

Data-Driven Identification of Dynamic Quality Models in Drinking Water Networks

Shen Wang[†], Ankush Chakrabarty[‡], and Ahmad F. Taha^{○,*}

Abstract

Traditional control and monitoring of water quality in drinking water distribution networks (WDN) rely on mostly model- or toolbox-driven approaches, where the network topology and parameters are assumed to be known. In contrast, system identification (SysID) algorithms for generic dynamic system models seek to approximate such models using only input-output data without relying on network parameters. The objective of this paper is to investigate SysID algorithms for water quality model approximation. This research problem is challenging due to (i) complex water quality and reaction dynamics and (ii) the mismatch between the requirements of SysID algorithms and the properties of water quality dynamics. In this paper, we present the first attempt to identify water quality models in WDNs using only input-output experimental data and classical SysID methods without knowing any WDN parameters. Properties of water quality models are introduced, the ensuing challenges caused by these properties when identifying water quality models are discussed, and remedial solutions are given. Through case studies, we demonstrate the applicability of SysID algorithms, show the corresponding performance in terms of accuracy and computational time, and explore the possible factors impacting water quality model identification.

Index Terms

Water distribution networks, water quality models, system identification

LIST OF ACRONYMS

WDN	Water Distribution Network
SysID	System Identification
SIM	Subspace Identification Method
CVA	Canonical Variate Analysis
MOESP	Multivariable Output-error State Space
N4SID	Numerical subspace state space System Identification
ERA	Eigensystem Realization Algorithm
OKID	Observer/Kalman filter Identification
SVD	Singular Value Decomposition
WQMI	Water Quality Model Identification

I. INTRODUCTION AND PAPER CONTRIBUTIONS

Water quality modeling describes the spatiotemporal evolution of disinfectant concentration (e.g., chlorine) in various network elements in drinking water distribution networks (WDN). However, water quality model can still be uncertain due to the complicated reaction dynamics that are severely affected by various factors including topology, parameters, uncertainty (e.g., sensor measurement noise, demand estimation errors), and leakage [1].

Many studies have investigated analysis-based water quality models [2], [3], [4], input-output models [5], [6], [7], [8], [9], and control-oriented state-space model [10]. However, the analysis-based methods either rely on cumbersome, extensive water/chemistry based modeling or are based on obtaining hydraulic and many other parameters—or require both. For example, studies [5], [6], [10] require a two-stage analysis to obtain the final water quality model, that is, water hydraulic analysis and water quality analysis (see Fig. 1). This is because the water quality analysis depends on the results of hydraulic analysis such as the flow rates in all links. Furthermore, all aforementioned studies have made more or less assumptions that simplify the true model such as single-species, first-order linear chlorine dynamics, a well-known WDN topology, perfectly accurate parameters, while not considering uncertainty or leakage. Specifically, inputs of the analysis-based method, marked as red in Fig. 1, contain uncertainties and assumed as known. The limitations existing in analysis-based methods cannot be ignored, but fortunately, are avoided by data-driven based methods; see Tab. I for the comparisons.

[†]School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China. Email: shen.wang@bupt.edu.cn

[‡]Research Scientist, Mitsubishi Electric Research Laboratories. Email: chakrabarty@merl.com

[○]Associate Professor, Department of Civil and Environmental Engineering at Vanderbilt University, Nashville, TN 37235, USA. Email: ahmad.taha@vanderbilt.edu

^{*}Corresponding author.

This work is partially supported by National Science Foundation under Grants 1728629 and 2151392.

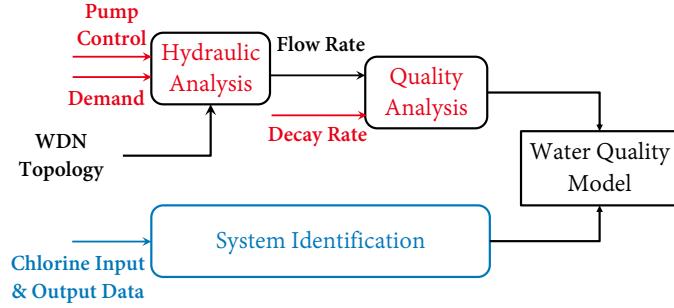


Fig. 1. Two approaches to obtain a water quality model in a WDN—analysis-based and data-driven based methods.

Tab. I. Comparison between analysis-based and data-driven based methods.

	Topology	Parameter	Uncertainty & leakage	Explainable	Hardware	Realtime
Analysis-based modeling	Full topology	Accurate parameters required	Sensitive & hard to model	Excellent	No needs	Yes
Data-driven based modeling	No requirements	No requirements	Robust	Poor	Controllers & sensors	Nearly

A branch of dynamic system sciences, namely data-driven *System Identification* (SysID) of dynamic system models, seek to obtain dynamic models via data-driven approaches, without relying on advanced modeling or knowledge of network parameters or topology. The paper’s objective is to thoroughly investigate, analytically and experimentally, the performance of mainstream identification methods in computing dynamic state-space models of water quality. To the best of our knowledge, this is the first study that seeks to analyze water quality dynamic models without relying on any network parameters, hydraulic simulation data, and only utilizing input and output experimental data (e.g., the control data sequences applied to the system and measurements collected from installed sensors) with the help of using mainstream SysID algorithms .

A. Literature review of system identification

The motivation of SysID is to construct system dynamics/models of a dynamical system from input-output experimental data. To this end, controllers and sensors are required to collect the necessary experimental data and corresponding well-designed algorithms are expected to estimate system dynamics. The models identified are found useful for control or estimation purposes; hence, SysID methods have been widely applied in industrial process [11], aerospace [12], [13], and other fields such as chemistry [14], as well as computational fluid dynamics [15], [16], [17], and power grids [18].

Various SysID algorithms have been developed in the literature of dynamic systems. The majority of these algorithms can be divided into two categories: time-domain methods and frequency-domain methods. Time-domain methods are usually useful for complex, large-scale systems with many inputs and output measurements. Frequency-domain methods are popular in engineering practice for smaller systems [19]. The paper’s focus is on time-domain methods as the water quality model in WDNs involves tens of thousands of states and tens of inputs and outputs [10]. In what follows, we summarize the state-of-the-art of time-domain SysID algorithms.

Time-domain methods can be categorized into various methods. Subspace identification methods (SIMs) [20], [21], [22] estimates linear state-space models directly from time-discrete input-output experimental data and only needs one parameter to be determined by users—the system order. The three famous SIMs from [20], [21], [22] are known as canonical variate analysis (CVA), multivariable output-error state space (MOESP), and numerical subspace state space system identification (N4SID). Overschee and Moor [23] propose a unifying theorem for these three SIMs. They point out that the three algorithms use the same subspace, but the weighting matrix, used to calculate a basis for the column space of the observability matrix, is slightly different.

In short, SIMs only assume that there exists a true underlying finite order linear time-invariant (LTI) system, and no a priori information about the system is needed. Moreover, SIMs are computationally attractive since only standard matrix operations such as singular value decomposition (SVD) are utilized and iterative optimization is not involved. Hence, SIMs have been successfully applied to numerous multi-input multi-output (MIMO) applications [11], [14].

SysID algorithms can also be viewed as the identification of a subspace spanned by an extended observability matrix, followed by the reconstruction of the corresponding state-space matrices [23]. Therefore, methods implementing these two goals can be used to identify system dynamics. A method called eigensystem realization algorithm (ERA) [12], [24] is developed for modal parameter identification and model reduction of dynamic systems from impulse response. However, the requirement of using impulse signals as inputs to obtain impulse response histories (i.e., Markov parameters) can not be met by most realistic systems. To solve such an issue, Juang *et al.* [25] develop an observer/Kalman filter identification (OKID) that adopts commonly-used system inputs (e.g., step signal) to efficiently estimate the Markov parameters. Then, Markov parameters are passed to the ERA or some improved variants of it such as the eigensystem realization algorithm using data correlations (ERA/DC) [26] to complete

the final SysID procedure [27]. The approach combining OKID and ERA is often referred to as OKID/ERA, which is one of the most successful SysID algorithms.

Recently, new system identification algorithms have been proposed such as dynamic mode decomposition (DMD) and sparse identification of nonlinear dynamics (SINDy). DMD [15], [28] is a linear dynamic regression technique to map high-dimensional fluid data sequence into a dynamical system of significantly fewer degrees of freedom in computational fluid dynamics. Proctor *et al.* [29] further extend DMD to include inputs and control, and establish deeper connections to SIMs. Bai *et al.* [30] combine the compressed sensing DMD [31] and DMD with input and control approaches [29], and propose a mathematical framework for compressive SysID. SINDy [16], another novel SysID approach, assumes the governing equations of nonlinear dynamics are sparse in the space of possible functions, and discovers governing equations from data by sparse identification.

With so many years of development, SysID is a classical but still challenging research problem even for linear systems. For example, Zheng and Li [32] focus on solving non-asymptotic identification of linear dynamical systems. Tsiamis and Pappas [33] investigate when SysID is statistically challenging or less so. Using tools from minimax theory, their study shows that classes of linear systems (e.g., under-actuated or under-excited systems with weak coupling among the states) can be hard to identify. There is extensive and rich literature on theoretical and algorithmic developments of SysID, see [34], [35], [36], [37], [32] for further details.

