Neural Network Approximation of Graph Fourier Transform for Sparse Sampling of Networked Dynamics

ALESSIO PAGANI, The Alan Turing Institute, UK

ZHUANGKUN WEI, School of Engineering, University of Warwick, UK

RICARDO SILVA, The Alan Turing Institute and University College London, UK

WEISI GUO*, School of Aerospace, Transport and Manufacturing, Cranfield University, UK

Infrastructure monitoring is critical for safe operations and sustainability. Like many networked systems, water distribution networks (WDNs) exhibit both graph topological structure and complex embedded flow dynamics. The resulting networked cascade dynamics are difficult to predict without extensive sensor data. However, ubiquitous sensor monitoring in underground situations is expensive and a key challenge is to infer the contaminant dynamics from partial sparse monitoring data. Existing approaches use multi-objective optimisation to find the minimum set of essential monitoring points, but lack performance guarantees and a theoretical framework. Here, we first develop a novel Graph Fourier Transform (GFT) operator to compress networked contamination dynamics to identify the essential principle data collection points with inference performance guarantees. As such, the GFT approach provides the theoretical sampling bound. We then achieve under-sampling performance by building auto-encoder (AE) neural networks (NN) to generalize the GFT sampling process and under-sample further from the initial sampling set, allowing a very small set of data points to largely reconstruct the contamination dynamics over real and artificial WDNs. Various sources of the contamination are tested and we obtain high accuracy reconstruction using around 5-10% of the network nodes for known contaminant sources, and 50%-75% for unknown source cases, which although larger than that of the schemes for contaminant detection and source identifications, is smaller than the current sampling schemes for contaminant data recovery. This general approach of compression and under-sampled recovery via NN can be applied to a wide range of networked infrastructures to enable efficient data sampling for digital twins.

CCS Concepts: • **Computer systems organization** → *Embedded and cyber-physical systems*; • **Networks** → *Cyber-physical networks*; **Network dynamics**.

Additional Key Words and Phrases: sampling theory, Graph Fourier Transform, neural networks.

1 INTRODUCTION

Contamination in drinking water supply arises from natural disasters [1, 2], industrial crime, and terrorism [3, 4]. Together, they pose serious risks to the safety and integrity of Water Distribution Networks (WDNs).

*Corresponding author, email: weisi.guo@cranfield.ac.uk

Authors' addresses: Alessio Pagani, apagani@turing.ac.uk, The Alan Turing Institute, London, UK; Zhuangkun Wei, School of Engineering, University of Warwick, Coventry, CV4 7AL, UK, zhuangkun.wei@warwick.ac.uk; Ricardo Silva, The Alan Turing Institute and University College London, London, UK; Weisi Guo, School of Aerospace, Transport and Manufacturing, Cranfield University, Milton Keynes, UK.

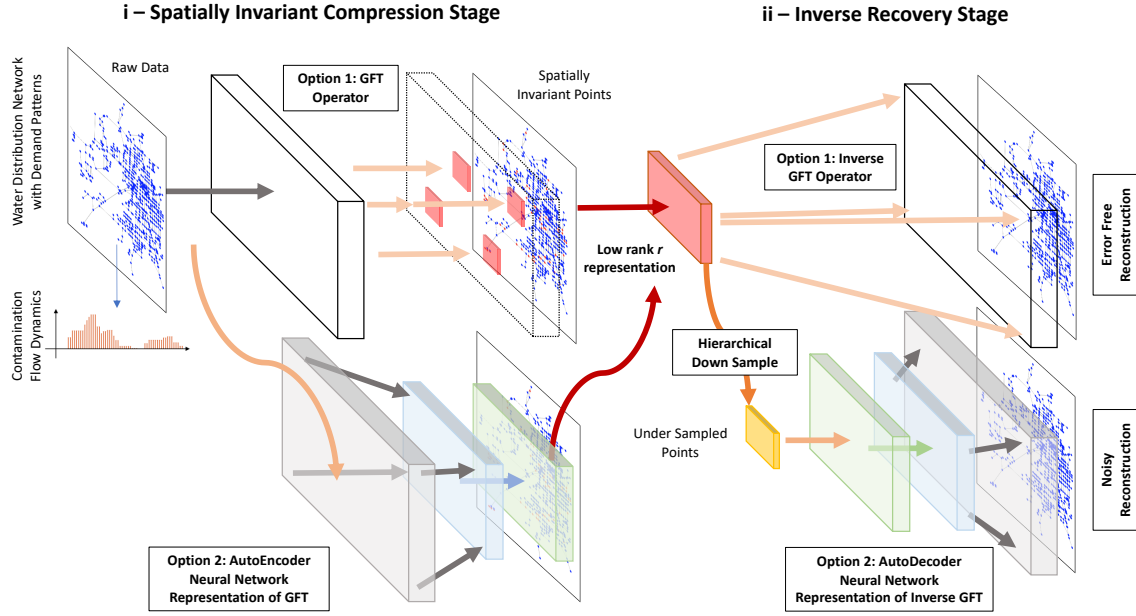**i – Spatially Invariant Compression Stage**

ii – Inverse Recovery Stage

Fig. 1. Compressing dynamic contamination data on a water distribution network to spatially invariant monitoring points and recovering the dynamics: (a) full GFT or NN compression and (b) recovery by either inverse-GFT or NN with hierarchical down-sampling option.

Contamination spreading in WDNs are governed by both Navier-Stokes dynamics and the topological structure of the network. Heterogeneous elements (e.g. pumps, filters, reservoirs), feedback loops, and the vast size of the network (100,000s of nodes) make the spread process difficult to predict without large-scale simulation and wide-spread data monitoring.

Using sparse sampling data, reconstructing the contaminant dynamics over a vast network could enable timely interventions and help to save lives. The simplest way to reconstruct the complete contaminant data would be to install probes in each junction and monitor various dynamic states. However, this is often not possible because of the high installation and maintenance cost and the difficulty in accessing underground pipes retrospectively. This raises the necessity of optimized sensor placement, detecting chemical intrusions and recovering the contamination spread in the shortest time and with the highest accuracy possible.

1.1 State of the Art

From the prospective of current sensor placement techniques, they can be categorized into: (i) warning system designs using essential key performance indicators (KPIs) [5, 6] (e.g., low time to detect chemical intrusion, low amount of contaminated water consumed or population affected), and (ii) reconstructing and predicting the spread of dynamics. The first approach leverages mostly on the rule-based (e.g., single-objective or multi-objective) optimisations, while the second is resorted to the graph sampling theory or data-driven approaches, e.g., the compressed sensing (CS). We will examine them one-by-one and this study focus on the second approach.

1.1.1 Numerical Optimization Approaches. Rule based multi-objective optimisation considers a number of performance metrics and factors related to both WDN dynamics, as well as accessibility and complexity aspects of the

cyber-physical interface [7]. For example, Berry et al. [8] tackled the problem of sensor placement formulation by optimizing the number of sensors that minimize the expected fraction of population at risk. The optimisation approaches include mixed-integer program (MIP), randomized contamination matrix [9], and genetic algorithms [10]. Other sensor placement approaches [11–14] revolve around multi-objective optimization frameworks (e.g., for source identification [13], and detecting deliberate contamination events [12]), this gives the capability to reduce the dimension of the network through a sensitivity-informed analysis [15], and thereby reduce the sampling points by outstanding KPIs for the designs of contamination warning system [11, 16]. When it comes to the dynamic reconstruction challenge, however, they may suffer from the lack of explicit relational knowledge between the topological structure and the underlying dynamics with the optimal sampling points. Furthermore, the aforementioned solutions become less feasible for large-scale networks, especially for multiple or diverse contamination dynamics. Whilst computational improvements for multi-objective optimisation in WDNs have been developed [17, 18], such as a progressive genetic algorithm (PGA), they offer less performance guarantees and theoretical insights.

