.; Farmani, R.; Sweetapple, C.; Butler, D. Topological attributes of network resilience: A study in water distribution systems. *Water Res.* **2018**, *143*, 376–386. [CrossRef]
8. Nunes, R.; Arraut, E.; Pimentel, M. Risk assessment model for the renewal of water distribution networks: A practical approach. *Water* **2023**, *15*, 1509. [CrossRef]
9. Torres, J.M.; Brumbelow, K.; Guikema, S.D. Risk classification and uncertainty propagation for virtual water distribution systems. *Reliab. Eng. Syst. Saf.* **2009**, *94*, 1259–1273. [CrossRef]
10. Ahmad, N.; Chester, M.; Bondank, E.; Arabi, M.; Johnson, N.; Ruddell, B.L. A synthetic water distribution network model for urban resilience. *Sustain. Resilient Infrastruct.* **2022**, *7*, 333–347. [CrossRef]
11. Hoff, R.; Chester, M. Preparing infrastructure for surprise: Fusing synthetic network, interdependency, and cascading failure models. *Environ. Res. Infrastruct. Sustain.* **2023**, *3*, 025009. [CrossRef]
12. Gegner, K.M.; Birchfield, A.B.; Xu, T.; Shetye, K.S.; Overbye, T.J. A methodology for the creation of geographically realistic synthetic power flow models. In Proceedings of the 2016 IEEE Power and Energy Conference, at Illinois (PECI), Urbana, IL, USA, 19–20 February 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–6.
13. Schweitzer, E.; Scaglione, A.; Monti, A.; Pagani, G.A. Automated generation algorithm for synthetic medium voltage radial distribution systems. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **2017**, *7*, 271–284. [CrossRef]
14. Sitzenfrei, R.; Fach, S.; Kleidorfer, M.; Urich, C.; Rauch, W. Dynamic virtual infrastructure benchmarking: DynaVIBe. *Water Sci. Technol. Water Supply* **2010**, *10*, 600–609. [CrossRef]
15. Sitzenfrei, R. A review on network generator algorithms for water supply modelling and application studies. In Proceedings of the World Environmental and Water Resources Congress 2016, West Palm Beach, FL, USA, 22–26 May 2016; pp. 505–515.
16. Duque, N.; Bach, P.M.; Scholten, L.; Fappiano, F.; Maurer, M. A simplified sanitary sewer system generator for exploratory modelling at city-scale. *Water Res.* **2022**, *209*, 117903. [CrossRef] [PubMed]
17. Mair, M.; Rauch, W.; Sitzenfrei, R. Spanning Tree-Based Algorithm for Generating Water Distribution Network Sets by Using Street Network Data Sets. In Proceedings of the World Environmental and Water Resources Congress 2014, Portland, OR, USA, 1–5 June 2014; pp. 465–474. [CrossRef]

18. Zauscher, E.; Berglund, E.Z. Validating a Methodology for Generating Water Infrastructure Network Models. In Proceedings of the World Environmental and Water Resources Congress 2024, Milwaukee, WI, USA, 19–22 May 2024; pp. 1380–1389. [CrossRef]
19. Sitzenfrei, R.; Möderl, M.; Rauch, W. Automatic generation of water distribution systems based on GIS data. *Environ. Model. Softw.* **2013**, *47*, 138–147. [CrossRef] [PubMed]
20. Möderl, M.; Sitzenfrei, R.; Fetz, T.; Fleischhacker, E.; Rauch, W. Systematic generation of virtual networks for water supply. *Water Resour. Res.* **2011**, *47*, W02502. [CrossRef]
21. Mair, M.; Zischg, J.; Rauch, W.; Sitzenfrei, R. Where to Find Water Pipes and Sewers?—On the Correlation of Infrastructure Networks in the Urban Environment. *Water* **2017**, *9*, 146. [CrossRef]
22. Mair, M.; Rauch, W.; Sitzenfrei, R. Improving incomplete water distribution system data. *Procedia Eng.* **2014**, *70*, 1055–1062. [CrossRef]
23. Paez, D.; Filion, Y. Generation and validation of synthetic WDS case studies using graph theory and reliability indexes. *Procedia Eng.* **2017**, *186*, 143–151. [CrossRef]
24. Momeni, A.; Chauhan, V.; Bin Mahmoud, A.; Piratla, K.R.; Safro, I. Generation of synthetic water distribution data using a multiscale generator-optimizer. *J. Pipeline Syst. Eng. Pract.* **2023**, *14*, 04022074. [CrossRef]
25. Rehm, I.S.; Friesen, J.; Pouls, K.; Busch, C.; Taubenböck, H.; Pelz, P.F. A Method for Modeling Urban Water Infrastructures Combining Geo-Referenced Data. *Water* **2021**, *13*, 2299. [CrossRef]
26. Nikolopoulos, D.; Kossieris, P.; Tsoukalas, I.; Makropoulos, C. Stress-Testing Framework for Urban Water Systems: A Source to Tap Approach for Stochastic Resilience Assessment. *Water* **2022**, *14*, 154. [CrossRef]
27. Pagano, A.; Sweetapple, C.; Farmani, R.; Giordano, R.; Butler, D. Water Distribution Networks Resilience Analysis: A Comparison between Graph Theory-Based Approaches and Global Resilience Analysis. *Water Resour. Manag.* **2019**, *33*, 2925–2940. [CrossRef]
28. Portal Oficial Del Gobierno De Puerto Rico. Infraestructuras. 2023. Available online: https://gis.pr.gov/ (accessed on 8 June 2023).