While the theoretical developments of SysID have been rich in the past decade, their application and investigation (analytically and experimentally) to challenging water network problems are rare. A related work from Polycarpou *et al.* [7] adopts input-output data to identify the unknown coefficients of input-output model for water quality using adaptive parameter estimation, and the input-output model structure is identified with the help of hydraulic and water quality analysis from Zierolf *et al.* [5]. However, the input-output model proposed in these two works is different from the state-space form proposed in this paper. To the best of our knowledge, this paper presents the first attempt to demystify SysID methods for estimating water quality dynamic models in the form of state-space representation using classical SysID methods.

B. Paper contribution

The major objective of the paper is to thoroughly explore the applicability of SysID algorithms in water quality models, investigate the corresponding performance, and possible factors affecting the performance of SysID methods in estimating water quality dynamics. The below list outlines in detail the contributions of the paper.

- 1) We present the first attempt to identify water quality dynamics in state-space representation form in WDNs using only input-output data collected from experiments and classical SysID methods such as SIMs and ERA-based methods without knowing *a priori* information about the system such as any parameters of WDNs.
- 2) To adopt the suitable SysID methods and apply these algorithms correctly, exploring the properties or characteristics of water quality dynamics is important and necessary. Hence, we point out several significant and unique properties in WDNs such as *linear-invariant*, *high-delay* and *overall uncontrollability and unobservability*. Moreover, the challenges caused by these properties when identifying water quality models (i.e., mismatches between SysID algorithms and the properties) are discussed, and possible solutions are proposed.
- 3) Case studies show the potential of applying classical system identification algorithms in water quality models/dynamics, the performance of these algorithms in terms of accuracy and computational time are tested, and the possible factors affecting algorithm performance are explored. The significance of this work is that it allows water system operators to (*i*) simulate water quality dynamics in the entire network without knowing network parameters, topology, or hydraulic and reaction parameters or even leaks, and (*ii*) potentially apply modern control/estimation algorithms to manage water quality.

The rest of the paper is organized as follows. Section II briefly describes the basics of water quality modeling and its state-space representation. Section III introduces SysID algorithms that are potentially applicable/useful to water quality models. Properties of water quality models are proposed for the first time in Section IV, followed by the challenges caused by these properties when identifying water quality models and the discussion of possible solutions. Section V presents case studies showing the applicability of various SysID algorithms, their performance, and possible factors impacting the final results. Conclusions and future research directions are given in Section VI.

Paper's Notation. Boldface characters represent matrices and column vectors: a is a scalar, \mathbf{a} is a vector, and \mathbf{A} is a matrix. Matrix \mathbf{I} denotes an identity square matrix, whereas $\mathbf{0}_{m \times n}$ denotes a zero matrix with size m -by- n . The notation \mathbb{R} denotes the set of real and positive real numbers, and the notations \mathbb{R}^n and $\mathbb{R}^{m \times n}$ denote a column vector with n elements and an m -by- n matrix in \mathbb{R} . For any vector $\mathbf{x} \in \mathbb{R}^n$, \mathbf{x}^\top is its transpose. For any two matrices \mathbf{A} and \mathbf{B} with same number of columns, the notation $\{\mathbf{A}, \mathbf{B}\}$ denotes $[\mathbf{A}^\top \mathbf{B}^\top]^\top$. The induced norm of \mathbf{A} is $\|\mathbf{A}\|_2$, and the Frobenius norm of \mathbf{A} is $\|\mathbf{A}\|_F$. The Moore-Penrose pseudoinverse of \mathbf{A} is denoted by \mathbf{A}^\dagger .

II. WATER QUALITY MODELING

In this section, we succinctly present the fundamentals of water quality modeling considering single-species interactions and the corresponding state-space representation.

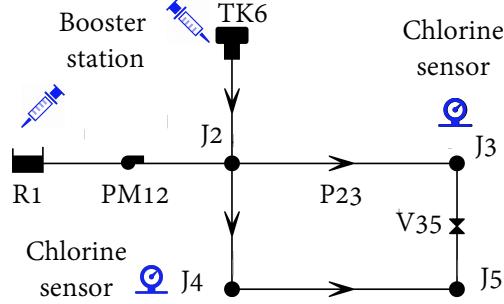


Fig. 2. Exemplar topology of a WDN (multi-input and multi-output system with two boosters (inputs) and two sensors (outputs) to inject and to detect chlorine).

A. Water quality dynamics

We model WDN by a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{L})$. The set \mathcal{N} defines the nodes and is partitioned as $\mathcal{N} = \mathcal{J} \cup \mathcal{T} \cup \mathcal{R}$ where \mathcal{J} , \mathcal{T} , and \mathcal{R} are collection of junctions, tanks, and reservoirs. Let $\mathcal{L} \subseteq \mathcal{N} \times \mathcal{N}$ be the set of links, and define the partition $\mathcal{L} = \mathcal{P} \cup \mathcal{M} \cup \mathcal{V}$, where \mathcal{P} , \mathcal{M} , and \mathcal{V} represent the collection of pipes, pumps, and valves.

An example of a WDN graph is shown in Fig. 2, and a disinfectant (e.g., chlorine) is injected into the network with a proper mass and concentration by booster stations installed at the reservoir source R_1 and the node TK_6 to disinfect the WDN. Each junction in a WDN is a type of node that connects two links. Junctions may consume water, i.e., they have water demands which are assumed as known and unchanged during a hydraulic time-step (an hour or two) in water research. The concentration sensors are installed at specific nodes for measurement purposes (see the sensors installed at J_3 and J_4 in Fig. 2), in this way the water system operators can obtain the concentrations around the network. Ultimately, a water quality model representing the movement of all chemical and/or microbial species within a WDN as they traverse various components of the network can be represented by state-space form equations [10].

B. State-space representation of water quality dynamics

A water quality state-vector \mathbf{x} defining the concentrations of the disinfectant (chlorine) in the network at time t is denoted as $\mathbf{x}(t) \triangleq \{\mathbf{c}^N(t), \mathbf{c}^L(t)\} \in \mathbb{R}^{n_x}$ and $n_x = n_N + n_L$ where vectors \mathbf{c}^N and \mathbf{c}^L stand for the concentration at nodes and in links, the number of elements in vectors are n_N and n_L . Specifically, pipes are split into segments, and each segment has a variable in \mathbf{c}^L standing for its concentration. After applying the principle of conversion of mass and assuming a constant decay rate, we can rewrite all component models (i.e., water quality model) using a discrete-time state-space model of the form [10]

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{A}(k)\mathbf{x}(k) + \mathbf{B}(k)\mathbf{u}(k) \\ \mathbf{y}(k) &= \mathbf{C}(k)\mathbf{x}(k) + \mathbf{D}(k)\mathbf{u}(k), \end{aligned} \quad (1)$$

where $\mathbf{x}(k) \in \mathbb{R}^{n_x}$ is the state vector, the vectors $\mathbf{u}(k) \in \mathbb{R}^{n_u}$ and $\mathbf{y}(k) \in \mathbb{R}^{n_y}$ are the system input and output vectors (i.e., n_u and n_y are the number of booster stations and sensors installed), and $\mathbf{A}(k)$, $\mathbf{B}(k)$, $\mathbf{C}(k)$, and $\mathbf{D}(k)$ are system matrices and time-varying.

Note that system matrices are usually updated from half hour to several hours according to the changing rate of demands. That is, during a specific period which is referred to as *hydraulic time-step*, the system matrices remain the same, and the discrete-time linear time-varying (DT-LTV) system (1) degenerates into a discrete-time linear time-invariant (DT-LTI) system, that is

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k) \\ \mathbf{y}(k) &= \mathbf{C}\mathbf{x}(k) + \mathbf{D}\mathbf{u}(k). \end{aligned} \quad (2)$$

Let us denote a generic n_x -th order DT-LTI system (2) as the tuple $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ for brevity. In this way, the DT-LTV system (1) comprises different tuples (DT-LTI systems). It is worthwhile to note that (i) water quality modeling is a DT-LTV system, (ii) the DT-LTV system can be decomposed into several DT-LTI systems, and (iii) we can apply SysID algorithms to each single DT-LTI system to obtain the identified model of the whole DT-LTV system.

III. SYSTEM IDENTIFICATION METHODS

From the literature review, various SysID algorithms including SIMs, ERA-based, DMD-based, and SINDy methods can theoretically be applied to *water quality model identification* (WQMI). However, we notice that not every method is suitable for our application. For example, DMD-based methods and SINDy are not applicable due to DMD assumes installing sensors at all nodes while SINDy depends on the choice of measurement variables, data quality, and the sparsifying function basis; this is thoroughly discussed in Section IV-B. Hence, only SysID algorithms that are potentially applicable to WQMI such as SIMs and ERA-based methods are introduced next.