1.1.2   Graph-Based Analytical Approaches. Graph Spectral Techniques (GSTs) that identify the most influential points on the base of the topological structure of the networks (e.g. via the Laplacian operator [19–21]) offer theoretical insight between the topological structure of the network and the key monitoring points. These approaches significantly reduce the computation complexity by removing the need of hydraulic simulations [22, 23], but tend to assume a homogeneous network (e.g. pumps, reservoirs, and junctions are treated equally). The main assumption in topological analysis that doesn't consider the underlying fluid dynamics is that it is assumed that the topology dominates. As such, it is important to create an approach that considers both the complex network topology and the contamination signals. The challenge with WDNs is that the underlying Navier-Stokes dynamics with dynamic Reynolds numbers is high dimensional and highly non-linear [24]. As such, an analysis of the optimal sampling points as a function of both the network topology and the dynamic equations is non-trivial.

1.1.3   Data-Driven Approaches. One approach that considers the data-structure is compressed sensing (CS) [25–27], which compresses the data by transforming them into a sparse domain. However, the main challenge lies in the unknown of positions of such sparse non-zeros elements in the transformed data, which will inevitably lead to an approximately $(N + K - r) \times r/K$ nodes for monitoring (for a data matrix $\mathbf{X}$ with N nodes, K time-step and $r = rank(\mathbf{X})$) [28, 29]. This is not to mention that most of the CS approaches do not guarantee the unchanged nodes for sensor deployment.

To further reduce the number of sampling nodes, our previous work in [29, 30] proposed a data-driven GFT sampling method, which is able to characterize the data matrix into an r-bandlimited space, and thereby ensures the recovery accuracy with the monitoring of only r orthogonal nodes. However, the overlooks of the unknown and latent transitions among states limit the further reduction of the number of monitoring nodes. This raises the necessity of the understanding of such relations underlying the data, and we do so by the run of a Neural Network.

1.1.4   Graph Neural Networks. Graph neural networks (GNNs) were initially proposed by Gori et al. [31] and Scarselli et al. [32]. These early studies fall into the category of recurrent graph neural networks (RecGNNs). Convolutional neural networks (ConvNNs) were then introduced by Bruna et al. [33] and gained popularity [34, 35]. In the last couple of years, several works focused on using machine learning techniques with arbitrarily structured graph data [36, 37]. A comprehensive survey of NN applied to graphs can be found in [38], including the most recent techniques based on Graph Autoencoders (GAEs): GAEs map nodes into a latent feature space and decode graph information from latent representations. They can be used to learn network embedding or generate new graphs.
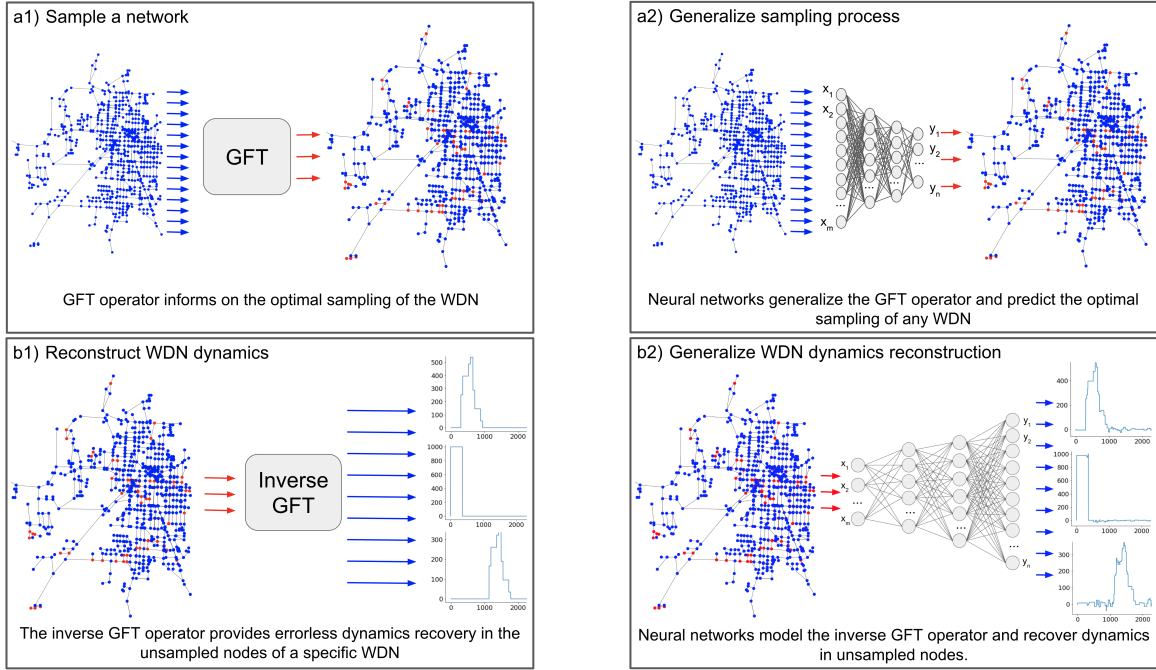
Fig. 2. Sampling process. The GFT informs on the optimal number of nodes required to reconstruct the dynamics without errors (a1). A Neural Network (NN) is used to generalize the process and reduce the number of sampled nodes (a2). The inverse GFT operator is used to reconstruct the dynamics in all the nodes of a specific network (b1). NN can be used to model the inverse GFT operator and reconstruct the dynamics for every given input (b2). Using NN to sub-sample the GFT ensemble or reconstruct unseen signals can introduce errors in the dynamics reconstruction, with a trade-off between size of the final ensemble and accuracy.

In our work, we improve the state of the art exploiting a GAE-based technique to optimally sample the network considering the graph- and dynamics- domain, with the aim to reconstruct the dynamics in all the nodes of the graph. NN alone, in fact, cannot inform us which nodes are optimal for sampling, nor how many nodes are needed. This is why we are using the GFT to inform and drive the NN, which is novel.

Neural Networks for WDN applications. Neural Networks (NN) are increasingly being used in WDN applications, especially for water supply issues and predictions relative to chemical disinfectants and contaminants. A synopsis of NN methods, including the design and operation of WDNs, is provided in review article [39]. Proposals for a neural network in the assessment of pressure losses in water pipes are proposed in [40, 41]. Cuesta Cordoba et al. [42] used an NN to predict chlorine decay using historical data. Similarly, Andrade et al. [43], estimated the disinfectant concentration at the relevant nodes using NN with the aim of improving WDNs design. These approaches, however, either work well only on a pre-selected subset of nodes considered important, or require a high number of sensors to work on the whole water infrastructure. Moreover, they are not designed for working in real time, excluding the capability to promptly predict chemical intrusion and spreading.

1.2 Contribution

We aim to use sparse data to recover the contaminant spreading dynamics in the WDNs. In this work, we first use GFT to derive the operator that reveals the minimum number of nodes needed to sample the WDN and recover its dynamics with minimum error (see Fig. 1). This approach was outlined by our recent work [29, 44] (Fig. 2(a)), which minimized sensor deployment number at the penalty of relying on the inverse GFT operator to recover the dynamics, only with the full set of sampled data. Here we expand on this by under-sampling the GFT derived set of minimum nodes to create noisy measurements (Fig. 2(b)). By developing a neural network (NN), we can sample from an extremely sparse set that can recover the essential noisy contamination dynamic trend. This is useful because often we are interested in whether a contamination has exceeded a threshold, but not necessarily its full dynamic response.

