Modeling, Identification and Control, Vol. 44, No. 2, 2023, pp. 69–82, ISSN 1890–1328

Subspace System Identification of a Pilot Tunnel System of a Combined Sewage System

Yongjie Wang   Finn Aakre Haugen

Faculty of Technology, Natural Sciences and Maritime Sciences. Department of Electrical Engineering, Information Technology and Cybernetics. University of South-Eastern Norway (USN), Porsgrunn, Norway. E-mail: {Yongjie.Wang, Finn.Haugen}@usn.no

Abstract

This paper presents an investigation into the potential use of subspace identification methods (SIMs) for model-based control of urban drainage systems (UDS) that play a crucial role in collecting and transporting stormwater runoff and domestic sewage to Water Resource Recovery Facilities (WRRF) in urban areas. To evaluate the feasibility of level control using model-based algorithms in UDS, a pilot tunnel system was constructed. Three linear state-space models were identified using the system identification toolbox in MATLAB and an open-source module in Python named SIPPY. The study finds that the identified models can predict the system output with acceptable accuracy thus for model-based control of the system. The findings of this study aim to contribute to the development of more efficient and effective control strategies for UDS.

Keywords: Urban drainage systems, subspace system identification, drainage tunnel, model-based control, Saint-Venant equations.

1 Introduction

Urban drainage systems (UDS) play an important role in maintaining a healthy and safe living environment in cities, collecting and transporting stormwater runoff and domestic sewage to wastewater treatment plants.

This work aims to study the mathematical modeling process of a pilot tunnel system, which is a simulator of an existing tunnel system in Norway. The actual tunnel is 42 km long and transports a total volume of up to 110 million m^3/year combined sewage overflow (CSO) to a Water Resource Recovery Facility (WRRF) named VEAS. A simplified version of the tunnel was designed and constructed to investigate the performance of real-time control (RTC) algorithms, for example, model-based predictive control (MPC) of the system (Lund et al., 2018; Aakre Haugen, 2018; Ocampo-Martinez et al., 2013; Breckpot et al., 2012). The goal of this work is to present the system identification process of the pilot tunnel system.

The pilot tunnel system can be viewed as an open channel flow. The mathematical modeling of the flow rate, velocity, depth, etc. of the flow in such a system can be derived from the conservation of mass and momentum (Chanson, 2004; Litrico and Fromion, 2009). A set of partial differential equations (PDEs) named Saint-Venant equations (SVE) for a one-dimensional open channel flow case such as in the pilot tunnel system is given in (1)-(2).

$$\frac{\partial a}{\partial t} + \frac{\partial q}{\partial x} = 0 \tag{1}$$

$$\frac{\partial q}{\partial t} + \frac{\partial}{\partial x}\left(\frac{q^2}{a}\right) + ga\frac{\partial h}{\partial x} + g(S_f - S_0) = 0 \tag{2}$$

In the SVE model, $h(x,t)$ [mm] is the depth from the free surface to the bed. $a(x,t)$ [mm^2] is the wetted area. $q(x,t)$ [L/s] is the discharge across section a. S_0

doi:10.4173/mic.2023.2.3

is the bed slope. S_f is the friction slope. t [s] is the time axis. g [mm/s^2] is the gravitational acceleration constant.

To facilitate the simulation of the SVE model for prediction purposes, various numerical methods can be utilized(Kurganov and Levy, 2002; Kamboh et al., 2016; Kurganov and Petrova, 2007). However, the accuracy of the simulation results is typically dependent on a high computational load, which can pose a challenge for RTC applications with short time steps. Additionally, for model-based control algorithms like MPC, multiple simulations are often needed for each optimization iteration, which can further lead to significant difficulties for the optimization solver. To simplify the model for real-time control applications, techniques such as model deduction and linearization can be applied to the SVE model (Xu et al., 2011, 2012; Cen et al., 2010).

This work presents the modeling process of the system using subspace identification methods and the performance evaluation of the identified models for predictive control applications.

The content is organized as follows: The pilot tunnel system and the system identification approach are described in Section 2, followed by an introduction to how to implement identified models for model-based control algorithms such as MPC. The identification results are presented and the prediction performance of different identified models are discussed in Section 3, yielding conclusions in Section 4.

2 Methods and materials

This section presents the pilot tunnel system and the modeling of the system.

2.1 Pilot tunnel system

Figure 1 and Figure 2 show the overview and a photo of the pilot tunnel system, respectively. The system consists of two sections, V1 and V2, with one inlet each. The outflow of V1 and V2 can be manipulated by a pump and a movable weir, respectively. Both sections are made of transparent acrylic circular tubes with inner diameter $d = 94$ mm, length $l = 4000$ mm, and slope $S_0 = 85$ mm/4000 mm $\approx 1.2°$.

This work focuses on the V1 section of the tunnel. The water level at the end of V1 is manipulated by a pump (Main Pump, MP) on the left end. V1 has one inlet at the right end and the flow rate is controlled by an inlet pump 1 (PV1). The left end is lower than the right end so that the inflow from PV1 can flow freely to the MP end. Both MP and PV1 are 12V DC pumps that can be manipulated using PWM (Pulse Width Modulation) from 0 to 100% duty cycle with a fixed pulse frequency 50 Hz. The water level is measured using a pressure sensor with a range of 0∼50 mbar, equivalent to a maximum 50 cm level measurement.

2.2 System identification

This section outlines the system identification problem and various methods that can be adopted, followed by a practical approach to implementing the identified models for model-based control applications.

2.2.1 State-space model

A general dynamic system with deterministic inputs and outputs is depicted in Figure 3.

In the system, $y_k \in \mathbb{R}^{n_y}, x_k \in \mathbb{R}^n, u_k \in \mathbb{R}^{n_u}, w_k \in \mathbb{R}^n, v_k \in \mathbb{R}^{n_y}$ are the system outputs, states, control inputs, process noise, and measurement noise vectors at discrete time index $k = 1, 2, ...$, respectively. n, n_y, n_u are integers representing the system order, number of outputs, and number of inputs, respectively.

The process can be represented using a variety of models. For example, Igreja et al. (2011) implemented distributed MPC based on a linear state-space form of an ARX (Auto Regressive Exogenous) model for controlling water level in a canal. Yang and Chang (2005) trained a neural networks model to estimate the velocity profile and discharge in an open channel flow using experimental data. A few other examples of system models are transfer functions, input-output polynomial models, state-space models (SSMs), neural networks, etc.