A. Fundamentals of system identification

1) *Subspace identification method*: Instead of introducing each algorithm of the SIMs, that are N4SID [20], MOESP [21], and CVA [22], we summarize the unifying theorem [23] that unifies the three SIMs. Note that the identification problem in [23] is considered in a combined deterministic-stochastic scenario. However, we only present the core ideas of the unifying SIM in a deterministic scenario for simplicity, since the key procedures considering state and measurement noises are similar for the stochastic scenario.

The unifying SIM organizes the state, input, and output data in a special way. Recall that $\mathbf{x}(k)$ in (2) stands for the system state, and it is collected into state sequence \mathbf{X} . That is

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_0 & \mathbf{X}_k \\ \mathbf{x}(0) \cdots \mathbf{x}(k) \cdots \mathbf{x}(m-1) \cdots \mathbf{x}(k+m-1) \cdots \end{bmatrix}, \quad (3)$$

where sub-sequence $\mathbf{X}_k \in \mathbb{R}^{n_x \times m}$ starts from index k and includes m -step data. The output sub-sequence $\mathbf{Y}_k \in \mathbb{R}^{n_y \times m}$ are defined similarly as \mathbf{X}_k , that is

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_0 & \mathbf{Y}_k \\ \mathbf{y}(0) \cdots \mathbf{y}(k) \cdots \mathbf{y}(m-1) \cdots \mathbf{y}(k+m-1) \cdots \end{bmatrix}.$$

Furthermore, the input data sequence with $2k + m$ steps, that is $\mathbf{U} = [\mathbf{u}(0) \ \mathbf{u}(1) \ \cdots \ \mathbf{u}(2k + m - 1)]$, is organized in the following form

$$\mathbf{U} = \begin{bmatrix} \mathbf{U}_0 \{ \underbrace{\mathbf{u}(0) \ \mathbf{u}(1) \cdots \mathbf{u}(k) \ \cdots \ \mathbf{u}(m-1) \ \cdots \ \mathbf{u}(k+m-1)}_{\mathbf{U}_p} \} \ \mathbf{u}_k \\ \mathbf{u}(1) \ \mathbf{u}(2) \cdots \mathbf{u}(k+1) \cdots \mathbf{u}(m) \ \cdots \ \mathbf{u}(k+m) \\ \vdots \ \vdots \ \vdots \ \vdots \ \vdots \ \vdots \\ \mathbf{U}_{k-1} \{ \underbrace{\mathbf{u}(k-1) \mathbf{u}(k) \cdots \mathbf{u}(2k-1) \cdots \mathbf{u}(k+m-2)}_{\mathbf{U}_f} \cdots \mathbf{u}(2k+m-1) \} \end{bmatrix},$$

where input sub-sequences \mathbf{U}_0 and \mathbf{U}_k are defined akin to \mathbf{X}_k , whereas Hankel blocks \mathbf{U}_p and \mathbf{U}_f in $\mathbb{R}^{(kn_x) \times (mn_u)}$ are considered as the past and future input data, that are $\{\mathbf{U}(0), \dots, \mathbf{U}(k-1)\}$ and $\{\mathbf{U}(k), \dots, \mathbf{U}(2k-1)\}$.

System identification from unifying SIM [23] is achieved by solving a least-square problem, that is

$$\underbrace{\begin{bmatrix} \mathbf{X}_{k+1} \\ \mathbf{Y}_k \end{bmatrix}}_{\text{known}} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \underbrace{\begin{bmatrix} \mathbf{X}_k \\ \mathbf{U}_k \end{bmatrix}}_{\text{known}}, \quad (4)$$

where input \mathbf{U}_k and output \mathbf{Y}_k data sequences are known, and if \mathbf{X}_k and \mathbf{X}_{k+1} which are also referred to as *Kalman state sequence* can be obtained, then system matrices $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ can be fully recovered from (4) by the least-square method. Next, we briefly present how unifying SIM estimates Kalman state sequence \mathbf{X}_k or \mathbf{X}_{k+1} .

With defining data sequences such as \mathbf{X}_k , \mathbf{Y}_k , \mathbf{U}_p , and \mathbf{U}_f , the system (2) can be rewritten as

$$\begin{aligned} \mathbf{X}_k &= \mathbf{A}^k \mathbf{X}_0 + \Delta_k \mathbf{U}_p \\ \mathbf{Y}_0 &= \Gamma_k \mathbf{X}_0 + \mathbf{H}_k \mathbf{U}_p \\ \mathbf{Y}_k &= \Gamma_k \mathbf{X}_k + \mathbf{H}_k \mathbf{U}_f, \end{aligned} \quad (5)$$

where extended observability matrix Γ_k and Toeplitz matrix \mathbf{H}_k are defined by

$$\Gamma_k = \begin{bmatrix} \mathbf{C} \\ \mathbf{C}\mathbf{A} \\ \vdots \\ \mathbf{C}\mathbf{A}^{k-1} \end{bmatrix}, \quad (6a)$$

$$\mathbf{H}_k = \begin{bmatrix} \mathbf{D} & & \\ \mathbf{C}\mathbf{B} & \mathbf{D} & \\ \vdots & \ddots & \ddots \\ \mathbf{C}\mathbf{A}^{k-2}\mathbf{B} & \cdots & \mathbf{D} \end{bmatrix}. \quad (6b)$$

If one observes the three subequations in (5) carefully, the connections among input and output data sequences \mathbf{Y}_k , \mathbf{Y}_0 , \mathbf{U}_p , and \mathbf{U}_f are achieved by state sequences \mathbf{X}_k , \mathbf{X}_0 , and other necessary matrices. The unifying SIM theorem points out that matrix $\mathcal{O}_k = \Gamma_k \mathbf{X}_k$ is equal to the oblique projection of \mathbf{Y}_k along \mathbf{U}_f direction on the past data $\begin{bmatrix} \mathbf{U}_p \\ \mathbf{Y}_0 \end{bmatrix}$, and each column of it is the autonomous response to initial states; see Appendix in [23] for more details. In short, $\mathcal{O}_k = \Gamma_k \mathbf{X}_k$ representing autonomous response can be easily found by calculating an oblique projection. Moreover, Kalman state sequence \mathbf{X}_k which is our goal can be extracted directly from SVD for \mathcal{O}_k , that is

$$\mathcal{O}_k = \mathbf{U} \Sigma \mathbf{V}^\top = \underbrace{\mathbf{U} \Sigma^{\frac{1}{2}}}_{\Gamma_k} \underbrace{\Sigma^{\frac{1}{2}} \mathbf{V}^\top}_{\mathbf{X}_k}. \quad (7)$$

We can adopt similar procedures to estimate Kalman state sequence \mathbf{X}_{k+1} , and after substituting them into (4), system matrices

can be identified. In addition to (4), another popular strategy to recover system matrices is introduced in [38] by taking advantage of the special structure of Γ_k and the special Toeplitz structure of \mathbf{H}_k .

The aforementioned procedures are only the core idea of unifying SIM, other details are simplified to help readers understand the big picture of SIMs. For example, (7) is actually a special case of SVD of $\mathbf{W}_1 \mathcal{O}_k \mathbf{W}_2$ when setting weight matrices $\mathbf{W}_1 = \mathbf{W}_2 = \mathbf{I}$. In this way, this unifying method generates the same results as the N4SID method published in [20]. When selecting different matrices \mathbf{W}_1 and \mathbf{W}_2 combinations, the unifying SIM can be viewed as CVA and MOESP methods as well. Moreover, to reduce the system order, we can choose the first n_r singular values when performing SVD for $\mathbf{W}_1 \mathcal{O}_k \mathbf{W}_2$, that is

$$\mathbf{W}_1 \mathcal{O}_k \mathbf{W}_2 = \mathbf{U} \boldsymbol{\Sigma} \mathbf{V}^\top = \underbrace{[\mathbf{U}_r \ \mathbf{0}]}_{\Gamma_k} \underbrace{\begin{bmatrix} \boldsymbol{\Sigma}_r^{\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}}_{\mathbf{X}_k} \underbrace{\begin{bmatrix} \boldsymbol{\Sigma}_r^{\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}}_{\mathbf{V}_r^\top} \underbrace{\begin{bmatrix} \mathbf{V}_r^\top \\ \mathbf{0} \end{bmatrix}}_{\mathbf{X}_k}. \quad (8)$$

As for how to determine a proper value for system order n_r , the users could use trial-and-error method or a binary search based algorithm provided at the end of Section III-B; see Algorithm 1 for detail. The unifying SIM are summarized as Procedure 1.