There is a natural trade-off between reconstruction accuracy and size of the sampled nodes ensemble used for the training, but we show that NNs can significantly reduce the number of required nodes (compared to the pure GFT approach) maintaining a high accuracy. The final nodes ensemble is the the minimal set of nodes that has to be equipped with probes and sensors for monitoring. We reconstruct the chemical spread dynamics using two scenarios: known potential contamination source (e.g., factories or plants, maintenance works, known dangerous areas) and unknown contamination source (e.g., terrorist attacks).

The final outcome of this work is a general framework able to optimally sample a network with complex flow dynamics, e.g. partial differential equations (PDEs) with dynamic parameters and feedback loops, and then reconstruct the flow dynamics in all the network nodes using only the sparse partial dynamics at optimal points. The framework is composed of two main parts: (1) an encoder (via GFT or NN) - that compress the network considering its topology and dynamics, and a decoder (via inverse-GFT or NN) - that reconstructs the flow dynamics. We achieved (i) the reduction of the number of sensors in the face of known contaminant sources, and (ii) a generalized NN method for node selection and signal recovery when any prior information (e.g., sparsity and graph bandlimitedness) is unavailable in the case of unknown contaminant source scenarios. In this view, the usages of 5%-10% nodes for known cases and 50%-75% for general cases are attractive, although larger than those schemes aiming for only contaminant detection and warning WDN systems.

The rest of paper is structured as follows. In Section II, we describe the initial sampling process based on GFT and the subsequent sampling reductions that is used as input for the NN models. We then introduce the NN architecture and the WDN simulations. In Section III, we show the results of the dynamics reconstructions using the NN models. In Section IV, we conclude the paper and discuss the potential future areas of research.

2 METHODS

This work aims to deploy two NNs to generalize the GFT sampling and recovery processes, for WDN with contaminant signal over time. In essence, the GFT reversibly maps the sampling node set to the band-limited signal space in a **linear** manner, making the minimum number of sensors equals to the dimension of band-limited signal derived from one contaminant injection-specific scenario. As more injection-specific cases considered (dimension of signal space increases), the GFT method will lead to either a poor recovery performance or exceedingly large number of sensor deployment. To overcome this, the NNs that generalize the GFT sampling and recovery will obtain a **high-dimensional** and **nonlinear** GFT manner, creating a nonlinear and reversibly mapping between the sampling node set and the high-dimensional signal-space derived from the general injection scenarios. This therefore enables the sampling node compression in the injection-specific scenario, and provides a feasible sampling and recovery paradigm for general cases.

We first use GFT to develop a linear operator that transforms the injection-specific band-limited WDN contamination spread dynamic into a set of sampling nodes for guaranteed reconstruction performance. Then, we design and train the nonlinear NNs (an encoder NN and a decoder NN in Fig. 2) from various GFT datasets,

.

for the purpose to generalize the GFT processes, and to (i) reduce the sampling nodes when contaminant sources are known, and (ii) derive an efficient sampling nodes and recovery method for all possible and unknown contamination scenarios.

2.1 GFT Sampling Process

Graph sampling theory over complex network [1] aim at sampling and recovering the time-varying networked signals, denoted as $\mathbf{X} \in \mathbb{R}^{N \times K}$ that are $(r < N)$-bandlimited to a given GFT operator, denoted as $\mathbf{F}^{-1}$. Here, we consider a *static topology* with *time-varying dynamic signals*. The N rows of $\mathbf{X}$ present the time-varying signals on N nodes, and $K > N$ denotes the total number of time-indices. We say $\mathbf{X}$ is r-bandlimited to $\mathbf{F}^{-1}$, if and only if its GFT response $\tilde{\mathbf{X}} = \mathbf{F}^{-1} \cdot \mathbf{X}$ has only r nonzero rows. Denote $\mathcal{R}$ is the set of subscripts of the nonzero rows in $\tilde{\mathbf{X}}$, and $\mathcal{V} = \{1, \cdots, N\}$. Then, we say there exists a subset $\mathcal{S} \subset \mathcal{V}$ such that:

$$\mathbf{X} = \mathbf{F}_{\mathcal{V}\mathcal{R}} \cdot (\mathbf{F}_{\mathcal{S}\mathcal{R}}^T \cdot \mathbf{F}_{\mathcal{S}\mathcal{R}})^{-1} \cdot \mathbf{F}_{\mathcal{S}\mathcal{R}}^T \cdot \mathbf{X}_{\mathcal{S}\mathcal{K}}, \tag{1}$$

if and only if:

$$rank(\mathbf{F}_{\mathcal{S}\mathcal{R}}) = |\mathcal{R}| = r. \tag{2}$$

In Eqs. (1)-(2), $\mathbf{X}_{\mathcal{S}\mathcal{K}}$ denotes the sample of $\mathbf{X}$ from nodes that belongs to $\mathcal{S}$. $\mathbf{F}_{\mathcal{S}\mathcal{R}}$ denotes the selection of the matrix $\mathbf{F}$ with row indices from set $\mathcal{S}$, and column indices from set $\mathcal{R}$. The sampling and recovering processes can be pursued after (I) designing the GFT operator, and (II) the selection of $\mathcal{S}$ satisfying Eq. (2).

2.1.1 GFT Operator. The derivation of the GFT operator $\mathbf{F}^{-1}$ is resort to the QR factorization, due to its ability to transform a r-rank data matrix to its bandlimited (upper-triangular with r nonzero rows) response matrix. The inverse of the transforming matrix is therefore the GFT operator, which embraces the information of both data-space, and network topology that is reflected by the spread of data-flow via underlying pipes. To be specific, as we derive the maximally linearly independent columns of $\mathbf{X}$, denoted as $\mathbf{X}_{\mathcal{V}M} = [\mathbf{x}_{m_1}, \cdots, \mathbf{x}_{m_r}]$, the GFT operator can be computed as:

$$\mathbf{F}^{-1} = \mathbf{Q}^{-1}, \tag{3}$$

where $\mathbf{X}_{\mathcal{V}M} = \mathbf{Q} \cdot \mathbf{R}$. This $\mathbf{F}^{-1}$ ensures the r-bandlimited property of $\mathbf{X}$, since:

$$\begin{aligned}
\tilde{\mathbf{X}} &= \mathbf{F}^{-1} \cdot \mathbf{X}, \\
&\overset{(a)}{=} \mathbf{F}^{-1} \cdot [\mathbf{X}_{\mathcal{V}M}, \ \mathbf{X}_{\mathcal{V}M} \cdot \mathbf{\Pi}], \\
&\overset{(b)}{=} [\mathbf{R}, \ \mathbf{R} \cdot \mathbf{\Pi}].
\end{aligned} \tag{4}$$

In Eq. (4), (a) holds for that each column of $\mathbf{X}$ can be expressed by the columns from $\mathbf{X}_{\mathcal{V}M}$ multiplied with an $r \times (K - r)$ matrix $\mathbf{\Pi}$, since $rank(\mathbf{X}_{\mathcal{V}M}) = rank(\mathbf{X}) = r$. (b) indicates that only the first r rows of $\tilde{\mathbf{X}}$ are non-zero, as $\mathbf{R}$ is the upper triangular matrix with $rank(\mathbf{R}) = r$.

2.1.2 Signal Recovery. After the computation of the GFT operator in Eq. (3), the guarantee of complete dynamic recovery is to select $\mathcal{S}$ that satisfies Eq. (2). One can refer to [45–55] for details. Here, in order to achieve a robust sampling scheme on nodes, we select $\mathcal{S}$ that maximizes the minimum singular of $\mathbf{F}_{\mathcal{S}\mathcal{R}}$, i.e.,

$$\mathcal{S}_{\text{opt}} = \underset{\mathcal{S} \subset \mathcal{V}}{\text{argmax}} \ \sigma_{\min}(\mathbf{F}_{\mathcal{S}\mathcal{R}}), \tag{5}$$

where $\sigma_{\min}(\cdot)$ denotes the smallest singular value. As such, the importance of the nodes in $\mathcal{S}$ can be ranked with the descending order of the singulars.