In this work, a linear time-invariant (LTI) SSM as in (3)-(4) is utilized to model the system:

$$x_{k+1} = Ax_k + Bu_k + w_k \quad (3)$$

$$y_k = Cx_k + Du_k + v_k \quad (4)$$

where, $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times n_u}, C \in \mathbb{R}^{n_y\times n}, D \in \mathbb{R}^{n_y\times n_u}$ are the system matrix, input matrix, output matrix, and feedthrough matrix. For LTI systems, A, B, C, D are constant matrices.

The covariance matrices of the noise sequences w_k, v_k are defined in appropriate shapes as in (5):

$$\mathbb{E}\left[\begin{pmatrix} w_p \\ v_p \end{pmatrix}\begin{pmatrix} w_q^T & v_q^T \end{pmatrix}\right] = \begin{pmatrix} Q & S \\ S^T & R \end{pmatrix}\delta_{pq} \geq 0 \quad (5)$$

2.2.2 System identification

Van Overschee and De Moor (1996) suggest that the system identification aims to:

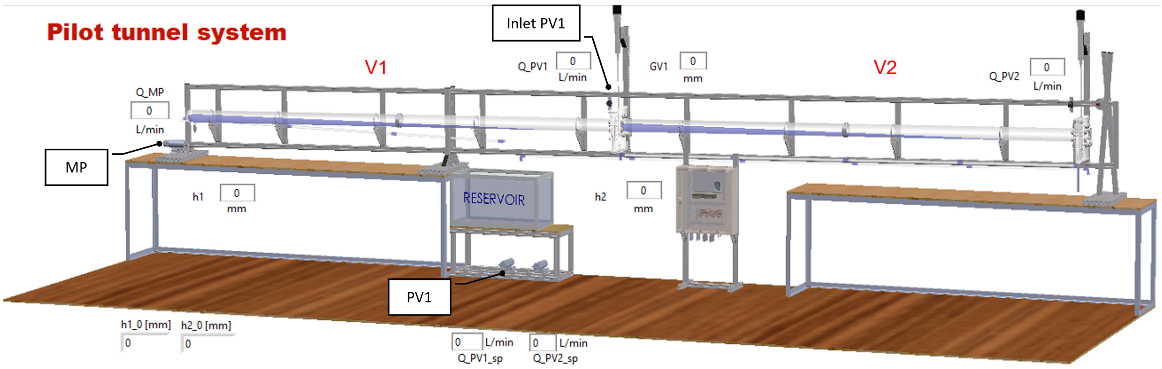

Figure 1: Overview of the pilot tunnel system

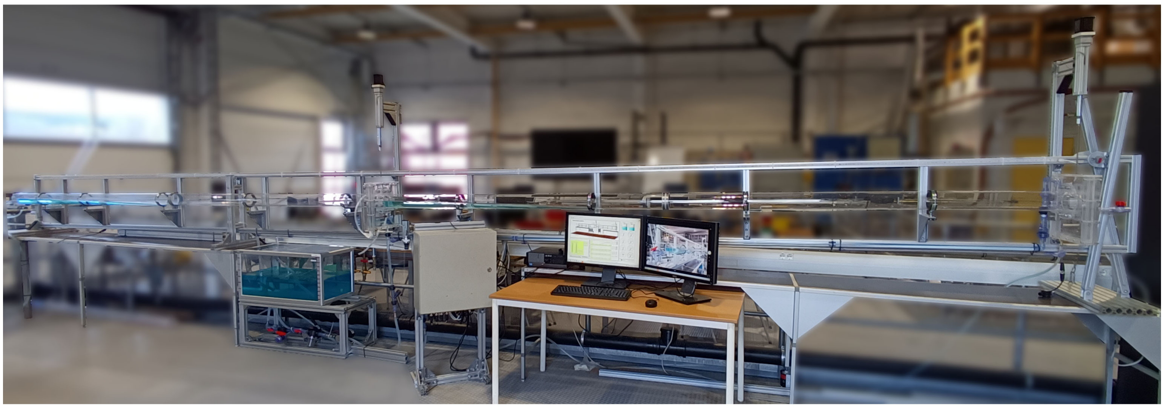

Figure 2: A photo of the pilot tunnel system

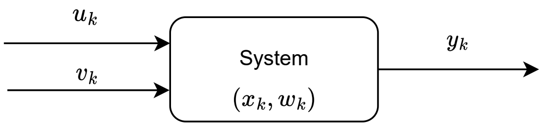

Figure 3: A dynamic system

> Given a set of input and output measurements $u_1, u_2, ..., u_s$, and $y_1, y_2, ..., y_s$, determine an appropriate system order n and estimate the system matrices A, B, C, D, Q, R, S.

This process is based on the following assumptions:

- The matrix pair A, C is assumed observable, indicating that all the modes in the system can be observed in the output y_k and identified.
- The matrix pair $A, [B \quad Q^{1/2}]$ is controllable, implying that all the modes of the system are exited by either the deterministic input u_k and/or the stochastic noise w_k.

The process and measurement noise are assumed to be white and zero-mean in the following sections, giving zeros in S.

Identification procedures

In general, the process of identification involves the following tasks (Ljung, 1998):

- Collection and pre-processing data obtained from, i.e., real experiments. A common practice is to split one experimental data set into two segments, one for model parameter estimation and the other one for validation. Data from different experiments can also be used.
- Selection of a set of candidate models.
- Definition of a criterion of fit, i.e., the sum of norms of the prediction error.
- Validation of the identified models. The identified models must be tested using experimental data sets that are different from the estimation process. Muroi and Adachi (2015) compared several validation criteria for system identification in the

time domain, such as fit ratio, correlation coefficient, index of agreement, etc. In MATLAB® System Identification Toolbox (MATLAB, 2023), MSE (mean squared error), NMSE (normalized MSE), or NRMSE (normalized root mean squared error) can be selected as the fitness criteria. See Eq. (6)-(8) for details. In addition, Ljung (1999) suggested a residual analysis consisting of a whiteness test of the residuals and an independence test between the residuals and the past inputs.

$$\mathrm{MSE} = \frac{1}{N}\sum_{k=1}^{N}(y_k - \hat{y}_k)^2 \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(y_k - \hat{y}_k)^2} = \sqrt{\mathrm{MSE}} \tag{7}$$

$$\mathrm{NRMSE} = \frac{\sqrt{\sum_{k=1}^{N}(y_k - \hat{y}_k)^2}}{\sqrt{\sum_{k=1}^{N}(y_k - \bar{y})^2}} = \frac{\mathrm{RMSE}}{\sigma_y} \tag{8}$$

where, N is the number of samples. y_k and $\hat{y}_k$ represent the measured and the corresponding predicted output at time step k, respectively. $\bar{y}$ and σ_y stand for the average and the standard deviation of the measured output, respectively.