Procedure 1: Unifying subspace identification method (SIM)

Input: Input block Hankel matrices \mathbf{U}_p and \mathbf{U}_f , ouput data sequences \mathbf{Y}_0 and \mathbf{Y}_k , system order n_r

Output: Identified system matrices $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$

- 1 Solve an oblique projection from \mathbf{Y}_k to $\begin{bmatrix} \mathbf{U}_p \\ \mathbf{Y}_0 \end{bmatrix}$ along \mathbf{U}_f direction and obtain autonomous response \mathcal{O}_k
- 2 Select proper weight matrices \mathbf{W}_1 and \mathbf{W}_2 in (8) to determine the use of CVA, MOESP, or N4SID
- 3 Choose system order n_r and perform SVD for $\mathbf{W}_1 \mathcal{O}_k \mathbf{W}_2$ as (8) to obtain Kalman state sequence \mathbf{X}_k , \mathbf{X}_{k+1} in (4), and extended observability matrix Γ_k in (6a)
- 4 Obtain $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ by (4) or the strategy in [38]

2) *ERA method:* As mentioned in Section I-A, ERA [12] is developed for a finite-dimensional, DT-LTI dynamical system such as (2). The core idea is to use impulse signals as system inputs to construct Markov parameters, that are \mathbf{D} and $\mathbf{C} \mathbf{A}^{k-1} \mathbf{B}$. Specifically, suppose that the unit impulse signal is defined as

$$u_i(k) = \begin{cases} 1, & k = 0 \\ 0, & k \neq 0 \end{cases},$$

where input $u_i(k)$ is the element of \mathbf{u} with index of (i, k) .

Then, the output sequence can be expressed by

$$\mathbf{y}(k) = \begin{cases} \mathbf{D}, & k = 0 \\ \mathbf{C} \mathbf{A}^{k-1} \mathbf{B}, & k \neq 0 \end{cases}. \quad (9)$$

Note that matrix \mathbf{D} in (2) is already identified and can be obtained directly from $\mathbf{y}(0)$. We present how to identify other matrices next.

First, the ERA algorithm begins by forming the *block Hankel matrix* \mathbf{H}_m :

$$\mathbf{H}_m = \begin{bmatrix} \mathbf{y}(1) & \mathbf{y}(2) & \cdots & \mathbf{y}(m_c) \\ \mathbf{y}(2) & \mathbf{y}(3) & \cdots & \mathbf{y}(m_c + 1) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{y}(m_o) & \mathbf{y}(m_o + 1) & \cdots & \mathbf{y}(m_c + m_o) \end{bmatrix} = \begin{bmatrix} \mathbf{C} \mathbf{B} & \mathbf{C} \mathbf{A} \mathbf{B} & \cdots & \mathbf{C} \mathbf{A}^{m_c} \mathbf{B} \\ \mathbf{C} \mathbf{A} \mathbf{B} & \mathbf{C} \mathbf{A}^2 \mathbf{B} & \cdots & \mathbf{C} \mathbf{A}^{m_c+1} \mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C} \mathbf{A}^{m_o} \mathbf{B} & \mathbf{C} \mathbf{A}^{m_o+1} \mathbf{B} & \cdots & \mathbf{C} \mathbf{A}^{m_c+m_o} \mathbf{B} \end{bmatrix},$$

where parameters m_c and m_o are the numbers of steps.

Second, ERA computes SVD of \mathbf{H}_m in (10) to obtain \mathbf{U}_r and \mathbf{V}_r .

$$\mathbf{H}_m = \mathbf{U} \boldsymbol{\Sigma} \mathbf{V}^\top = \underbrace{[\mathbf{U}_r \ \mathbf{0}]}_{\Gamma_m} \underbrace{\begin{bmatrix} \boldsymbol{\Sigma}_r^{\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}}_{\mathbf{X}_m} \underbrace{\begin{bmatrix} \mathbf{V}_r^\top \\ \mathbf{0} \end{bmatrix}}_{\mathbf{X}_m}, \quad (10)$$

where singular value $\boldsymbol{\Sigma}_r$ is in $n_r \times n_r$, and constant n_r is chosen by users as the dimension of identified system or obtained by Algorithm 1 at the end of Section III-B.

Finally, the identified system matrices $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ are

$$\begin{aligned} \mathbf{A} &= \boldsymbol{\Sigma}_r^{-\frac{1}{2}} \mathbf{U}_r^\top \mathbf{H}_m' \mathbf{V}_r \boldsymbol{\Sigma}_r^{-\frac{1}{2}}, \\ \mathbf{B} &= \text{the first } n_u \text{ columns of } \boldsymbol{\Sigma}_r^{\frac{1}{2}} \mathbf{V}_r^\top, \\ \mathbf{C} &= \text{the first } n_y \text{ rows of } \mathbf{U}_r \boldsymbol{\Sigma}_r^{\frac{1}{2}}, \\ \mathbf{D} &= \mathbf{y}(0), \end{aligned} \quad (11)$$

where

$$\mathbf{H}'_m = \begin{bmatrix} CAB & CA^2B & \cdots & CA^{m_c+1}B \\ CA^2B & CA^3B & \cdots & CA^{m_c+2}B \\ \vdots & \vdots & \ddots & \vdots \\ CA^{m_o+1}B & CA^{m_c+2}B & \cdots & CA^{m_c+m_o+2}B \end{bmatrix}.$$

The general ERA procedure from [12] are summarized as Procedure 2.

Procedure 2: Eigensystem realization algorithm (ERA)

Input: Impulse signal, system order n_r

Output: Identified system matrices $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$

- 1 Collect output data using (12). Obtain Markov parameters \mathbf{D} and $CA^{k-1}B$ using impulse signal
- 2 Construct block Hankel matrix \mathbf{H}_m using (10)
- 3 Perform SVD of \mathbf{H}_m via (10) with chosen system order n_r
- 4 Obtain $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ by (11)

Note that ERA requires impulse signal as system inputs, and this may be not applicable to WQMI since the amplitude of impulse response is not large enough to be detected by sensors. We investigate that in the ensuing section and the case study.

3) *OKID/ERA method:* To avoid using the impulse signal and make ERA more general, OKID [25] is developed to efficiently estimate the Markov parameters (i.e., \mathbf{D} and $CA^{k-1}B$). With an assumption of zero-initial conditions, the system equation (2) can be rewritten as

$$\underbrace{\begin{bmatrix} \mathbf{y}(0) \\ \mathbf{y}(1) \\ \vdots \\ \mathbf{y}(m) \end{bmatrix}}_{\mathbf{y}_m} = \underbrace{\begin{bmatrix} \mathbf{D} & \mathbf{C}B & \cdots & \mathbf{C}A^mB \end{bmatrix}}_{\mathbf{Y}_m} \underbrace{\begin{bmatrix} \mathbf{u}(0) & \mathbf{u}(1) & \cdots & \mathbf{u}(m) \\ \mathbf{u}(0) & \cdots & \mathbf{u}(m-1) \\ \ddots & & \vdots \\ \mathbf{u}(0) & & & \mathbf{u}(0) \end{bmatrix}}_{\mathbf{U}_m}, \quad (12)$$

where parameter m represent the number of steps. All Markov parameters are collected in \mathbf{Y}_m that connects m -step input matrix \mathbf{U}_m and output sequence \mathbf{y}_m .

Then, the Markov parameters located in \mathbf{Y}_m can be simply solved by $\mathbf{y}_m \mathbf{U}_m^\dagger$. As for large-scale systems or a large m , \mathbf{Y}_m can also be solved by SVD techniques. After Markov parameters are obtained, the standard ERA procedure is called to complete the final system identification; the general OKID/ERA procedure is presented in Procedure 3.

Procedure 3: OKID followed by eigensystem realization algorithm (OKID-ERA)

Input: System input sequence $\mathbf{u}(0), \mathbf{u}(1), \dots, \mathbf{u}(m)$; output sequence $\mathbf{y}(0), \mathbf{y}(1), \dots, \mathbf{y}(m)$, system order n_r

Output: Identified system matrices $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$

- 1 Construct input-output relationship (12) and solve for each Markov parameter in matrix \mathbf{Y}_m
- 2 Call Steps 2-5 in Procedure 2 to obtain $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$

B. System order determination algorithm

As mentioned in the above procedures, system order n_r is the only parameter to be determined by users. The users could finish this task according to their experiences and adopt the trial-and-error method. However, setting system order by intuition is cumbersome and is not a stable way when the singular values vary significantly for different matrices to be decomposed using the SVD method. Hence, we provide a stable indicator (i.e., energy level) for determining system order next.

Singular values define the “energy” of each state in a system $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$. Thus, keeping larger energy states of a system preserves most of its characteristics in terms of stability, frequency, and time responses. Note that each SysID algorithm mentioned in this paper includes SVD procedure and an available singular value matrix Σ ; see (8) or (10). To determine the system order automatically and efficiently, we use the concept of energy level—the ratio of the sum of the n_r largest singular values to the sum of all singular value—that is

$$l = \frac{\sum_{i=1}^{n_r} \Sigma_{i,i}}{\sum_{i=1}^n \Sigma_{i,i}}. \quad (13)$$

To preserve most of the characteristics in a system, the users simply set a minimal energy level goal (e.g., $l > 95\%$) which corresponds to a certain n_r . We provide a binary search algorithm to determine the system order n_r when the energy level is set by users, and the detailed steps are given in Algorithm 1.

Algorithm 1: Binary search based algorithm to determine system order n_r .