---

[1]a complex network is a graph (network) with non-trivial topological features

As discussed, the initial sampling is conducted using the GFT analysis that exploits the spatial and temporal dependency induced low-rank property to optimally sample junction nodes in WDNs. Using GFT, it is possible to fully recover network dynamics of a specific injection scenario using a subset of data sampled at the identified nodes. The identified nodes are used as initial subset, hereafter called the **GFT dataset**. Note that the GFT dataset is different for each possible source of contamination (injection location). However all the scenarios are used to train the NNs network, therefore the final solution is able to identify the best common subset of nodes to reconstruct the water dynamics in each possible scenario. We next elaborate the encoder and decoder NNs to identify this common subset of nodes and to reconstruct the dynamics minimizing the error.

2.2 Deep Neural Network based GFT Sampling

We deploy two NNs (the encoder NN and decoder NN) that aim to mimic and generalize the GFT sampling and GFT data reconstruction processes. Two following scenarios are considered.

Injection-specific case. accounts for the known locations (nodes) of the main sources of contamination, as one can know exactly the locations of factories that may pollute the area given a specific region (e.g., city). In such a case, it is true that a sensor placement via the downstream the sources is able to recover the contaminant dynamic on each node. However, this requires the deployment of sensors on every downstream nodes. Our NN approach here helps to reduce the number of sensors while reconstructing the dynamics with very low error.

General case. refers to as the scenario with unknown contaminant sources. In such a case, our NN approach aims to generalize the sampling node selection and contaminant data reconstruction procedures.

2.2.1 Encoder Neural Network Architecture. This process of encoder NN here is similar to encoders used in autoencoders (AE) [56], with the difference that our final compression has to be related with physical nodes in the network. As such, a feed-forward neural network is used for the encoder NN, where both the input and output layers are equal to the number of nodes in the WDN, and the activation function is the Rectified Linear Unit (ReLU). The purpose of the NN is to classify the network nodes: each node is associated with a neuron in the input and output layer. While the input layer is fed with node dynamics time series ($\mathbf{X}_i$), the output layer classifies the nodes as important (1) or not important (0) ($\mathbf{y}_i$). For a given NN the input layer is therefore defined as:

$$\mathbf{X} = [\mathbf{X}_1, ..., \mathbf{X}_N]$$

while the output layer is defined as:

$$\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_N]$$

where N is the number of junctions in network. The Encoder contains two or three hidden layers. It is noteworthy that for different WDN topology, the NN may change its network topology (e.g., using an additional hidden layers for some complex WDN, or increasing/decreasing the number of input and output neurons when WDN has more/less nodes). Nevertheless, compared with the data-driven GFT that requires to generate different operators and sampling node sets for different injection-specific scenarios, the NN, learnt from all possible injection scenarios, is able to identify the best common subset. This attributes to its ability to create nonlinear and reversible mapping between the compressed sampling node set and the high-dimensional signal-space induced by general injection cases.

2.2.2 Training for Encoder Neural Network. The encoder NN is trained via the gradient descent algorithm, using the derived GFT datasets. For the injection-specific scenarios, the input layer is with the simulated contaminant matrix given known contamination sources, and the output layer is assigned in accordance with the corresponding data-driven GFT determined sampling nodes. For the general cases, simulated contaminant spreading data matrices

.

with unknown contamination sources and their data-driven GFT determined sampling nodes are used for the training of the encoder NN.

It is noteworthy that the number of GFT datasets increases with the number of injection points in the network, making the training process become computationally very intensive. In this paper, we filter the GFT datasets by removing those that are subsets of other GFT datasets. For example one GFT dataset may recommend to monitor 3 nodes A, B and C and another GFT dataset may recommend to monitor 2 nodes A and B. In such a case, the second scenario is removed from the training dataset used in the neural network training. This optimization helps to maintain a balanced dataset and improves the training time.

As such, when a large number of GFT datasets is provided, the NN Encoder is able to learn to generalize the selection process and to detect the important nodes for any possible injection in a WDN. This process is based only on the data relative to the spreading of pollution over time, the WDN topology is therefore implicitly learnt by the neural network. We aim to emphasize here that the neural network has no explicit information regarding the topology of the network (e.g., the adjacency matrix). Instead, it automatically learns it using the networked contaminant spreading data. For example, when the pollution spreads in a short amount of time from a node to another, it suggests that there is a connection between the two nodes.

2.2.3 Decoder Neural Network Architecture. We train the decoder NN for the reconstruction of the chemical spread dynamics for each sampling dataset generated. The NN is feed-forward deep neural networks, with hidden layers and an increasing number of neurons equipped by ReLU activation function. Here, two hidden layers are used, the first with 1/3 of the neurons in the output layer and the second with 2/3 of the neurons in the output layer. As is extensively analyzed and evaluated, two layers are generally enough to reconstruct the inverse GFT and reconstruct the dynamics in all the nodes of the WDN. It is worth mentioning that the topology of the WDN is implicit in the training dataset, hence learned by the NNs without need of NN architectures specifically designed for graph learning.

A sensor is installed in each node that belongs to the important nodes determined by the encoder NN ($\mathbf{x}_i$). Then the NN is fed with the sensor readings over time. Hence, the input layer is a set of neurons, one for each probe installed in the WDN. The output layer is the estimated concentration of chemical in all the junctions ($\mathbf{y}_i$) of the WDN. For a given NN the input layer is therefore defined as:

$$\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_m]$$

where m is the size of the sampling subset. The output layer is defined as:

$$\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_N]$$

yielding contaminant time-series of total N nodes in network. As such, the decoder NN is to recover the contaminant time-series on each node from the feeding of sampling readings, which suggests its ability to recover and detect a contaminant in real-time manner.

2.2.4 Training for Decoder Neural Network. We then elaborate the training data for decoder NN. Here, the train algorithm is the gradient descend. For injection-specific scenarios (when the chemical source is known), the relative data-driven GFT determined sampling nodes can be used as initial sampling subset. In a GFT dataset the nodes are ranked in order of importance for the reconstruction of the signal. For this reason, the **injection-specific sampling datasets** are generated removing from the GFT dataset, one by one, the nodes with the lower rank. Each newly created sampling dataset is used to train the decoder NN and the model performance is evaluated. The reduction process is repeated until the reconstruction accuracy drops below a given threshold.

When the injection sources are unknown, we use two techniques to generate the training dataset, one based on the nodes frequency in the GFT datasets, the other based on their importance (rank) in each GFT dataset.

GFT Frequent Nodes Dataset. this approach revolves around the selection of the more frequent nodes in the GFT datasets of the different injection scenarios. The **GFT frequent nodes dataset (GFT-F)** is created by counting the times each node appears in the different GFT datasets and selecting only the nodes which appear more than a given threshold. Different thresholds are discussed in the results.

GFT Important Nodes Dataset. this approach concerns the selection of the most important nodes in each injection scenario. In order to create a **GFT important nodes dataset (GFT-I)**, we consider only the n most important nodes of each GFT dataset. For example, for $n = 1$, we select the node with the highest ranking in each GFT dataset (i.e., each injection scenario). This could lead to big sampling datasets even with a small n, however the most important nodes are usually shared among different GFT datasets, for this reason the final GFT-I datasets are significantly smaller than the number of different injection scenarios.