In addition to the above criteria, the *compare()* function in MATLAB utilizes the *FitPercent* or fit ratio (FIT) in Eq. (9) to compare different models' quality (MathWorks, 2023; Muroi and Adachi, 2015).

$$\mathrm{FIT} = (1 - \mathrm{NRMSE}) \times 100\% \tag{9}$$

Data collection and processing

System identification relies on input and output data from either simulations or real experiments. The quality of the identified model is highly dependent on the design of the experiment and the data quality.

Open-loop or closed-loop can be adopted for system identification experiments. Using a closed-loop approach the system states can be safely under control. However, the approach has a fundamental issue of correlation between the noise and the input, making subspace methods such as CVA, N4SID, and MOESP biased. In contrast, the open-loop approach assumes the data are independent of the past noise (Qin, 2006). To implement an open-loop experiment, the inputs should excite the system dynamics adequately. See (Ljung, 1999, p. 412) for details about the *persistence of excitation*. A disadvantage of open-loop experiments is that the system can be at risk of being driven away from its operating point and thus becomes unsafe, for example, an empty/ full tunnel system.

Pre-processing of the raw data sets including filtering, removing trend and offset, and resampling should be applied when necessary.

Subspace Identification Methods (SIMs)

SIMs identify directly state-space models of the system using known input-output data pairs.

A variety of SIMs have been developed for identifying the system model, such as traditional methods (N4SID, MOESP, and CVA) (Qin, 2006) and parsimonious methods (PARSIM-K(Pannocchia and Calosi, 2010), PARSIM-P, PARSIM-S (Qin et al., 2005), and PARSIM-E (Qin and Ljung, 2003)) and DSR (Di Ruscio, 1997), etc. Detailed mathematical derivation and expression of these methods can be found in literature such as (Qin, 2006; Ljung, 1999; Di Ruscio, 1997).

Although the detailed procedures may differ, there are common steps involved in system identification, such as regression or projection, model reduction, parameter estimation, and iteration. Overall, three discrete representations of the system models used in SIMs are (Qin, 2006):

- Process form:

$$x_{k+1} = \hat{A}x_k + \hat{B}u_k + w_k \tag{10}$$
$$y_k = \hat{C}x_k + \hat{D}u_k + v_k \tag{11}$$

where, $\hat{A}$ is the identified system matrix, which can be seen as an estimate of the "true" matrix. The "ˆ" symbol is normally ignored. The same applies to the other matrices.

- Innovation form:

$$\hat{x}_{k+1} = A\hat{x}_k + Bu_k + Ke_k \tag{12}$$
$$\hat{y}_k = C\hat{x}_k + Du_k + e_k \tag{13}$$

where, K is the steady-state *Kalman filter gain* matrix, which can be obtained by solving the (discrete) algebraic Ricatti equation (Qin, 2006). $\hat{x}_k$ and $\hat{y}_k$ are the (optimal) estimates of the process states and output. $e_k = y_k - \hat{y}_k$ is the prediction error. Again, normally the symbol "ˆ" is ignored.

- Predictor form:

$$\bar{x}_{k+1} = A_K\bar{x}_k + B_Ku_k + Ky_k \tag{14}$$
$$\bar{y}_k = C\bar{x}_k + Du_k + e_k \tag{15}$$

with relations:

$$A_K = A - KC \tag{16}$$
$$B_K = B - KD \tag{17}$$

Table 1: Comparison of identification software packages.

	MATLAB	**SIPPY**
Subspace models	N4SID, CAV, MOESP	N4SID, MOESP, CVA; PARSIM-K, PARSIM-P, PARSIM-S.
Initial state x_0 estimate	Yes.	Only for parsimonious algorithms.
Model type	Continuous and discrete	Discrete
Data pre-processing	Detrend, input/output offset, resample, merge experiments.	Input/output offset.

These three forms serve different purposes in system identification and implementation of the system model.

Unlike the SVE model in (1)-(2), the obtained models typically are black-box models, meaning that the state vector x_k lacks physical meaning in most cases.

Toolboxes available

Several toolboxes are available for system identification, such as the MATLAB® system identification toolbox (MATLAB, 2023), DSR toolbox (MATLAB) (Di Ruscio, 1997) and SIPPY (Armenise et al., 2018). SIPPY is an open-source package for system identification, which requires several dependencies to access the full functionality. For example, solving the discrete Riccati equation requires Scipy/Slycot package.

Table 1 presents a brief comparison of these software packages for subspace identification methods.

The system order n is decided by the user, depending on the complexity of the system and acceptance criterion. Software packages in Table 1 offer default orders if not specified.

Estimation of the time delay, τ, between the control action and the actual process response, is recommended before estimating other parameters as it determines how to align the input-output data (Di Ruscio, 2001; Ljung, 1998). The MATLAB function *delayest()* can help to determine the time delay if no prior information is available (MATLAB, 2023).

2.3 Implementation of identification in this work

In this work, open-loop experiments are to be conducted using different excitation such as PRBS (Pseudo-random binary sequence), and chirp-type signals. A first-order low-pass filter with filter time-constant $T_f = 4$ seconds is to be applied to the raw measurements to reduce the measurement noise. In addition, all data sets are resampled with a fixed time step $T_s = 0.5$ s.

As presented, the pilot tunnel system (V1) has:

- One output variable: The water level y [mm].
- One controlled input variable: The main pump control signal u^{MP} [%].
- One inlet flow: Q^{PV1} [L/min]. This is a disturbance to the real system but will be treated as an input variable in this case. The value is normally unknown and should be obtained from estimation or forecasting. In this work, the inlet flow is assumed known until the current time step. u^{PV1} [%] will be used instead of Q^{PV1} since the inlet pump is controlled by control signal u^{PV1}.

From a system identification point of view, the pilot tunnel system is a MISO (Multi-Input-Single-Output) system with $n_y = 1, n_u = 2$. Letting $u_k = [u_k^{\text{MP}}, u_k^{\text{PV1}}]^T$, the system model (3) can be rewritten as:

$$x_{k+1} = Ax_k + B^{\text{MP}}u_k^{\text{MP}} + B^{\text{PV1}}u_k^{\text{PV1}} + w_k \tag{18}$$

$$= Ax_k + \underbrace{\begin{bmatrix} B^{\text{MP}} & B^{\text{PV1}} \end{bmatrix}}_{B} \underbrace{\begin{bmatrix} u_k^{\text{MP}} \\ u_k^{\text{PV1}} \end{bmatrix}}_{u_k} + w_k \tag{19}$$

Two algorithms, N4SID (in MATLAB and SIPPY) and PARSIM-K (in SIPPY) will be tested and compared. The following parameters are found to be appropriate according to preliminary tests and analysis, thus used for all identification algorithms:

- Model order $n = 2$.
- Time delay $\tau^{\text{MP}} = 2 \times T_s = 1$ second and $\tau^{\text{PV1}} = 15 \times T_s = 7.5$ seconds.
- Forward- and backward-prediction horizons for subspace algorithms: $S_f = S_p = 15$ samples.
- No feedthrough. Therefore for all models:

$$D = \begin{bmatrix} 0 & 0 \end{bmatrix}$$

The Kalman filter gain matrix K will be obtained as part of the system identification results in software packages shown in Table 1.