Input: Energy level goal g , singular value matrix Σ in (8) or (10)
Output: System order n_r

- 1 Let $n \leftarrow$ number of elements in singular value matrix Σ , $left \leftarrow 1$, $right \leftarrow n$
- 2 Compute the sum of singular values $S = \sum_{i=1}^n \Sigma_{i,i}$
- 3 while $|left - right| > 1$ do
 - 4 $n_r \leftarrow \lfloor \frac{left+right}{2} \rfloor$ // Find the middle number as current order
 - 5 Obtain current energy level $l \leftarrow \sum_{i=1}^{n_r} \Sigma_{i,i}/S$ // Find the current energy level
 - 6 if $l > g$ then
 - 7 $right \leftarrow n_r$
 - 8 else
 - 9 $left \leftarrow n_r$
 - 10 end if
- 11 end while

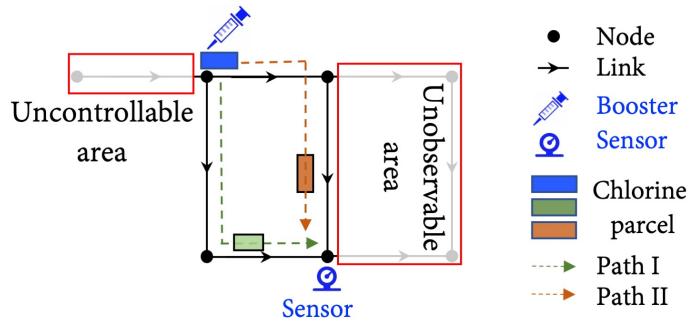


Fig. 3. An example to illustrate the properties of water quality dynamics.

IV. PROPERTIES OF WATER QUALITY MODEL, CHALLENGES OF IDENTIFICATION, AND POSSIBLE SOLUTIONS

Unlike other applications, the water quality model has special characteristics/properties. Hence, we show these special properties, present challenges caused by them, and discuss corresponding solutions while performing WQMI.

To illustrate these issues clearly, we use the network comprised of an output sensor and a booster station controller as an example; see Fig. 3. As the aforementioned topology in Section II-A, the WDN in Fig. 3 is expressed by a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{L})$. To simplify, only junctions are included in set \mathcal{N} , and the set \mathcal{L} only collects pipes in this example. Furthermore, the amount of chlorine injected by the booster station controller in Fig. 3 is viewed as the system input data; the chlorine concentration measured by the sensor in Fig. 3 is the corresponding system output data.

A. Properties of water quality model

Let us present the process of collecting the input-output data for SysID algorithms at first, and then the special properties can be summarized from the process intuitively.

During a fixed hydraulic time-step, junction demands remain the same, the hydraulics in a WDN are unchanged, and the water quality model can be described as a DT-LTI system (2). After a chlorine parcel with a certain mass and concentration is injected at the booster station (i.e., the blue parcel in Fig. 3), it travels to the sensor with various velocities in different links (i.e., the Paths I and II in Fig. 3). During traveling, both mass and concentration of the chlorine parcel reduce gradually due to consumption by junctions and decay of chlorine. Furthermore, parcels with different masses and concentrations (see the light green and brown parcels in Fig. 3) arrive at the sensor at various times. Data collection starts in all booster stations and sensors from the moment when the booster station injects the chlorine parcel, and stops at the moment when the last chlorine parcel reaches the sensor.

From the above input-output data collection process, we note several properties of the water quality model below.

- The water quality model is a special DT-LTV system with regular and consistent properties. That is, it degenerates into a simple DT-LTI system during each hydraulic time-step. We refer this property as *linear-invariant property*.
- In practice, the water quality model is stable. This *stable property* is intuitive because the injected chlorine would gradually decay and disappear eventually in a WDN after enough time. A system operator continuously injects chlorine to ensure that the chlorine concentrations do not decay to zero, which would violate federal water quality mandates.
- The overall water quality model, in general, is uncontrollable and unobservable from a control-theoretic perspective due to the limited number of booster stations (i.e., controllers) and measurement sensors that are allowed to be installed in many real-world WDNs. That is, there exist certain components that cannot be affected by booster stations, and certain components

have no effect on the sensors; see the uncontrollable and unobservable areas marked by red, rectangular boxes in Fig. 3. This property is referred to as *overall uncontrollability and unobservability*.

- Significant delays exist in the water quality model. The delay, caused by the slow traveling time from booster stations to sensors, is a natural attribute of the water supply system. This property is named as *high-delay property*.
- Sensors are assumed to be only installed at nodes in reality and installing sensors in links (e.g., in the center of a pipe) is not allowed. This *partial-state access property* indicates that C is not allowed to be I . That is, full-state access is impossible in a water quality scenario.

B. Challenges and solutions for water quality model identification

From above property analysis, we note that the *linear-invariant* and *stable* properties are supportive and friendly for WQMI using SysID algorithms since most SysID algorithms are designed for stable LTI systems. However, the *overall uncontrollability and unobservability*, *high-delay*, and *partial-state access* properties would cause problems and still are challenges for current state-of-the-art SysID algorithms.

First, SIMs introduced in this paper assume a controllable and observable system [22], [21], [20], [23] and ERA-related methods [12], [25] also make similar assumptions. Although standard DMD has no such requirements, it assumes full-state access and cannot deal with control inputs. Compressed sensing DMD [39] and random sampling strategy DMD [40] require a special measurement matrix C . For example, the C must be incoherent to the sparse basis [39]. It indicates DMD-related methods are not applicable. Hence, the *overall uncontrollability and unobservability* property indicates that the SysID of water quality model for overall WDN is infeasible.

However, there are always both controllable and observable areas in a WDN, that is the part “surrounded” by controllers and sensors (see the middle area in Fig. 3). The area is controllable because the chlorine concentrations of all pipes and junctions in that area can be affected by the booster station; all chlorine concentrations have effects on the sensors indicating the existence of observable property. From a control-theoretical perspective, the state-space representation of the middle part is the controllable and observable subsystem after performing Kalman decomposition for the overall/full state-space representation. Therefore, we are allowed to adopt the current popular SysID algorithms to estimate that part of the water quality model in a WDN. Moreover, the identified subsystem would have the same behavior as the part of the original full system surrounded by controllers and sensors.

Second, the *high-delay property* also brings challenges to WQMI, which is reflected in the completeness of the data received by sensors. Suppose that the traveling time from the booster station controllers to sensors is larger than the hydraulic time-step. Sensors have not received enough/complete data for SysID yet, and the current DT-LTI system has been switched to the next one because of the updated demand. This would cause a direct failure for SysID algorithms. Hence, the duration of the signal selected as input has to be as short as possible or finite length such that the data completeness can be guaranteed.

There are two possible solutions for this high-delay issue. The simple one is to install the controllers and sensors close enough to each other so that the traveling time is always less than a hydraulic time-step. But this clearly would limit the size of the system we can identify. Moreover, this might be infeasible since the system operator might not have enough budget to install more sensors or controllers. The complicated one is to use repeaters to cascade the system when the controllers and sensors are far away from each other. The repeaters are actually sensors but viewed as system inputs. That is, the chlorine concentration data measured by repeater sensors is treated as the input data injected by booster stations. In this way, a large-scale system cannot be identified due to lack of data completeness is solved by cascading two identified two subsystems. Suppose we have two identified small-scale subsystems (A_1, B_1, C_1, D_1) and (A_2, B_2, C_2, D_2) . After cascading, the final identified system is

$$\left(\begin{bmatrix} A_1 \\ B_2 C_1 & A_2 \end{bmatrix}, \begin{bmatrix} B_1 \\ B_2 D_1 \end{bmatrix}, [D_2 C_1 \quad C_2], D_2 D_1 \right).$$

Third, the challenge from *partial-state access property* is tricky. If WQMI for the overall system is expected, then there is no solution since it is impossible to install sensors at all junctions and in all links to meet the controllable and observable requirements in most SysID algorithms. But if the goal is only to identify the subsystem discussed in the aforementioned *overall uncontrollability and unobservability property* instead of the whole system, then the C for the overall system does not have to be the identity matrix, meaning that the subsystem can be identified using only a few sensors installed.

V. CASE STUDIES

We present two examples (i.e., Three-node and Net1 networks [41], [42]) to illustrate the properties (e.g., stability) and the performance of different SysID methods in terms of accuracy and computational time. The testing environment is an Ubuntu 16.04 Precision 5810 Tower with Intel(R) Xeon(R) CPU E5-1620 v3 @ 3.50GHz and 8G. Specially, in this numerical study we attempt to answer the following questions:

- *Q1*: Can the water system operator only utilize the collected input-output experimental data to identify the water quality model without knowing any other network parameters?
- *Q2*: How do the aforementioned algorithms perform in terms of performance (e.g., accuracy and running time) of tested SysID methods for different networks?