2.2.5 Analysis of NN based GFT Sampling. Here, our proposed NN based sampling and recovery method relies on the spatial and temporal dependencies (i.e., the graph bandlimitedness) of the networked contaminant dynamics, which are mainly determined by the network topology and the underlying flow dynamic time-evolution. In this view, trained by all possible contaminant injection locations and the regular day-night water-demand cycle in a fixed WDN topology, the encoder and decoder NNs are able to capture the graph bandlimitedness governed by the underlying flow dynamic time-evolution, and therefore capable of identifying the sampling locations and successfully recovering the complete networked contaminant dynamics.

2.3 Case Study: Chemical Injection Simulations

The simulations are executed using WNTR (Water Network Tool for Resilience) [57]. WNTR is an EPANET [58] compatible Python package designed to simulate and analyse resilience of WDNs, it performs extended-period simulation of hydraulic and water-quality behaviour within pressurized pipe networks. This package also supports the simulation of spatially and temporally varying water demand, constant or variable speed pumps, and the minor head losses for bends and fittings. The modelling provides information such as flows in pipes, pressures at junctions, propagation of a contaminant, chlorine concentration, water age, and even alternative scenario analysis.

In our case-study, the EPANET2 simulated WDN system is configured as $N = 102$ nodes (see Fig. 2), including 100 junctions and 2 reservoirs. A contaminant is injected at a random starting time over 24 hours, for a predefined amount of time (from the beginning time-step $k = 0$ to $k = 100$), in a specific junction of the network. The water demands of junctions are assigned ranging from 77.2 L/s (in the morning) to 7.6 L/s (at night time). The contaminant spreads in the network following the water dynamics and the water demand and it is finally expelled by demand junctions. The discrete contaminant data of each node are collected every 30 seconds, for a period of 24 hours, and the total time-steps are $K = 2880$. One contaminant dataset is the matrix of size $N \times K$, and is therefore composed of the discrete contaminant data on N nodes with K time-step.

The data-driven GFT sampling method selects the sampling node for each data matrix, which will then be used for the training processes of the encoder and decoder NNs. For the injection-specific scenarios, we assume a known injection point and the contaminant datasets for training and testing the NNs are with this known injection point. For general cases, the contaminant datasets are generated with all possible injection points, and are used for the training and the test of the NNs.

For both the cases, the NNs uses 10000 datasets for training and 10000 corresponding datasets for testing. The training time is around 30 minutes to 4 hours.

GFT dataset.

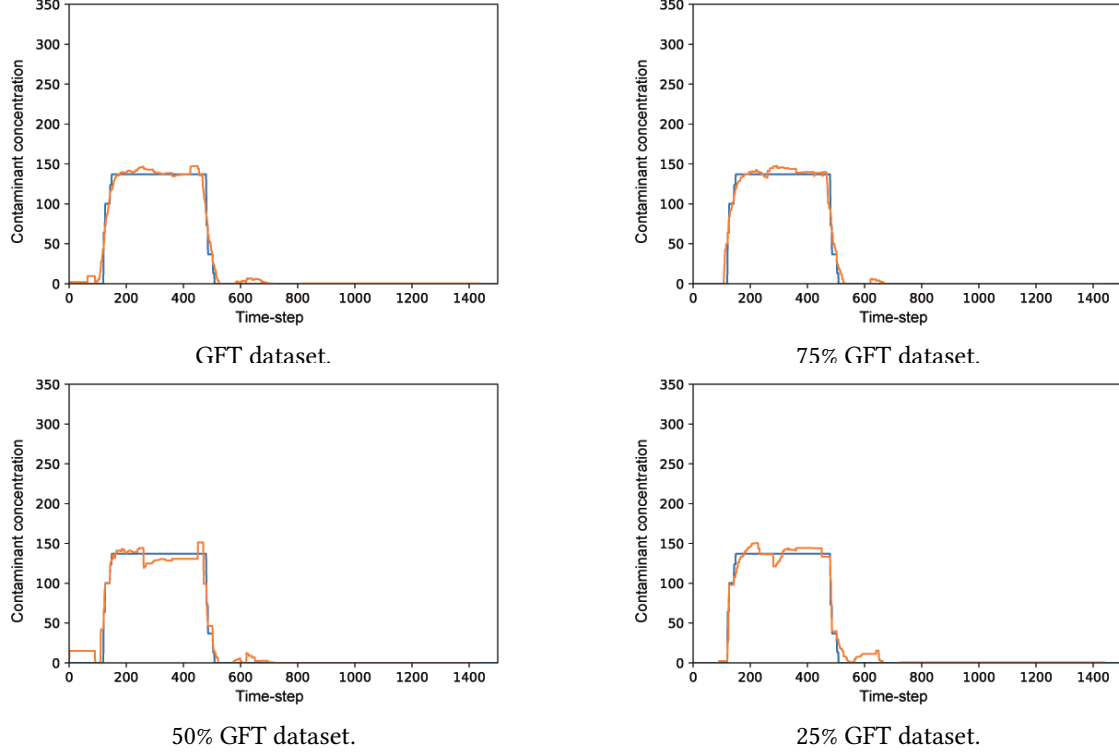75% GFT dataset.

50% GFT dataset.

25% GFT dataset.

Fig. 3. Example 1 - Reconstruction of the dynamics using different portions of the GFT dataset. Original signal in blue, reconstructed signal in orange.

3 RESULTS AND DISCUSSION

In this section we discuss the sampling and recovery results obtained using NN encoder and NN decoder, for injection-specific scenarios, and general cases. One illustration of the WDN layout, the monitored nodes and the reconstructed contaminant signals using NNs is provided in Fig. 2 a2) and b2), where a2) gives the sampling node selection by the encoder NN, and b2) shows the signal reconstruction via the sampling nodes using decoder NN. The results are compared with two sampling baselines: random sampling and Laplacian sampling. In the **random sampling** a predefined percentage of WDN nodes is randomly selected. In the **Laplacian sampling**, the WDN nodes are selected (as $\mathcal{S}_L \subset \{1, \cdots, N\}$) by the r smallest of eigenvalues, i.e., $rank(\mathbf{U}_{\mathcal{S}_L \mathcal{R}}) = |\mathcal{R}|$, where $\mathbf{U}$ is composed by the eigenvectors of the Laplacian matrix $\mathbf{L}$, ranked by the ascending eigenvalues, and $\mathcal{R} = \{1, \cdots, r\}$.