Eq. (6)-(9) will be adopted for validation.

2.4 Using identified models for Model-based control

After the system model is obtained, it can be used for the design of real-time control algorithms, such as MPC for water level reference tracking, in which the model prediction of the process forms the basis of the optimization problem (Di Ruscio and Foss, 1998).

2.4.1 Optimal prediction from identified models

For a prediction horizon $j = 1, 2, ..., M$, the prediction of the system states and output can be obtained by simulating the process form in (10)-(11), by omitting the (white) noise terms, resulting (20)-(21):

$$\bar{x}_{k+j+1} = A\bar{x}_{k+j} + Bu_{k+j} \tag{20}$$

$$\bar{y}_{k+j} = C\bar{x}_{k+j} + Du_{k+j} \tag{21}$$

where, the system matrices A, B, C, and D are obtained from the identification. The input sequence is $u_{k+j} = [u^{\text{MP}}_{k+j}, u^{\text{PV1}}_{k+j}]^T$ with u^{MP}_{k+j} to be obtained from solving the MPC optimization problem at every time step.

To validate the model performance, iterative simulations of (20)-(21) can be done. In addition, functions such as *sim*() in MATLAB or *Scipy.signal.dlsim()* (SciPy.org, 2023) can also be used. The time delay of the inputs τ must be taken into account.

2.4.2 Initial conditions

Simulating (20)-(21) requires the initial state vector (at the first time step in each prediction horizon) $\bar{x}_{k+1}$, which is unknown but can be estimated using the innovation form and the present measurement and states:

$$\begin{aligned} \bar{x}_{k+1} &= \hat{x}_{k+1} && (22)\\ &= A\hat{x}_k + Bu_k + Ke_k && (23)\\ &= A\hat{x}_k + Bu_k + K(y_k - \hat{y}_k) && (24)\\ &= A\hat{x}_k + Bu_k + K(y_k - (C\hat{x}_k + Du_k)) && (25) \end{aligned}$$

All terms on the right-hand side in (25) are available up to time step k. The initial state $\hat{x}_0$ can be estimated or assumed to be zeros.

In this approach (known as *output feedback*), achieving feedback of the process output y_k is realized through model prediction utilizing the estimated process states. A state observer such as the Kalman gain K plays an essential role in this task.

2.4.3 M-step-ahead prediction

The 1-step-ahead prediction is the forced response of the system under u_{k+1}, as expressed in Eq. (26):

$$\begin{aligned} \bar{y}_{k+1} =& C\bar{x}_{k+1} + Du_{k+1} && (26)\\ =& C(A - KC)\bar{x}_k + CKy_k + \\ & + C(B - KD)u_k + Du_{k+1} && (27) \end{aligned}$$

Once new process measurements are available, the procedure, including estimating $\bar{x}_{k+1}$ from the updated output and recalculating the model prediction, is repeated.

For a system with no feedthrough (strictly proper system), Eq. (28) is the M-step-ahead prediction expressed in a matrix form, given current state $\hat{x}_k$, current input u_k, and future inputs u_{k+j}. See more details in (Di Ruscio, 2001, p. 150-151).

$$\bar{\boldsymbol{Y}}_k = \mathcal{H}^x(A - KC)\hat{x}_k + \mathcal{H}^y K y_k + \mathcal{H}^u \boldsymbol{U}_k \tag{28}$$

where,

$$\bar{\boldsymbol{Y}}_k = \begin{bmatrix} \bar{y}_{k+1} \\ \bar{y}_{k+2} \\ \vdots \\ \bar{y}_{k+M} \end{bmatrix}, \boldsymbol{U}_k = \begin{bmatrix} u_k \\ u_{k+1} \\ \vdots \\ u_{k+M-1} \end{bmatrix},$$

$$\mathcal{H}^x = \mathcal{H}^y = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{M-1} \end{bmatrix},$$

$$\mathcal{H}^u = \begin{bmatrix} CB & 0 & \dots & 0 \\ CAB & CB & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ CA^{M-1}B & CA^{M-2}B & \dots & CB \end{bmatrix}$$

According to (28), the predicted values $\bar{y}_{k+1}, \bar{y}_{k+2}, \dots, \bar{y}_{k+M}$ can all be related to $\hat{x}_k$ and y_k. Therefore, the estimation of the process states $\hat{x}_k$ deserves an investigation. Note that predicting $\bar{y}_{k+M}$ requires inputs up to u_{k+M-1}, as there is always one step of delay from the input to the output in RTC.

3 Results and discussion

A few experiments were conducted to examine the system responses under two different types of excitation signals. The results of identification and validation are presented and discussed.

3.1 Data sets for identification

The data set for model parameter estimation is presented in Figure 4 and the validation data sets are shown in Figure 5 and 6, respectively. Linear SSMs are not able to model the system offset. Therefore, subtracting the operating equilibrium or the sample mean from offline transient data is recommended (Ljung, 1999, pp. 458-460). In practice, this can be done manually or in MATLAB by using "*OutputOffset*" and in SIPPY by setting "*centering= InitVal*". In this work, instead of raw values y_{raw}, the modeling will use the low-pass filtered data y_{LP}, with output offset applied:

$$y_{\text{id}} = y_{\text{LP}} - y_{\text{LP}}[0] \tag{29}$$

where $y_{\text{LP}}[0]$ is the first element in the filtered output sequence. In fact, all experiments started around the operating point $h_1^* = 60$ mm. The same applied to the validation data sets.

Note that a few step changes of u^{MP} and u^{PV1} were applied to avoid a full/empty tunnel. This is a disadvantage of conducting open-loop experiments as discussed in Section 2.2.2.

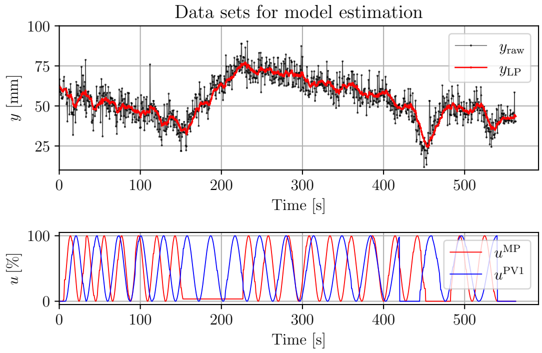

Figure 4: Data sets for model estimation. $u = [u^{\text{MP}}, u^{\text{PV1}}]^T$: Input; y_{raw}: Raw output; y_{LP}: Low-pass filtered output.