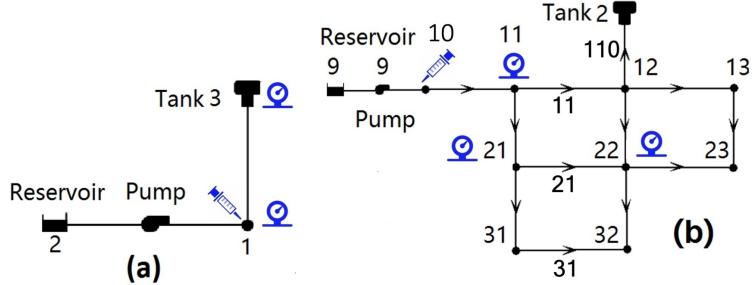


Fig. 4. Three tested networks: (a) Three-node network; (b) Net1 network.

Tab. II. Basic information of two tested networks.

Networks	# of components*	Full order n_x	Booster station locations (n_u)	Sensors location (n_y)
Three-node network	{1,1,1, 1,1,0}	154	J2 (1)	J2, TK3 (2)
Net1	{9,1,1, 12,1,0}	1,293	J10 (1)	J11, J21, J22 (3)

*Number of each component in WDN: $\{n_J, n_R, n_{TK}, n_P, n_M, n_V\}$.

- *Q3:* How can SysID methods help water system operators to control the chlorine concentration in the interested area?

A. Network settings and validation

1) *Setup:* The topologies of tested networks are shown in Fig. 4. Three-node network (see Fig. 4a) is a simple, self-designed network for the illustrative purpose that includes one junction J1, one pipe P13, one pump PM21, one tank TK3, and one reservoir R2. A booster station is installed at J1 and two chlorine concentration sensors are put at J1 and TK3 to increase observability, that is $n_u = 1$ and $n_y = 2$. Similarly, we also tested the Net1 in Fig. 4b which is a looped and more complicated network. Note that Algorithm 1 is used to determine n_r first after setting the energy level goal, and then we select the largest n_r as a common one that works for all SysID algorithms for comparison purpose.

Note that we assume the hydraulic time-steps of tested networks are large enough so that the tested networks can be viewed as an LTI system. The system inputs $u(k)$, implemented by controlling booster stations, could be an impulse, rectangular, or random signal according to the SysID methods we choose to use in this study; see Fig. 5. Moreover, the inputs are divided into test and validation inputs based on the purpose. For a short hydraulic time-step that might cause data completeness issues, we prefer to explore the case in the future even though a possible solution has been proposed in Section IV-B. Furthermore, zero-initial conditions and zero measurement noises are assumed in this case study.

2) *Validation:* The core idea of the designed experiments in this numerical study is applying the same input signals (for validation purpose) to the original and identified state-space models simultaneously, and then comparing output differences. The accuracy of identified models is validated by the overlapping degree of responses. Mathematically, we use $\text{RMSE} = \sqrt{\frac{1}{m} \sum_{k=1}^m \|\mathbf{y}_e(k)\|_2^2}$ to quantify the error, where $\mathbf{y}_e(k)$ is defined by $\mathbf{y}(k) - \hat{\mathbf{y}}(k)$ is the absolute error, and vectors $\mathbf{y}(k)$ and $\hat{\mathbf{y}}(k)$ are measured outputs of the original and identified models.

The original state-space models for the tested two networks available in [10] are built by analysis-based methods and only used to simulate the tested networks and generate output data (i.e., measurement $\mathbf{y}(k)$) when the system input $u(k)$ is given. Note that original state-space models are derived when all parameters are perfectly known and system equations based on the mass and energy balance law are listed. The full order of the original models n_x is high even for the small tested networks; see Tab. II for details. However, the system order n_r of identified system generated by data-driven based methods is unknown, and it is the only parameter to be determined by users according to the identification procedures introduced in Section III. We would like to stress that SysID algorithms do not need any other parameters except n_r .

B. Performance of SysID algorithms

1) *Three-node network:* According to the input of each procedure in Section III-A, the input signals (i.e., $u(k)$) for all tested SysID algorithms are shown in Fig. 5. In particular, the first subplot is the input for the ERA method that is required to be an impulse signal; inputs for SIMs and OKID/ERA methods can be any kind of signals, but to ensure the data completeness aforementioned, we use a rectangular signal (the second subplot) for the Three-node network. After the system is identified, a random signal shown in the last subplot is a validation input for identified systems.

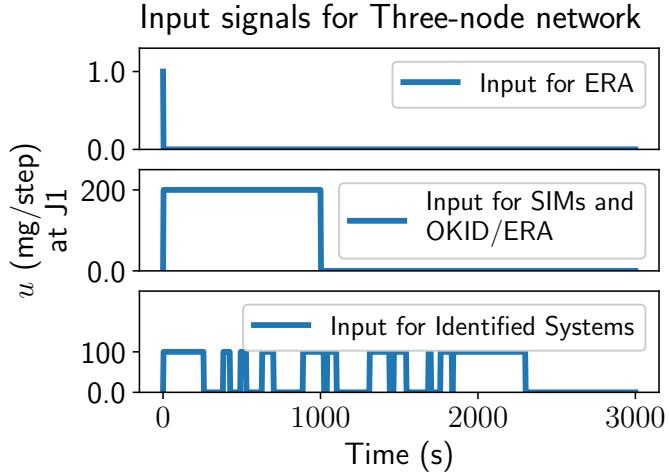


Fig. 5. Input signals for tested SysID algorithms (generated by booster station J1 in Fig. 4a). The first two signals are the inputs for SysID algorithms whereas the last one is for validation of identified systems.

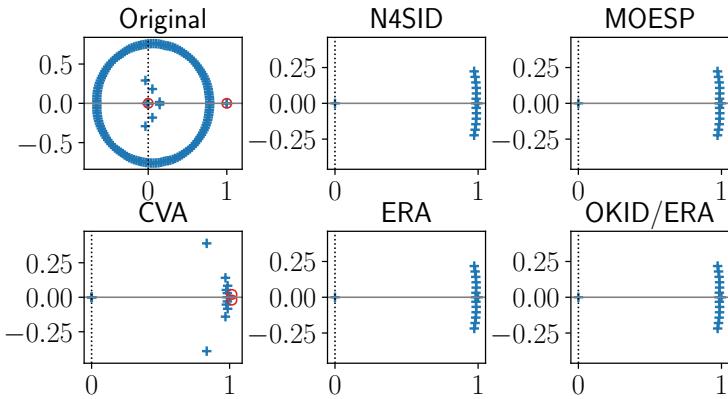


Fig. 6. Poles (blue "+") and zeros (red "o") in complex plane of original and identified systems via various SysID algorithms. The solid/dashed line is the real/imaginary axis of complex plane.

SIMs and ERA-based methods are tested for this Three-node network, and system order n_r is set to 15 for all SysID algorithms. The model properties of the original and identified systems (e.g., stability, system frequencies, and delay) can be implicitly and roughly compared by pole-zero plot (see Fig. 6).

Poles represent the behavior of a dynamical system; zeros on the other hand stands for how the input signal affects the system itself. The locations of the poles and zeros of the original system in Fig. 6 indicate that (i) the original system is stable since all the 154 poles are inside the unit circle (stability property), (ii) the original system has low and high frequencies since the poles appear around the whole unit circle in the complex plane, and (iii) huge delay exists in the original system due to several poles are close to the $(0, 0)$ in the complex plane (high-delay property). The poles and zeros of the identified systems using SIMs and ERA-based methods are shown in Fig. 6. It can be seen that (i) the 15 poles of all identified systems are inside the unit circle and stable except the one from the CVA method, (ii) all SysID methods keep the low frequencies and high frequencies are neglected, and (iii) the high-delay property of the original system is retained by keeping a pole near the $(0, 0)$ in the complex plane.

Another explicit and accurate way to verify the performance is to compare the actual response to the same input validation signal, that is, simulate the network with chlorine injection. We choose to use the same random signal as a validation input (i.e., the last subplot in Fig. 5) for all SysID algorithms (including ERA) such that the results are comparable. To illustrate clearly, the results are grouped by their categories; see Fig. 7. The left column shows the results from SIMs whereas results from ERA-based methods are in the right column. Observed from the results (sensors are installed at J1 and TK3), we conclude that all methods successfully identified the original state-space system (order $n_x = 154$) but with different inaccuracies. The ERA-based methods are more accurate than the SIMs for the Three-node network. In particular, the RMSE of N4SID is 0.0712 which is the smallest while the largest RMSE = 0.323 is from CVA method; the RMSEs for other methods are in Tab. IV.