29. Moglen, R.L.; Barth, J.; Gupta, S.; Kawai, E.; Klise, K.; Leibowicz, B.D. A nexus approach to infrastructure resilience planning under uncertainty. *Reliab. Eng. Syst. Saf.* **2023**, *230*, 108931. [CrossRef]
30. Jones, C.B.; Bresloff, C.J.; Darbali-Zamora, R.; Lave, M.S.; Bezares, E.E.A. Geospatial Assessment Methodology to Estimate Power Line Restoration Access Vulnerabilities After a Hurricane in Puerto Rico. *IEEE Open Access J. Power Energy* **2022**, *9*, 298–307. [CrossRef]
31. Azad, S.; Ghandehari, M. A Study on the Association of Socioeconomic and Physical Cofactors Contributing to Power Restoration After Hurricane Maria. *IEEE Access* **2021**, *9*, 98654–98664. [CrossRef]
32. Jordahl, K.; den Bossche, J.V.; Fleischmann, M.; Wasserman, J.; McBride, J.; Gerard, J.; Tratner, J.; Perry, M.; Badaracco, A.G.; Farmer, C.; et al. *geopandas/geopandas*, v0.8.1; Zenedo: Geneva, Switzerland, 2020. [CrossRef]
33. Boeing, G. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Comput. Environ. Urban Syst.* **2017**, *65*, 126–139. [CrossRef]
34. OpenStreetMap Contributors. Planet Dump Retrieved from https://planet.osm.org. 2017. Available online: https://www.openstreetmap.org (accessed on 1 November 2022).
35. U.S. Geological Survey. 3D Elevation Program 1/3 Arc Second Digital Elevation Model. 2023. Available online: https://apps.nationalmap.gov/downloader/ (accessed on 1 February 2023).
36. OpenStreetMap Contributors. HOTOSM Puerto Rico Buildings (OpenStreetMap Export). 2020. Available online: https://data.humdata.org/dataset/hotosm_pri_buildings (accessed on 1 November 2022).
37. Staff, T.S. Mayagüez interim mayor looks to accelerate repair of failed water supply pipe. *The San Juan Daily Star*, 23 April 2023, p. 5.
38. Staff, T.S. Mayagüez declares state of emergency over drinking water shortage. *The San Juan Daily Star*, 5 January 2024.
39. Yazdani, A.; Otoo, R.A.; Jeffrey, P. Resilience enhancing expansion strategies for water distribution systems: A network theory approach. *Environ. Model. Softw.* **2011**, *26*, 1574–1582. [CrossRef]
40. Kabir, G.; Tesfamariam, S.; Hemsing, J.; Sadiq, R. Handling incomplete and missing data in water network database using imputation methods. *Sustain. Resilient Infrastruct.* **2019**, *5*, 365–377. [CrossRef]
41. Shi, F.; Liu, Z.; Li, E. Prediction of Pipe Performance with Ensemble Machine Learning Based Approaches. In Proceedings of the 2017 International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC), Shanghai, China, 16–18 August 2017; pp. 408–414. [CrossRef]
42. Sitzenfrei, R.; Mair, M.; Diao, K.; Rauch, W. Assessing Model Structure Uncertainties in Water Distribution Models. In Proceedings of the World Environmental and Water Resources Congress 2014, Portland, OR, USA, 1–5 June 2014; pp. 515–524. [CrossRef]
43. Hagberg, A.; Swart, P.; S Chult, D. *Exploring Network Structure, Dynamics, and Function Using NetworkX*; Technical Report; Los Alamos National Lab. (LANL): Los Alamos, NM, USA, 2008.