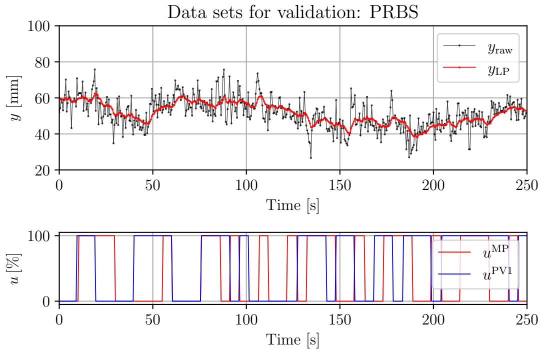

Figure 5: Data sets for validation using PRBS excitation. $u = [u^{\text{MP}}, u^{\text{PV1}}]^T$: Input; y_{raw}: Raw output; y_{LP}: Low-pass filtered output.

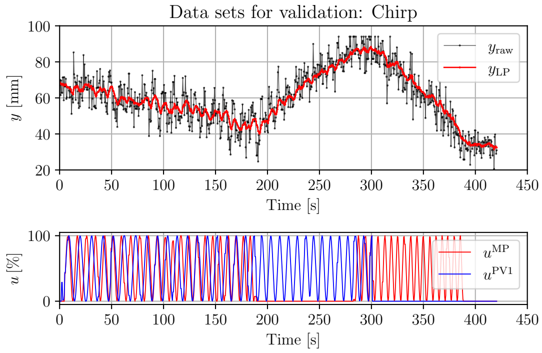

Figure 6: Data sets for validation using chirp excitation. $u = [u^{\text{MP}}, u^{\text{PV1}}]^T$: Input; y_{raw}: Raw output; y_{LP}: Low-pass filtered output.

3.2 Identified model parameters

Table 2 presents the parameters of three models.

3.3 Validation of models

With $\hat{x}_0 = [0, 0]^T$, Figure 7 presents the validation results and Table 3.3 compares the model fits.

Overall, the results in Figure 7 and Table 3.3 show that SSM-1 presented the best fit with the smallest MSE and highest FIT ($\approx$ 80%) under both types of excitation. As a comparison, SSM-2 and SSM-3 predict the process output equally well, yet worse than SSM-1, particularly in the long-term horizon.

All models performed well in predicting the variations with relatively lower frequencies in the data sets. However, none of the three caught the system dynamics with relatively higher frequencies. This is partially due to the relatively small system order ($n = 2$) used, in addition to the nonlinear dynamics that are challenging for linear models to catch.

When the system was driven too far away from the operating point, the performance of all models deteriorated. For instance, in the chirp excitation case, the identified models failed to capture the free response of the system after both pumps were deactivated at $t =$ 380 s, even though SSM-1 still performed the best.

3.4 Prediction using identified models

The system models are intended for deployment in model-based control of the pilot tunnel system. In this section, more detailed prediction performances of the identified models were evaluated, followed by a discussion on state estimation. Figure 8 shows the results with prediction horizons of 20 seconds.

Table 2: Identified discrete-time SSMs of the pilot tunnel system

Parameter	SSM-1 (MATLAB, N4SID)	SSM-2 (SIPPY, N4SID)	SSM-3 (SIPPY, PARSIM-K)
A	$\begin{bmatrix} 0.9999 & 0.0196 \\ -0.0034 & 0.9136 \end{bmatrix}$	$\begin{bmatrix} 0.9929 & -0.0566 \\ -0.0127 & 0.9046 \end{bmatrix}$	$\begin{bmatrix} 0.9953 & -0.0596 \\ -0.0071 & 0.9043 \end{bmatrix}$
B	$\begin{bmatrix} -6.99\text{E}-6 & 1.42\text{E}-5 \\ -3.88\text{E}-5 & -8.49\text{E}-6 \end{bmatrix}$	$\begin{bmatrix} 1.73\text{E}-4 & -3.52\text{E}-5 \\ -5.79\text{E}-5 & 2.36\text{E}-4 \end{bmatrix}$	$\begin{bmatrix} -2.54\text{E}-5 & -2.49\text{E}-5 \\ 1.37\text{E}-4 & -1.86\text{E}-4 \end{bmatrix}$
C	$\begin{bmatrix} 393.3 & -2.818 \end{bmatrix}$	$\begin{bmatrix} -36.1534 & -4.4308 \end{bmatrix}$	$\begin{bmatrix} 66.5356 & 1.2653 \end{bmatrix}$
K	$\begin{bmatrix} 0.00225 \\ -0.00499 \end{bmatrix}$	$\begin{bmatrix} -0.00816 \\ -0.00115 \end{bmatrix}$	$\begin{bmatrix} 0.01443 \\ 0.00807 \end{bmatrix}$

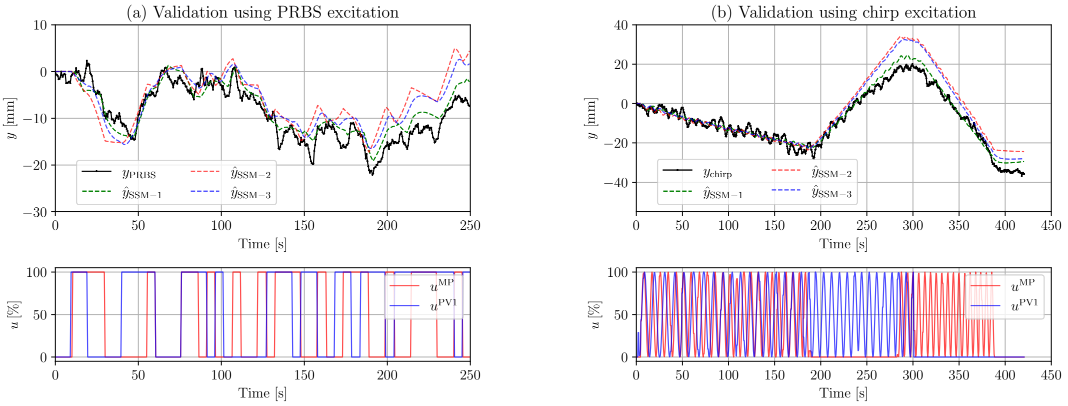

Figure 7: Validation of identified models. The measured outputs y_{PRBS} and y_{chirp} are compared with the predicted outputs $\hat{y}_{\text{SSM-1}}$, $\hat{y}_{\text{SSM-2}}$, and $\hat{y}_{\text{SSM-3}}$.