The corresponding computational time is shown in Tab. III, and the MOESP method takes only 0.21 seconds which is the fastest whereas OKID/ERA takes 3.34 seconds and is the slowest. The computational times indicate the applicability and practicality of

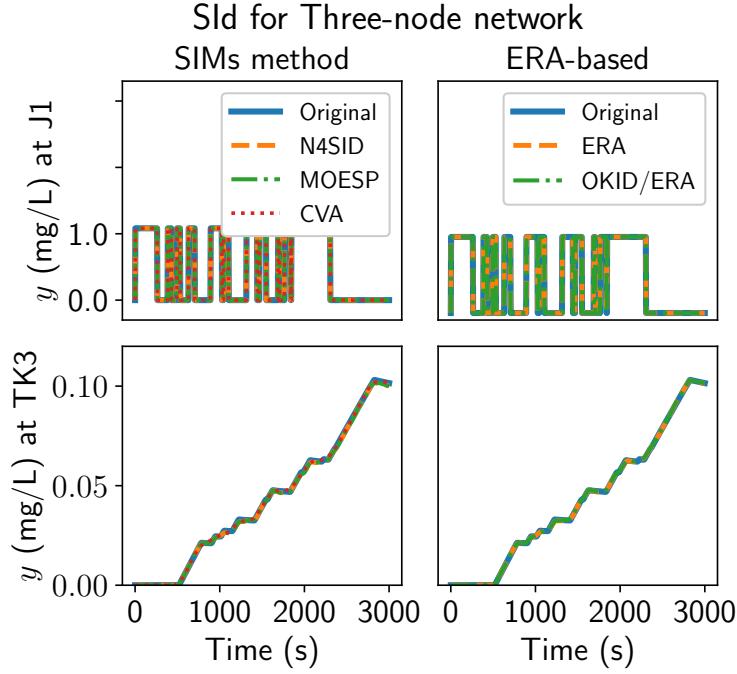


Fig. 7. Output comparisons between various SysID algorithms and original system (sensors are installed at J1 and TK3).

Tab. III. Computational time (in seconds) of various SysID algorithms for all tested networks.

	SIMs			ERA-based	
	N4SID	MOESP	CVA	ERA	OKID/ERA
Three-node	0.21	0.27	0.51	0.78	3.34
Net1	-	-	-	90.00	20.97

using data-driven based methods to identify system models.

2) *Net1 network*: We also perform five SysID algorithms to Net1, but all SIMs fail due to the wrong estimation of Kalman state sequence \mathbf{X}_k or \mathbf{X}_{k+1} . In the end, they cannot ensure stable identified state-space models (i.e., stable \mathbf{A} matrices) when solving the least-square problem (4) for large-scale problems. Hence, only results from ERA-based methods (i.e., ERA and OKID/ERA) are presented in Fig. 8. Instead of presenting the pole-zero map, we show the response to the same input signal directly. For the ERA method, the test input $u(k)$ at J10 is an impulse signal similar to the first figure in Fig. 5, and we do not repeat the figure here. For the OKID/ERA method, the test input at J10 is chosen as a rectangle signal lasting for 600 seconds (see the first subplot in Fig. 8) whereas the validation input is a different rectangle signal lasting for 1200 seconds (see the second subplot in Fig. 8). That is, the booster station injects 80 steps and each step contains 500 mg chlorine (1 step lasts for 15 seconds). The injected chlorine parcels travel along the Net1 and reach J11, J21, and J22. The sensors installed at these junctions receive corresponding concentration values as output measurements y for SysID algorithms (the left column of the last three rows in Fig. 8). Note that there are two paths to reach J22, and that's why two impulses are observed at J22 (see the last row in Fig. 8). Furthermore, the corresponding measurement error y_e is shown in the right column.

From the last three rows in Fig. 8, we conclude that ERA and OKID/ERA successfully identified the water quality models for Net1 since all outputs are overlapping yet with different errors/inaccuracies. The RMSEs for ERA and OKID/ERA are 3.104 and 3.402. Hence, the ERA is more accurate. The system order n_r for ERA and OKID/ERA are 115 and 127 which is much smaller than its original full system order $n_x = 1293$. This is due to only the part of the network (controllable and observable part) is identified according to the discussion we mentioned in Section IV. Furthermore, this water quality model also has the high-delay property, because after the chlorine is injected at J10, and it takes 4680, 7290, and $\{14235, 16860\}^*$ seconds (high-delay) to reach all sensors installed at J11, J21, and J22.

As for the computational time for Net1, the ERA and OKID/ERA take 90.00 seconds and 20.97 seconds indicating OKID/ERA is much faster than ERA. This is due to the way obtaining Markov parameters in the OKID/ERA procedure is faster. With that in mind, the system identification methods and the corresponding state-space matrices generated from the methods in this paper can still be used to perform water quality regulation or monitoring.

*J22 has two impulses detected; hence, two traveling times values exist.

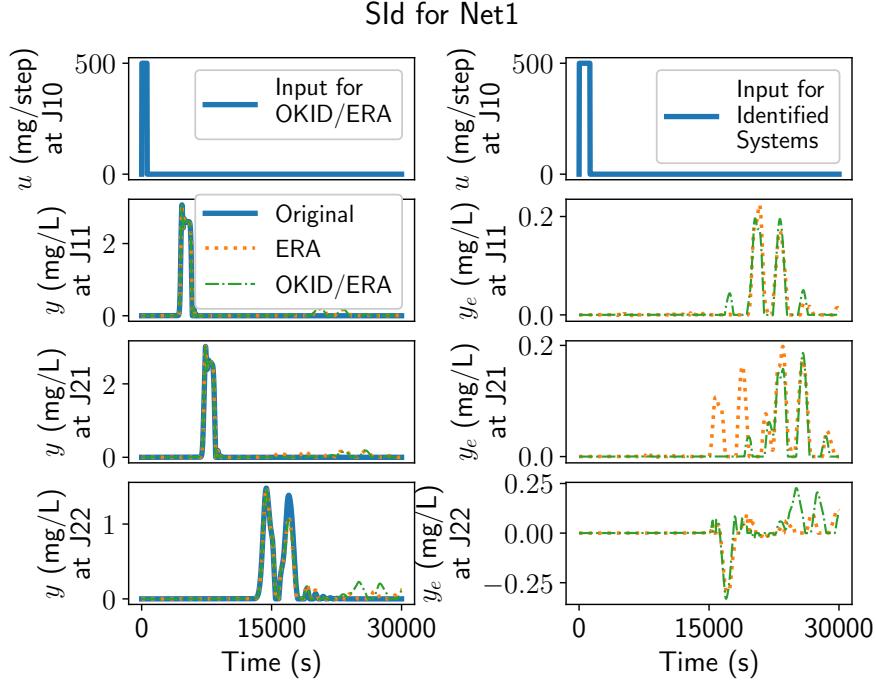


Fig. 8. Test and validation inputs for Net1 and output comparisons between SysID algorithms and original system (see booster station and sensors locations in Tab. II).

Tab. IV. RMSEs of SysID algorithms for Three-node network under different scenarios.

		Three-node		
		Scenario 1	Scenario 2	Scenario 3
SIMs	<i>N4SID</i>	0.0332	0.0712	0.0107
	<i>MOESP</i>	0.0332	0.0516	0.0105
	<i>CVA</i>	0.0183	0.323	—
ERA-based	<i>ERA</i>	0.0224	0.0152	0.0037
	<i>OKID/ERA</i>	0.0234	0.0154	0.0037

C. Possible factors affecting SysID algorithms

From the procedures, we know that there are two main factors determined by users that might have an impact on the final SysID results: system inputs and system order n_r . Hence, we use Three-node network as an example to illustrate the impacts. To this end, three scenarios are designed as

- Scenario 1: u = rectangular signal; $n_r = 15$
- Scenario 2: u = random signal; $n_r = 15$
- Scenario 3: u = rectangular signal; $n_r = 40$.

Note that the input signal is not changed for ERA for all three scenarios since it always requires an impulse signal as input.

The final results are shown in Tab. IV. Compared with Scenarios 1 and 2, we notice that the results are similar indicating the input signal does not impact the results significantly. Compared with Scenarios 2 and 3, all RMSEs are apparently reduced which indicates improving the system order n_r helps to improve the accuracy of SysIDs. However, we also would like to emphasize that the larger n_r causes stability problems for SIMs, that is the SIMs tend to be unstable after n_r is increased to a certain number (see CVA in Tab. IV fails to give a stable A matrix). Even though the phenomena are not observed for the ERA-based methods, we still prefer to keep the n_r as small as possible.

VI. CONCLUSIONS AND FUTURE WORK

Given the thorough tests in the previous section, the following observations are made, thereby answering the posed research questions in Section V:

- A1: Water system operators are able to obtain water quality models of a WDN only utilizing the collected input-output data.
- A2: In general, ERA-based methods are more accurate and stable than SIMs (especially for high-order systems).

- A3: After obtaining the system model, the water system operator can potentially use them to simulate the water quality directly and perform quality control easily using popular control algorithms such as model predictive control.

In summary, this paper explores the applicability of classical system identification methods in water quality dynamics for the first time. Furthermore, the properties of the water quality models are studied and proposed, its ensuing challenges caused by these properties and the corresponding solutions are discussed. The numerical case study shows the performance of each tested SysID algorithm, and the possible factors affecting the results. To make SysID algorithms applicable, a constant decay rate in WDNs and the large-enough hydraulic time-steps of tested networks are assumed so that the tested networks can be viewed as an LTI system. Moreover, zero-initial conditions and zero measurement noises are assumed to meet the requirements of SysID algorithms. Future work will focus on solving the DT-LTV system identification for water quality dynamics considering significant measurement noise, uncertainties, leaks, and multi-species and giving sensor placement strategy when solving for high-delay issue.