3.1 Dynamics Reconstruction

When the source is known, the GFT dataset can be hugely reduced without significant loss of accuracy. In fact, the NN is able to reconstruct the signal using a few dynamics in specific nodes. Accurate reconstructions are obtained, on average, using 20% of the GFT dataset (i.e., around 5-10% of all the WDN nodes has to be monitored, depending on the injection point). Two examples of signals reconstructed using NN models are shown in Fig. 3 and Fig. 4: the concentration over time of a chemical component in two (not-monitored) junctions are reconstructed (in
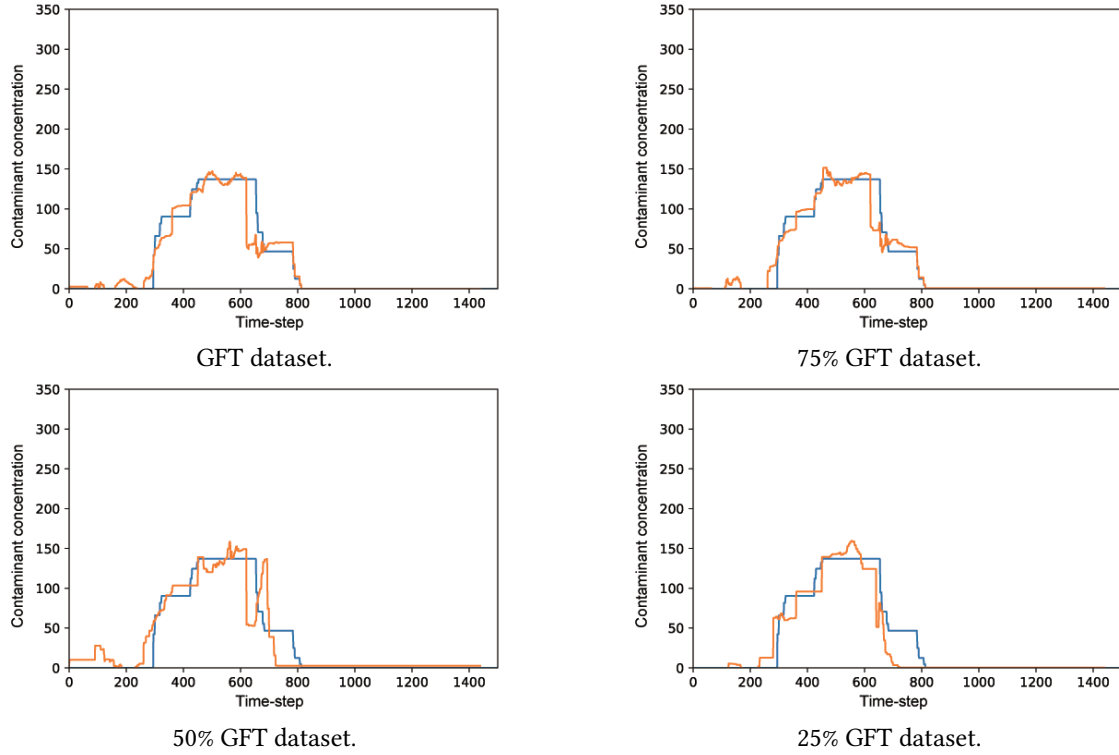
Fig. 4. Example 2 - Reconstruction of the dynamics using different portions of the GFT dataset. Original signal in blue, reconstructed signal in orange.

blue the original signal, in orange the reconstructed one). Besides, Figs. 3-4 show the immediate contaminant detection. This is because the decoder NN is able to recover the networked contaminant data from samples in a real-time manner, therefore indicating the immediate contaminant detection.

When the chemical source is unknown, or a general platform able to reconstruct the dynamics in the whole network is required, the injection specific approach cannot be used, instead one of the general approaches is required. While a source specific approach requires, on average, 5-10% of the WDN nodes to reconstruct the dynamics, the general approaches require several more nodes: among the tested sampling techniques, the best one is the selection of the most frequent nodes in the GFT datasets (GFT-F approach): given a GFT-F dataset, the NN requires, on average, to monitor 50-55% of the WDN nodes (junctions) for near error-less dynamic reconstruction. The same accuracy is reached using the GFT-I dataset and monitoring 70-75% of the WDN nodes. More details on the percentage of dynamics correctly reconstituted for a given percentage of WDN nodes is provided in Section 3.2. Both the approaches performed better than the baselines: the signal reconstruction requires respectively 75-80% of the WDN nodes using the Laplacian ranking and 80-90% of the WDN nodes using random sampling.

Fig. 5 shows the recovery error of the reconstructed dynamics, using 10000 datasets of different size and different sampling techniques. Here, the recovery error is measured by the root mean square error normalized by the average contaminant magnitude (referred to as normalised RMSE). Fig. 5(a) presents the average results over 10000 datasets, and Fig. 5(b) and Fig. 5(c) show the worst and best case derived from one of the general case given the unknown contaminant source. It is intuitively seen that, the higher the acceptable error in the

Average RMSE.

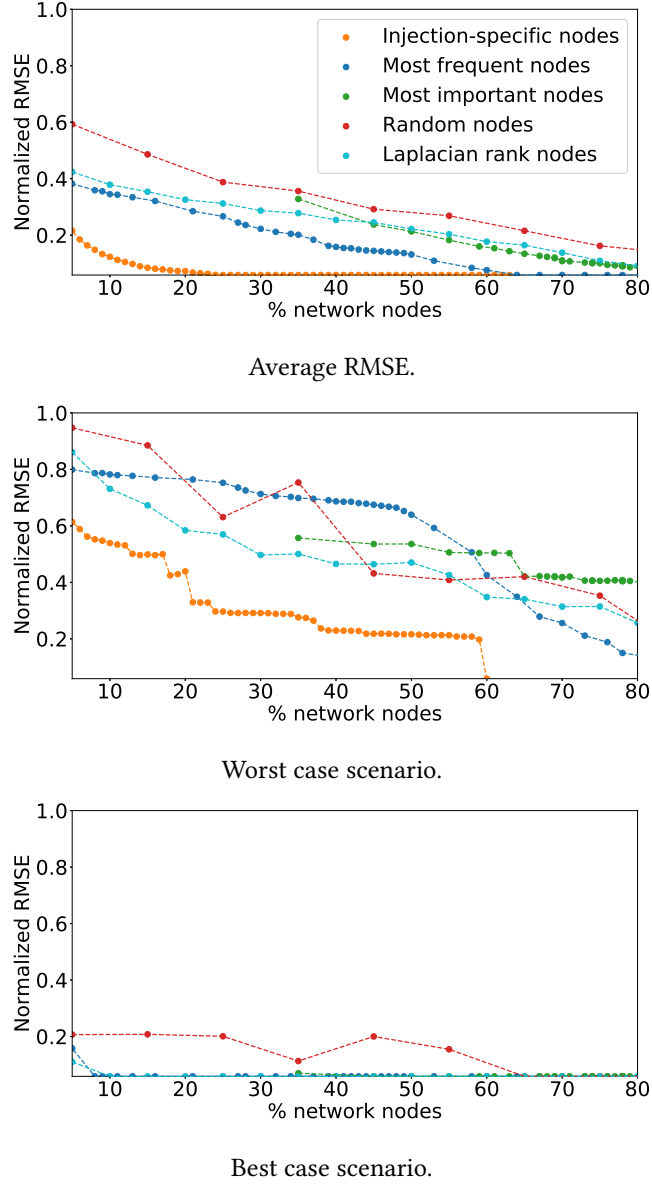Worst case scenario.

Best case scenario.

Fig. 5. Normalised RMSE of the reconstructed dynamics using different sampling approaches and different dataset size.

dynamics reconstruction, the lower the required number of nodes to be monitored. Then, the proposed NNs using GFT-F outperforms the other sampling schemes (e.g., the Laplacian schemes), demonstrating the feasibility of the proposed scheme.

High accuracy reconstruction

	75% WDN	50% WDN	30% WDN	10% WDN
GFT-F	51%	47%	44%	19%
GFT-I	40%	38%	15%	N/A
Lapl.	32%	19%	18%	10%

Medium accuracy reconstruction

	75% WDN	50% WDN	30% WDN	10% WDN
GFT-F	88%	85%	63%	26%
GFT-I	81%	55%	22%	N/A
Lapl.	72%	41%	36	15%

Low accuracy reconstruction

	75% WDN	50% WDN	30% WDN	10% WDN
GFT-F	98%	90%	70%	36%
GFT-I	92%	72%	43%	N/A
Lapl.	91%	72%	68%	37%

Table 1. Sensitivity (true positives): percentage of polluted junctions with correctly reconstructed dynamics using different sensor placement techniques: GFT frequent nodes dataset (GFT-F), GFT important nodes dataset (GFT-I) and Laplacian ranking. 3 different levels of accuracy are used.