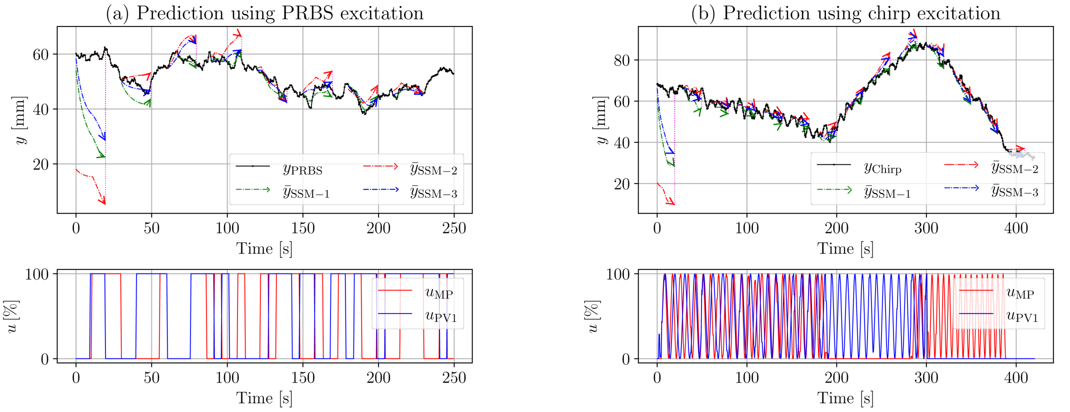

Figure 8: Model prediction for a given prediction horizon of 20 seconds. Every line segment with an arrowhead represents the prediction in the corresponding horizon. Note that a step-by-step prediction was implemented but a few horizons are shown for clear visual effect.

Table 3: Comparing the fitness of identified models

Model	PRBS excitation				Chirp excitation			
	MSE	RMSE	NRMSE	FIT [%]	MSE	RMSE	NRMSE	FIT [%]
SSM-1	8.16	2.57	0.204	79.6	6.87	2.62	0.187	81.3
SSM-2	52.85	7.27	0.518	48.2	38.59	6.21	0.44	55.7
SSM-3	76.40	6.81	0.486	51.4	44.20	6.65	0.47	52.6

3.4.1 Trends across the prediction

First of all, based on the results, the predictions under both types of excitation signals generally followed the trends of the actual measurements across most prediction horizons, particularly for the chirp excitation cases. However, wrong predictions exist. For example, in Figure 8 (a), SSM-2 and SSM-3 calculated increasing trends of the prediction from $t = 60$ s to 80 s, while the real measurement decreased in the time interval.

Furthermore, comparing results between PRBS excitation and chirp excitation, the PRBS cases were more challenging for the models in terms of predicting the system outputs. This is acceptable because the model parameter estimation relying on only chirp-type excitation can not fully capture the characteristics of the PRBS excitation case. Using more complex excitation signals, for example, combined chirp and PRBS inputs can help to improve the model accuracy. In addition, improved performance can be expected using longer data sets for identification.

Lastly, it is worth noting that a longer prediction horizon is more challenging, particularly in the PRBS excited cases. This tells that the length of the prediction horizon shall be carefully selected for model-based control algorithms such as MPC as it is one of the key factors in determining the overall performance.

3.4.2 Initial condition and state estimation

In the discussion above, the model prediction can follow the trends in the measurement for most prediction horizons. However, the predicted outputs $\bar{y}$ do not match the measured values precisely. In fact, large prediction errors can be observed at the beginning stages for all models. This is partially due to the inaccurate state estimates at the beginning.

As given in the matrix form of the M-step prediction in Eq. (28), the predicted values $\bar{y}_{k+1}, \bar{y}_{k+2}, \ldots, \bar{y}_{k+M}$ can be derived from $\hat{x}_k, y_k, u_k$, and future inputs u_{k+j}. Inaccurate state estimate $\hat{x}_k$ will result in inaccurate $\hat{x}_{k+1} (= \bar{x}_{k+1})$ therefore errors in the prediction.

Using $\hat{x}_0 = [0, 0]^T$ as the initial condition, the results of state estimation are presented in Figure 9.

As shown in Figure 9, all the estimated process output converged to the real measurements in a few time steps. However, compared with SSM-1 and SSM-3 which converged within only approximately 1 second (2 steps) under both types of excitation signals, SSM-2 took more than 5 seconds (10 steps). The results imply that SSM-1 and SSM-3 are more appropriate choices than SSM-2 for faster convergence and output estimation. For implementation of the identified models in MPC, it is recommended to wait for a few seconds before switching on the controller, so that the beginning phase of the estimation is skipped, if $\hat{x}_0$ is assumed zeros rather than estimated. For example, if SSM-2 is to be used, a waiting time of 5 to 10 seconds is necessary, so that the estimated output is close to the measurement. It is worth noting that changing the signs of both K and C will result in inverted $\hat{x}_1$ and $\hat{x}_2$ with respect to the Time axis with the same estimated output.

4 Conclusions

This work introduces a pilot tunnel system that was developed to explore real-time level control in urban drainage systems. The process dynamics are nonlinear, as described in the SVE model. In this work, three LTI SSMs identified using two subspace algorithms, N4SID and PARSIM-K have been carried out and compared for prediction performances for implementation in model-based control algorithm. The SSM-1 model estimated using the N4SID algorithm presented the best overall results. In RTC, the initial condition is mostly unknown and assumed to be zeros. To avoid inaccurate state estimation, it is recommended to wait a few seconds before switching on the real-time controller when implementing the identified models.

Additional insights into the dynamics of the system under study could be gained through further research into nonlinear model structures and online identification. A linear model structure such as SSM used in this work is incapable of capturing the nonlinear dynamics far from the operating point. Identification based on offline data sets cannot compensate for altered dynamics due to changes in system parameters, such as hardware wear out over time (García et al., 1989). Improving accuracy and precision of model predictions can be expected using longer data sets for identification.