DATA AVAILABILITY STATEMENT

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request (e.g., the tested Three-node and Net1 networks and their discrete-time state-space water quality models, the code of SIM, ERA, and OKID-ERA algorithms).

A. Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants 1728629 and 2151392.

REFERENCES

- [1] S. Wang, A. F. Taha, N. Gatsis, L. Sela, and M. H. Giacomoni, "Probabilistic state estimation in water networks," *IEEE Transactions on Control Systems Technology*, vol. 30, no. 2, pp. 507–519, 2022.
- [2] F. Hua, J. West, R. Barker, and C. Forster, "Modelling of chlorine decay in municipal water supplies," *Water research*, vol. 33, no. 12, pp. 2735–2746, 1999.
- [3] R. M. Clark, W. M. Grayman, J. A. Goodrich, R. A. Deininger, and K. Skov, "Measuring and modeling chlorine propagation in water distribution systems," *Journal of water resources planning and management*, vol. 120, no. 6, pp. 871–887, 1994.
- [4] W. M. Grayman, R. M. Clark, and R. M. Males, "Modeling distribution-system water quality; dynamic approach," *Journal of Water Resources Planning and Management*, vol. 114, no. 3, pp. 295–312, 1988.
- [5] M. L. Zierolf, M. M. Polycarpou, and J. G. Uber, "Development and autocalibration of an input-output model of chlorine transport in drinking water distribution systems," *IEEE transactions on control systems technology*, vol. 6, no. 4, pp. 543–553, 1998.
- [6] F. Shang, J. Uber, and M. Polycarpou, "Input-output model of water quality in water distribution systems," in *Building Partnerships*, 2000, pp. 1–8.
- [7] M. M. Polycarpou, J. G. Uber, Z. Wang, F. Shang, and M. Brdys, "Feedback control of water quality," *IEEE Control Systems Magazine*, vol. 22, no. 3, pp. 68–87, 2002.
- [8] Z. Wang, M. M. Polycarpou, J. G. Uber, and F. Shang, "Adaptive control of water quality in water distribution networks," *IEEE transactions on control systems technology*, vol. 14, no. 1, pp. 149–156, 2005.
- [9] K. Duzinkiewicz, M. Brdys, and T. Chang, "Hierarchical model predictive control of integrated quality and quantity in drinking water distribution systems," *Urban Water Journal*, vol. 2, no. 2, pp. 125–137, 2005.
- [10] S. Wang, A. F. Taha, and A. A. Abokifa, "How effective is model predictive control in real-time water quality regulation? state-space modeling and scalable control," *Water Resources Research*, p. e2020WR027771, 2020.
- [11] W. Favoreel, B. De Moor, and P. Van Overschee, "Subspace state space system identification for industrial processes," *Journal of process control*, vol. 10, no. 2-3, pp. 149–155, 2000.
- [12] J. N. Juang and R. S. Pappa, "An eigensystem realization algorithm for modal parameter identification and model reduction," *Journal of Guidance, Control, and Dynamics*, vol. 8, no. 5, pp. 620–627, 1985.
- [13] G. Horta, W. Longman, and R. W. L. Jer-Nan Juang, Minh Phan, Lucas G. Horta, "Identification of Observer / Kalman Filter Markov Parameters : Theory," *Nasa Technical Memorandum 104069*, 1991.
- [14] C. D. Schaper, W. E. Larimore, D. E. Seborg, and D. A. Mellichamp, "Identification of chemical processes using canonical variate analysis," in *29th IEEE Conference on Decision and Control*. IEEE, 1990, pp. 605–610.
- [15] P. Schmid, "Dynamic mode decomposition of numerical and experimental data," *Journal of Fluid Mechanics*, vol. 656, pp. 5–28, 2008.
- [16] S. L. Brunton, J. L. Proctor, J. N. Kutz, and W. Bialek, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 15, pp. 3932–3937, 2016.
- [17] S. Vijayshankar, S. Nabi, A. Chakrabarty, P. Grover, and M. Benosman, "Dynamic mode decomposition and robust estimation: Case study of a 2d turbulent boussinesq flow," in *2020 American Control Conference (ACC)*. IEEE, 2020, pp. 2351–2356.
- [18] I. Kamwa and L. Gerin-Lajoie, "State-space system identification-toward mimo models for modal analysis and optimization of bulk power systems," *IEEE Transactions on Power Systems*, vol. 15, no. 1, pp. 326–335, 2000.
- [19] Z. Y. Shi, S. S. Law, and H. N. Li, "Subspace-based identification of linear time-varying system," *AIAA Journal*, vol. 45, no. 8, pp. 2042–2050, 2007.
- [20] P. Van Overschee and B. De Moor, "N4sid: Subspace algorithms for the identification of combined deterministic-stochastic systems," *Automatica*, vol. 30, no. 1, pp. 75–93, 1994.
- [21] M. Verhaegen, "Identification of the deterministic part of mimo state space models given in innovations form from input-output data," *Automatica*, vol. 30, no. 1, pp. 61–74, 1994.
- [22] W. E. Larimore, "Canonical variate analysis in identification, filtering, and adaptive control," in *29th IEEE Conference on Decision and control*. IEEE, 1990, pp. 596–604.
- [23] P. Van Overschee and B. De Moor, "A unifying theorem for three subspace system identification algorithms," *Automatica*, vol. 31, no. 12, pp. 1853–1864, 1995.
- [24] J.-N. Juang and R. S. Pappa, "Effects of noise on modal parameters identified by the eigensystem realization algorithm," *Journal of Guidance, Control, and Dynamics*, vol. 9, no. 3, pp. 294–303, 1986.
- [25] J. N. Juang, M. Phan, L. G. Horta, and R. W. Longman, "Identification of observer/kalman filter markov parameters - theory and experiments," *Journal of Guidance, Control, and Dynamics*, vol. 16, no. 2, pp. 320–329, 1993.
- [26] J.-N. Juang, J. Cooper, and J. Wright, "An eigensystem realization algorithm using data correlations (era/dc) for modal parameter identification," 1987.

- [27] F. Vicario, M. Q. Phan, R. Betti, and R. W. Longman, “OKID as a unified approach to system identification,” *Advances in the Astronautical Sciences*, vol. 152, pp. 3443–3460, 2014.
- [28] C. W. Rowley, I. Mezi, S. Bagheri, P. Schlatter, and D. S. Henningson, “Spectral analysis of nonlinear flows,” *Journal of Fluid Mechanics*, vol. 641, no. Rowley 2005, pp. 115–127, 2009.
- [29] J. L. Proctor, S. L. Brunton, and J. N. Kutz, “Dynamic mode decomposition with control,” *SIAM Journal on Applied Dynamical Systems*, vol. 15, no. 1, pp. 142–161, 2016.
- [30] Z. Bai, E. Kaiser, J. L. Proctor, J. N. Kutz, and S. L. Brunton, “Dynamic mode decomposition for compressive system identification,” *AIAA Journal*, vol. 58, no. 2, pp. 561–574, 2020.
- [31] S. L. Brunton, J. L. Proctor, J. H. Tu, and J. N. Kutz, “Compressed sensing and dynamic mode decomposition,” *Journal of computational dynamics*, vol. 2, no. 2, p. 165, 2015.
- [32] Y. Zheng and N. Li, “Non-asymptotic Identification of Linear Dynamical,” pp. 1–22, 2020.
- [33] A. Tsiamis and G. J. Pappas, “Linear Systems can be Hard to Learn,” 2021. [Online]. Available: <http://arxiv.org/abs/2104.01120>
- [34] L. Ljung, “Perspectives on system identification,” *Annual Reviews in Control*, vol. 34, no. 1, pp. 1–12, 2010. [Online]. Available: <http://dx.doi.org/10.1016/j.arcontrol.2009.12.001>
- [35] ———, *System Identification: An Overview*. London: Springer London, 2015, pp. 1443–1458. [Online]. Available: https://doi.org/10.1007/978-1-4471-5058-9_100
- [36] S. J. Qin, “An overview of subspace identification,” *Computers and Chemical Engineering*, vol. 30, no. 10-12, pp. 1502–1513, 2006.
- [37] G. Favier, “An overview of system modeling and identification,” 2010.
- [38] M. Verhaegen, “Identification of the deterministic part of MIMO state space models given in innovations form from input-output data,” *Automatica*, vol. 30, no. 1, pp. 61–74, 1994.
- [39] S. L. Brunton, J. L. Proctor, and J. N. Kutz, “Compressive sampling and dynamic mode decomposition,” 2013. [Online]. Available: <http://arxiv.org/abs/1312.5186>
- [40] N. Benjamin Erichson, L. Mathelin, J. Nathan Kutz, and S. L. Brunton, “Randomized dynamic mode decomposition,” *SIAM Journal on Applied Dynamical Systems*, vol. 18, no. 4, pp. 1867–1891, 2019.
- [41] L. A. Rossman *et al.*, “EPANET 2: users manual,” 2000.
- [42] F. Shang, J. G. Uber, and M. M. Polycarpou, “Particle backtracking algorithm for water distribution system analysis,” *Journal of environmental engineering*, vol. 128, no. 5, pp. 441–450, 2002.