3.2 Sensitivity and Specificity of the General Approaches

Sensitivity (also called true positive rate or probability of detection) measures the percentage of junctions with correctly reconstructed dynamics when there is contamination. **Specificity** (also called true negative rate) measures the proportion of not polluted junctions (thus with constant dynamics, equal to 0) that are correctly identified as such (thus, indirectly, the probability of false alarms). For this experiment, we classify the accuracy as high, medium and low by the normalized-RMSE $< \gamma$ with $\gamma = 0.1, 0.2, 0.5$ respectively. As such, the sensitivity and specificity with respect to the accuracy level are expressed as:

$$sensitivity(\gamma) = \frac{\text{Number of nodes with contaminant and } N-RMSE < \gamma}{\text{Number of nodes with contaminant}} \tag{6}$$

$$specificity(\gamma) = \frac{\text{Number of nodes with no contaminant and } N-RMSE < \gamma}{\text{Number of nodes with no contaminant}} \tag{7}$$

The results of this analysis are shown in Table 1 (sensitivity) and Table 2 (specificity). This study confirms that NNs using GFT frequent nodes dataset (GFT-F) outperforms the other sampling techniques. While the specificity is always high (probability of false alarms is very low), the sensitivity varies significantly. In particular, if 75% of the WDN nodes is monitored, 51% of the dynamics are correctly reconstructed with high accuracy. The percentage of correctness reaches 88% with medium accuracy and 98% with low accuracy. When 50% of the network is monitored, we can correctly reconstruct the dynamics in 47% of the nodes with high accuracy. The percentage increases if medium accuracy (85%) or low accuracy (90%) are acceptable. When a low accuracy reconstruction of the signal is sufficient, we obtain good percentage of correct dynamics reconstructed even with 30% of monitored nodes (70% of dynamics correctly reconstructed) and 10% of the WDN nodes (36% of

High accuracy reconstruction

	75% dataset	50% dataset	30% dataset	10% dataset
Freq. nodes	94%	94%	95%	65%
Imp. nodes	90%	92%	79%	N/A
Lap. nodes	91%	91%	92%	92%

Medium accuracy reconstruction

	75% dataset	50% dataset	30% dataset	10% dataset
Freq. nodes	99%	98%	97%	81%
Imp. nodes	99%	97%	88%	N/A
Lapl. nodes	99%	97%	96%	96%

Low accuracy reconstruction

	75% dataset	50% dataset	30% dataset	10% dataset
Freq. nodes	100%	99%	99%	99%
Imp. nodes	99%	98%	98%	N/A
Lap. nodes	99%	99%	98%	98%

Table 2. Specificity (true negatives): percentage of not polluted junctions correctly identified using different sensor placement techniques: GFT frequent nodes dataset (GFT-F), GFT important nodes dataset (GFT-I) and Laplacian ranking. 3 different levels of accuracy are used.

dynamics correctly reconstructed). It is noteworthy that, without the prior information of the contaminant data (e.g., sparsity and graph bandlimitedness), there is no current way to place sensors for contaminant recovery. As such, although the percentage of sampling nodes may appear high for the guarantee of contaminant data recovery (around 50%-75%), our NN based scheme is able to provide a generalized GFT sampling method for sampling node selection and signal recovery.

The sensitivity and specificity also indicate the detection likelihood, where the former embraces the probability of correct detection among all contaminated nodes, and the $1 - specificity$ indicates the probability of false-alarm (which is always small as shown by Table 2). In this view, even if some of the percentages of dynamics correctly reconstructed (sensitivity) may appear low, it is still capable of presenting a high correct detection likelihood, since low reconstruction accuracy does not seriously influence the contaminant existence detection. Besides, another goal of the WDN monitoring is to guard the most sensible parts of the network (e.g., the most sensible to attacks or accidental contamination). For this reason, even the solution that monitors 36% of the network installing sensors in 10% of it may be extremely useful, especially when keeping the sensor installation and maintenance cost low is a priority. Reconstruct the dynamics in central nodes may be sufficient to understand when there is a chemical contamination and where it is spreading. This permits to act consequently and take the required precautions (e.g., closing gates to protect part of the network, alert population). Moreover, if significantly important junctions are known (e.g., junctions close to highly populated areas, chemical plants that may accidentally discharge contamination, etc.), the described approaches can be designed to reconstruct the dynamics with high accuracy in these junctions. Another possible solution could be an hybrid approach that uses more nodes belonging to GFT datasets of potentially dangerous nodes, or that prioritize the dynamics reconstruction of sensible nodes.

For the RMSE threshold γ, it is true that the different setting affects the sensitivity and specificity, e.g., a larger γ leads to larger sensitivity and specificity, but the exact selection of γ is out of the scope of this work. Here, the aim is to show the statistical results of the contaminant recovery on each node, i.e., the ratio of nodes whose RMSE$< \gamma$, which, combined with the average RMSE provided in Fig. 5, is to demonstrate the sampling and recovery performance, and to guide the number and locations of the sensors for guaranteed recovery performance. For specific monitoring scenarios, the selection of γ may be complicated. For instance, when the sensor network life-time is of superiority [59] (given the difficulty to recharging the underground sensors), one should consider the trade-off between the sensor network energy consumption and the required accuracy. For instance, when the contaminant has a low level of danger (e.g., the phosphate concentration [60], dissolved oxygen), one can use a large γ to obtain an approximate recovery. Otherwise, when monitoring and tracking the harmful chemical spreading, a small γ is required to ensure an accurate estimation result.

4   CONCLUSION

In this work, we proposed an innovative methodology to reconstruct the dynamics of the chemical diffusion in the WDNs, optimising the number of sensor required. This methodology outperforms the current state of the art both in terms of number of sensors and reconstruction accuracy. The dynamics are reconstructed using graph Fourier transform driven neural networks.

Two main approaches have been analysed: known and unknown source of possible contamination. On one hand, when the possible source is known (e.g., monitoring industrial areas), we optimise the sensor placement identifying the optimal nodes through GFT analysis and we further reduce the number of monitored points using neural networks. With this approach, we are able to reconstruct the dynamics with high accuracy using 20% of the GFT dataset. In other words, sensors have to be installed in around 5-10% of all the WDN nodes, depending on the injection point and the network structure. On the other hand, when the possible contamination source is unknown (e.g., general water contamination monitoring), several GFT datasets are used to identify an initial ensemble of important nodes. Also in this case, the number of monitored points is further reduced using a neural networks approach. By monitoring 75% of the WDN nodes, we are able to reconstruct the dynamics with a sensitivity up to 98% (low accuracy reconstruction) and up to 51% (high accuracy reconstruction). In applications where a low accuracy reconstruction of the signal is sufficient, we obtain good sensitivity even with 30% of WDN nodes monitored (70% of dynamics correctly reconstructed) and 10% of the WDN nodes monitored (36% of dynamics correctly reconstructed). As already stated, in many real applications, reconstruct the dynamics in core parts of the WDNs may be sufficient to understand when and where there is a contamination and where it is spreading. This allows to act consequently (e.g., closing gates to protect part of the network, alert population). For this reason, an approach that provides low accuracy but with much less sensors to be installed and monitored could be the best trade-off between cost and monitoring performance.

It is noteworthy that the digital and advanced management of urban water network is still in development, as the digital twin modelling may require substantial samples/measurements to represent the complete networked dynamic pattern. Our proposed NN based GFT sampling is able to provide a smaller number sampling set that embraces the spatial and temporal correlated dynamic information, which can help bring that dream closer. The proposed sampling techniques are also useful beyond the application of WDNs and they can be applied to a variety of infrastructure sensing. Future work will focus on how to improve the prediction accuracy using different machine learning techniques and how to further reduce the number of sensors, for example optimising the reconstruction of the dynamics only in the most important and sensible areas of the networks.

[19] F. Archetti, A. Candelieri, and D. Soldi, "Network analysis for resilience evaluation in water distribution networks," *Environmental Engineering and Management Journal*, vol. 14, 2015.

[20] C. Giudicianni, A. Nardo, M. Natale, R. Greco, G. Santonastaso, and A. Scala, "Topological taxonomy of water distribution systems," *Water*, vol. 10, 2018.