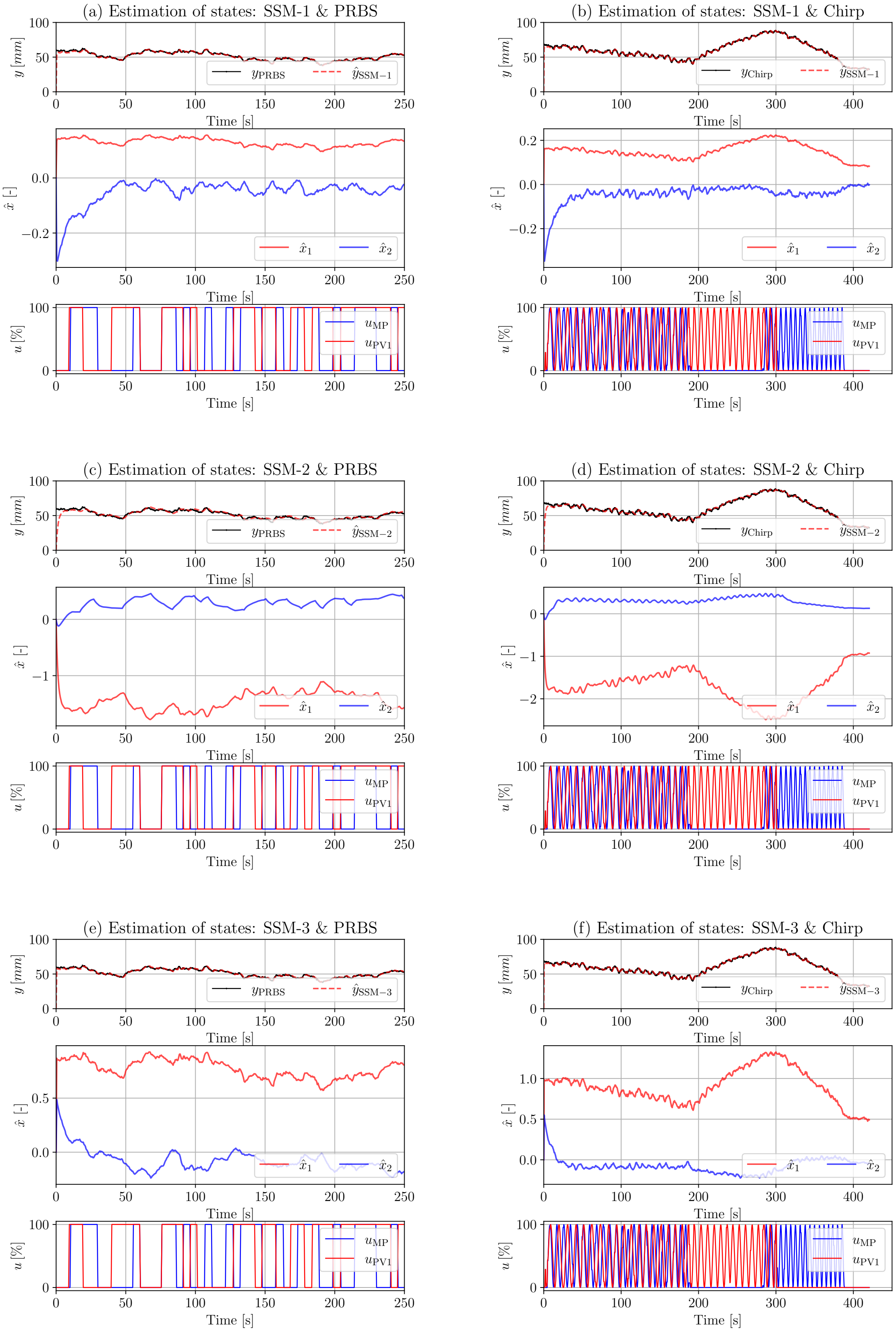

Figure 9: Estimation of states $\hat{x} = [\hat{x}_1, \hat{x}_2]^T$ using three identified models.

References

Aakre Haugen, F. Simulations and real applications of PI and MPC averaging level control in a water resource recovery facility. In *Proceedings of The 59th Conference on Simulation and Modelling (SIMS 59), 26-28 September 2018, Oslo Metropolitan University, Norway*. pages 297–302, 2018. doi:10.3384/ecp18153297.

Armenise, G., Vaccari, M., Di Capaci, R. B., and Pannocchia, G. An Open-Source System Identification Package for Multivariable Processes. In *2018 UKACC 12th International Conference on Control (CONTROL)*. IEEE, Sheffield, pages 152–157, 2018. doi:10.1109/CONTROL.2018.8516791.

Breckpot, M., Agudelo, O. M., and De Moor, B. Model Predictive Control applied to a river system with two reaches. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*. IEEE, Maui, HI, USA, pages 4549–4554, 2012. doi:10.1109/CDC.2012.6426501.

Cen, L., Xi, Y., and Li, D. Aggregation-based model predictive control of open channel networks. In *Proceedings of the 29th Chinese Control Conference*. IEEE, pages 3170–3175, 2010. URL `https://ieeexplore.ieee.org/abstract/document/5572239`.

Chanson, H. *Hydraulics of open channel flow*. Elsevier, 2004. doi:10.1016/B978-0-7506-5978-9.X5000-4.

Di Ruscio, D. Subspace system identification: Theory and applications. *Lecture notes, Telemark University College, Porsgrunn, Norway*, 1997. URL `http://davidr.no/iia2217/pensum/main_00.pdf`.

Di Ruscio, D. Model predictive control and optimization. *Lecture notes, System and Control Engineering Department of Technology Telemark University College*, 2001. URL `https://web01.usn.no/~davidr/sce4106/syllabus/main_mpc.pdf`.

Di Ruscio, D. and Foss, B. On state space model based predictive control. *IFAC Proceedings Volumes*, 1998. 31(11):301–306. doi:10.1016/S1474-6670(17)44945-7. Publisher: Elsevier.

García, C. E., Prett, D. M., and Morari, M. Model predictive control: Theory and practice—A survey. *Automatica*, 1989. 25(3):335–348. doi:10/br9kgh.

Igreja, J. M., Cadete, F. M., and Lemos, J. M. Application of distributed model predictive control to a water delivery canal. In *2011 19th Mediterranean Conference on Control & Automation (MED)*. IEEE, Corfu, Greece, pages 682–687, 2011. doi:10/fgwvjh.

Kamboh, S. A., Sarbini, I. N., Labadin, J., and Eze, M. O. Simulation of 2D Saint-Venant equations in open channel by using MATLAB. *Journal of IT in Asia*, 2016. 5(1):15–22. doi:10/ghcsm9.

Kurganov, A. and Levy, D. Central-Upwind Schemes for the Saint-Venant System. *ESAIM: Mathematical Modelling and Numerical Analysis*, 2002. 36(3):397–425. doi:10.1051/m2an:2002019.

Kurganov, A. and Petrova, G. A Second-Order Well-Balanced Positivity Preserving Central-Upwind Scheme for the Saint-Venant System. *Communications in Mathematical Sciences*, 2007. 5(1):133–160. doi:10.4310/CMS.2007.v5.n1.a6.