[21] A. Simone, L. Ridolfi, D. Laucelli, L. Berardi, and O. Giustolisi, "Centrality metrics for water distribution networks," *EPiC Series in Engineering*, vol. 3, 2018.

[22] A. Di Nardo, C. Giudicianni, R. Greco, M. Herrera, G. Santonastaso, and A. Scala, "Sensor placement in water distribution networks based on spectral algorithms," *13th International Conference on Hydroinformatics (HIC2018)*, 07 2018.

[23] K. Diao, R. Farmani, G. Fu, and D. Butler, "Vulnerability Assessment Of Water Distribution Systems Using Directed And Undirected Graph Theory," *International Conference on Hydroinformatics*, 2014.

[24] J. Hart, I. Guymer, F. Sonnenwald, and V. Stovin, "Residence Time Distributions for Turbulent, Critical, and Laminar Pipe Flow,," *Journal of Hydraulic Engineering*, vol. 142, 2016.

[25] R. Du, L. Gkatzikis, L. Gkatzikis, C. Fischione, and M. Xiao, "Energy Efficient Sensor Activation for Water Distribution Networks Based on Compressive Sensing,," *IEEE Journal on Selected Areas in Communications*, vol. 33, 2015.

[26] S. Kartakis, G. Tzagkarakis, and J. McCann, "Adaptive Compressive Sensing in Smart Water Networks,," *MDPI 2nd International Ele. Conf. on Sensors and Applications*, vol. 6, 2019.

[27] X. Xie, Q. Zhou, D. Hou, and H. Zhang, "Compressed sensing based optimal sensor placement for leak localization in water distribution networks,," *Journal of Hydroinformatics*, vol. 20, 2017.

[28] E. J. Candes and Y. Plan, "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2342–2359, 2011.

[29] Z. Wei, A. Pagani, G. Fu, I. Guymer, W. Chen, J. McCann, and W. Guo, "Optimal sampling of water distribution network dynamics using graph fourier transform," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 3, pp. 1570–1582, 2020.

[30] Z. Wei, B. Li, C. Sun, and W. Guo, "Sampling and inference of networked dynamics using log-koopman nonlinear graph fourier transform," *IEEE Transactions on Signal Processing*, vol. 68, pp. 6187–6197, 2020.

[31] M. Gori, G. Monfardini, and F. Scarselli, "A new model for learning in graph domains," in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, 2005, pp. 729–734 vol. 2.

[32] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2009.

[33] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[34] M. Henaff, J. Bruna, and Y. LeCun, "Deep convolutional networks on graph-structured data," 2015.

[35] R. Levie, F. Monti, X. Bresson, and M. M. Bronstein, "Cayleynets: Graph convolutional neural networks with complex rational spectral filters," *IEEE Transactions on Signal Processing*, vol. 67, no. 1, pp. 97–109, 2019.

[36] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *ICLR 2017*, 2016.

[37] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," 2016.

[38] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," 2019.

[39] T. Mosetlhe, Y. Hamam, S. Du, and Y. Alayli, "Artificial neural networks in water distribution systems: A literature synopsis," *International Conference on Intelligent and Innovative Computing Applications (ICONIC)*, pp. 1–5, 12 2018.

[40] A. Czapczuk, J. Dawidowicz, and J. Piekarski, "Application of multilayer perceptron for the calculation of pressure losses in water supply lines," *Rocznik Ochrona Srodowiska*, vol. 19, 11 2017.

[41] J. Dawidowicz, A. Czapczuk, and J. Piekarski, "The application of artificial neural networks in the assessment of pressure losses in water pipes in the design of water distribution systems," *Rocznik Ochrona Srodowiska*, vol. 20, pp. 292–308, 10 2018.

[42] G. C. Cordoba, L. Tuhovȟák, and M. Tauȟ, "Using artificial neural network models to assess water quality in water distribution networks," *Procedia Engineering*, vol. 70, pp. 399 – 408, 2014, 12th International Conference on Computing and Control for the Water Industry, CCWI2013.

[43] M. A. Andrade, D. Kang, C. Y. Choi, and K. Lansey, "Heuristic postoptimization approaches for design of water distribution systems," *Journal of Water Resources Planning and Management*, vol. 139, no. 4, pp. 387–395, 2013.

[44] Z. Wei, B. Li, and W. Guo, "Optimal sampling for dynamic complex networks with graph-bandlimited initialization," *IEEE Access*, vol. 7, pp. 150 294–150 305, 2019.

[45] S. Chen, R. Varma, A. Sandryhaila, and J. Kovacevic, "Discrete signal processing on graphs: Sampling theory<? pub _newline=""?" *IEEE Transactions on Signal Processing*, vol. 63, no. 24, pp. 6510–6523, 2015.

[46] I. Pesenson, "Sampling in Paley-Wiener spaces on combinatorial graphs," *Transactions of the American Mathematical Society*, vol. 360, no. 10, pp. 5603–5627, 2008.

[47] A. Sandryhaila and J. Moura, "Discrete signal processing on graphs: frequency analysis," *IEEE Transactions on Signal Processing*, vol. 62, 2014.

[48] A. Anis, A. Gadde, and A. Ortega, "Towards a sampling theorem for signals on arbitrary graphs." in *ICASSP*, 2014, pp. 3864–3868.

[49] S. Chen, R. Varma, A. Sandryhaila, and J. Kovacevic, "Discrete Signal Processing on Graphs: Sampling Theory," *IEEE Transactions on Signal Processing*, vol. 63, 2015.

[50] X. Wang, J. Chen, and Y. Gu, "Generalized graph signal sampling and reconstruction," in *Signal and Information Processing (GlobalSIP), 2015 IEEE Global Conference on.* IEEE, 2015, pp. 567–571.

[51] A. Anis, A. Gadde, and A. Ortega, "Efficient sampling set selection for bandlimited graph signals using graph spectral proxies," *IEEE Transactions on Signal Processing*, vol. 64, no. 14, pp. 3775–3789, 2016.

[52] S. Chen, R. Varma, A. Sandryhaila, and J. Kovacevic, "Signal Recovery on Graphs: Fundamental Limits of Sampling Strategies," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 4, 2016.

[53] F. Wang, Y. Wang, and G. Cheung, "A Optimal Sampling and Robust Reconstruction for Graph Signals via Truncated Neumann Series," *arXiv preprint arXiv:1803.03353*, 2018.

[54] L. F. Chamon and A. Ribeiro, "Greedy sampling of graph signals," *IEEE Trans. Signal Process.*, vol. 66, no. 1, pp. 34–47, 2018.

[55] A. Ortega, P. Frossard, J. Kovačević, J. M. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 808–828, 2018.

[56] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE Journal*, vol. 37, no. 2, pp. 233–243, 1991.

[57] K. Klise, R. Murray, and H. T., "An overview of the water network tool for resilience (wntr)," *Computing and Control in the Water Industry*, 2018.

[58] L. Rossman, "Epanet 2 users manual," *U.S. Environmental Protection Agency, Washington, D.C., EPA/600/R-00/057*, 2000.

[59] R. Du, L. Gkatzikis, C. Fischione, and M. Xiao, "Energy efficient sensor activation for water distribution networks based on compressive sensing," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 12, pp. 2997–3010, 2015.

[60] B. O'Flynn, R. Martinez-Catala, S. Harte, C. O'Mathuna, J. Cleary, C. Slater, F. Regan, D. Diamond, and H. Murphy, "Smartcoast: A wireless sensor network for water quality monitoring," in *32nd IEEE Conference on Local Computer Networks (LCN 2007)*, 2007, pp. 815–816.

Neural network approximation of graph Fourier transform for sparse sampling of networked dynamics

Pagani, Alessio

Association for Computing Machinery (ACM)