Litrico, X. and Fromion, V. *Modeling and control of hydrosystems*. Springer, London, 2009. doi:10.1007/978-1-84882-624-3.

Ljung, L. System identification. In A. Procházka, J. Uhlíř, P. W. J. Rayner, and N. G. Kingsbury, editors, *Signal analysis and prediction*, pages 163–173. Birkhäuser Boston, Boston, MA, 1998. doi:10.1007/978-1-4612-1768-8_11.

Ljung, L. *System identification: theory for the user*. Prentice Hall information and system sciences series. Prentice Hall PTR, Upper Saddle River, NJ, 2nd ed edition, 1999. doi:10.1016/S0005-1098(01)00214-X.

Lund, N. S. V., Falk, A. K. V., Borup, M., Madsen, H., and Steen Mikkelsen, P. Model predictive control of urban drainage systems: A review and perspective towards smart real-time water management. *Critical Reviews in Environmental Science and Technology*, 2018. 48(3):279–339. doi:10.1080/10643389.2018.1455484.

MathWorks. Loss Function and Model Quality Metrics. 2023. URL `https://se.mathworks.com/help/ident/ug/model-quality-metrics.html`.

MATLAB. System Identification Toolbox version: 10.0 (r2022b). 2023. URL `https://www.mathworks.com`.

Muroi, H. and Adachi, S. Model Validation Criteria for System Identification in Time Domain. *IFAC-PapersOnLine*, 2015. 48(28):86–91. doi:10.1016/j.ifacol.2015.12.105.

Ocampo-Martinez, C., Puig, V., Cembrano, G., and Quevedo, J. Application of Predictive Control Strategies to the Management of Complex Networks in the Urban Water Cycle. *IEEE Control Systems Magazine*, 2013. 33(1):15–41. doi:10.1109/MCS.2012.2225919.

Pannocchia, G. and Calosi, M. A predictor form PARSIMonious algorithm for closed-loop subspace identification. *Journal of Process Control*, 2010. 20(4):517–524. doi:10.1016/j.jprocont.2010.01.004.

Qin, S. J. An overview of subspace identification. *Computers & Chemical Engineering*, 2006. 30(10-12):1502–1513. doi:10.1016/j.compchemeng.2006.05.045.

Qin, S. J., Lin, W., and Ljung, L. A novel subspace identification approach with enforced causal models. *Automatica*, 2005. 41(12):2043–2053. doi:10.1016/j.automatica.2005.06.010.

Qin, S. J. and Ljung, L. Closed-loop subspace identification with innovation estimation. *IFAC Proceedings Volumes*, 2003. 36(16):861–866. doi:10.1016/S1474-6670(17)34868-1.

SciPy.org. scipy.signal.dlsim. 2023. URL `https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.dlsim.html`.

Van Overschee, P. and De Moor, B. *Subspace Identification for Linear Systems*. Springer US, Boston, MA, 1996. doi:10.1007/978-1-4613-0465-4.

Xu, M., Negenborn, R., van Overloop, P., and van de Giesen, N. De Saint-Venant equations-based model assessment in model predictive control of open channel flow. *Advances in Water Resources*, 2012. 49:37–45. doi:10/f4d992. ZSCC: 0000029.

Xu, M., van Overloop, P., and van de Giesen, N. On the study of control effectiveness and computational efficiency of reduced Saint-Venant model in model predictive control of open channel flow. *Advances in Water Resources*, 2011. 34(2):282–290. doi:10.1016/j.advwatres.2010.11.009.

Yang, H.-C. and Chang, F.-J. Modelling combined open channel flow by artificial neural networks. *Hydrological Processes*, 2005. 19(18):3747–3762. doi:10.1002/hyp.5858.

ABBREVIATIONS

Abbr.	**Full name**
ARX	Auto Regressive Exogenous
CSO	Combined Sewage Overflow
CVA	Canonical Variate Algorithm
LTI	Linear time-invariant
FIT	Fit ratio
MISO	Multi-Input-Single-Output
MSE	Mean squared error
MOESP	Multivariable Output Error State Space algorithm
MPC	Model-based predictive control
NRMSE	Normalized root mean squared error
NMSE	Normalized mean squared error
PDE	Partial differential equation
PWM	Pulse Width Modulation
PRBS	Pseudo-random binary sequence
RTC	Real-time control
SIMs	Subspace Identification Methods
SSM	State-space model
SVE	Saint-Venant equations
UDS	Urban drainage system
PARSIM	Parsimonious algorithm for system identification, such as PARSIM-K, PARSIM-S, PARSIM-P, PARSIM-E, etc.
DSR	Deterministic and Stochastic Realization algorithm for system identification

NOMENCLATURE

Symbol	[Unit]	Default value	Description
a	[mm^2]		The wetted area of open channel flow.
A, B, C, D			The system, input, output, and feedthrough matrix of SSMs.
d	[mm]	94	The inner diameter of the circular tunnel.
e			The prediction error.
g	[mm/s^2]		The gravitational constant.
h	[mm]		The water depth of open channel flow.
h^*	[mm]	60	The operation equilibrium in the open-loop experiments.
K			The steady-state Kalman filter gain matrix.
l	[mm]	4000	The length of the tunnel (V1 section).
M			The length of the prediction horizon.
n, n_y, n_u			The number of system states, outputs, and inputs.
q	[L/s]		The flow rate of the flow inside the tunnel.
Q^{PV1}	[L/min]		The inlet pump flow rate.
Q, R, S			The covariance matrices of the noise sequences.
r	[mm]		The reference signal.
S_0		1.2°	The bed slop of the pilot tunnel system.
S_f			The friction slope of open channel flow.
t, k	[s]		The continuous and discrete time index.
T_s	[s]	0.5	The sampling interval.
$u^{\mathrm{MP}}, u^{\mathrm{PV1}}$	[%]		The control signal to the main pump and inlet pump.
w, v			The process noise and measurement noise vectors.
x, y, u			The system state, output, and input vector.
$\bar{x}, \bar{y}$			The predicted process states and output.
$\hat{x}, \hat{y}$			The estimated process states and output. $\hat{x} = [\hat{x}_1, \hat{x}_2]^T$.
x_0			The initial state of the identified process model.
$y_{\mathrm{raw}}, y_{\mathrm{LP}}$	[mm]		The raw and low-pass filtered measurements.
$\tau^{\mathrm{MP}}, \tau^{\mathrm{PV1}}$	[s]		Time delay of main pump and inlet pump.
$\bar{\boldsymbol{Y}}, \bar{\boldsymbol{U}}$			Output and input vector in the prediction.
$\mathcal{H}^x, \mathcal{H}^y, \mathcal{H}^u$			M-step-ahead prediction matrix form coefficients.