Recovering nonlinear dynamics from non-uniform observations: A physics-based identification approach with practical case studies

Control Engineering Practice 164 (2025) 106411

Contents lists available at ScienceDirect

Control Engineering Practice

journal homepage: www.elsevier.com/locate/conengprac

# Recovering nonlinear dynamics from non-uniform observations: A physics-based identification approach with practical case studies

Cesare Donati [a,b,*], Martina Mammarella [b], Fabrizio Dabbene [b], Carlo Novara [a], Constantino Lagoa [c]

[a] Department of Electronics and Telecommunications, Politecnico di Torino, Corso Duca degli Abruzzi, 24, Torino, 10129, Italy
[b] Cnr-Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni, c/o Politecnico di Torino, Corso Duca degli Abruzzi, 24, Torino, 10129, Italy
[c] School of Electrical Engineering and Computer Science, The Pennsylvania State University, Electrical Engineering West, University Park, 16802, PA, USA

ARTICLE INFO

*Keywords:*
Nonlinear system identification
Non-uniform observations
Multiple runs
Missing measurements
Aggregated observations

ABSTRACT

Uniform and smooth data collection is often infeasible in real-world scenarios. In this paper, we propose an identification framework to effectively handle the so-called non-uniform observations, i.e., data scenarios that include missing measurements, multiple runs, or aggregated observations. The goal is to provide a general method for recovering the dynamics of nonlinear systems from non-uniform data, enabling accurate tracking of system behavior over time. The approach integrates domain-specific physical principles with black-box models, overcoming the limits of traditional linear or purely black-box methods. The description of this novel framework is supported by a theoretical study on the effect of non-uniform observations on the accuracy of parameter estimation. Specifically, we demonstrate the existence of upper bounds on the parametric error resulting from missing measurements and aggregated observations. Then, the effectiveness of the approach is demonstrated through two case studies. These include a practical application with missing samples, i.e., the identification of a continuous stirred-tank reactor using real data, and a simulated Lotka–Volterra system under aggregated observations. The results highlight the ability of the framework to robustly estimate the system parameters and to accurately reconstruct the model dynamics despite the availability of non-uniform measurements.

## 1. Introduction

Dynamic system identification is a well-established problem in control theory, with a wide range of applications, spanning from engineering to biological systems. When output observations are non-uniform (Aguero et al., 2007; Sleem & Lagoa, 2024), conventional identification methods – which assume regularly sampled data – become ineffective. In this situation, it becomes crucial to adopt identification techniques capable of integrating non-uniform observations and ensuring consistency despite experimental variability.

### 1.1. Case studies and motivations

The term *non-uniform observations* refers to data scenarios that may include: (i) missing measurements, (ii) multiple runs, i.e., repeated simulations of the system with different initial conditions, or (iii) aggregated outputs, where only averaged or accumulated system outputs over a time window are available, instead of individual readings. For example, sensor failures or varying sampling rates can create data gaps, while multiple runs may result from different experimental setups or varying external conditions (Yılmaz et al., 2018). On the other hand, aggregated outputs commonly arise in fields where continuous sampling is impractical. In such cases, monitoring the evolution of certain quantities can involve sampling measurements at extended intervals, providing only average values or accumulated information over these periods. In atmospheric or meteorological modeling, for example, weather stations record average temperature, humidity, or precipitation levels over several hours or days rather than continuously (Amin & Mourshed, 2024; Huntley & Hakim, 2010). In ecological and biological studies, averaged samples are used to study long-term ecological changes (Kidwell & Tomasovych, 2013) or to improve the robustness of statistical analyses and provide a better understanding of population variability (Nakagawa & Freckleton, 2011; Wangersky,

* Corresponding author at: Department of Electronics and Telecommunications, Politecnico di Torino, Corso Duca degli Abruzzi, 24, Torino, 10129, Italy.
*E-mail addresses:* cesare.donati@polito.it (C. Donati), martina.mammarella@cnr.it (M. Mammarella), fabrizio.dabbene@cnr.it (F. Dabbene), carlo.novara@polito.it (C. Novara), cml18@psu.edu (C. Lagoa).

https://doi.org/10.1016/j.conengprac.2025.106411
Received 20 December 2024; Received in revised form 29 April 2025; Accepted 15 May 2025
Available online 5 June 2025

1978). In economics and finance, data such as gross domestic product, growth rates, or quarterly earnings are typically collected at extended intervals to provide a broader view of the economic trends and financial health over time (see, e.g., Givoly and Palmon (1982)). Similarly, in chemical industries, processes like the simulated moving bed involve the collection of samples at extended intervals due to a time-consuming and costly analysis (Grossmann et al., 2009).

### 1.2. State-of-the-art

Within this context, the majority of the literature focuses on system identification techniques under the assumption of missing data and multiple runs, i.e., measurements are either sporadically absent or completely missing for certain time steps. For example, an expectation maximization-based strategy is presented in Raghavan et al. (2006) for data-driven identification with missing output observations. In this work, model parameters and missing observations are simultaneously and iteratively estimated using linear state-space models. In Isaksson (1993), several reconstruction methods for ARX models with missing measurements were compared, including Kalman filtering, maximum likelihood estimation, and iterative reconstruction. All these methods often rely on the exploitation of specific model structures, such as ARX or linear state-space models, which may not generalize well to a large variety of real systems. Furthermore, with this class of approaches, the computational cost for reconstructing missing data becomes increasingly expensive as the amount of missing data grows.

Another class of approaches dealing with missing measurements leverages nuclear norm subspace identification methods, as in Grossmann et al. (2009), Liu et al. (2013), Varanasi and Jampana (2020), where a convex optimization problem is formulated to estimate, in one step, both missing data and model parameters. Despite this class of techniques demonstrates robustness in handling incomplete dataset, their reliance on linear state-space models may jeopardize the model identification if the systems governing equations are highly nonlinear. Alternatively, an expectation–maximization algorithm that employs a particle filter and a particle smoother is employed in Gopaluni et al. (2009) and Gopaluni (2010) for the identification of a nonlinear black box model under missing observations. Analogously, solutions based on black-box neural networks are proposed in Demeester (2020) and Yuan et al. (2023) to effectively handle missing observations while identifying a system. Nonetheless, the effectiveness of black-box methods in modeling complex systems is limited by the lack of interpretability (Ljung, 2010; Pillonetto et al., 2025). This drawback compromises the possibility to rely on, e.g., prior knowledge of the model structure and physical constraints to compensate for the information lost due to missing measurements. Other recent approaches to missing data recovery include statistical and graph-based methods, such as kernel-based fault detection (Fan et al., 2021) and spatio-temporal graph convolutional networks (Yu et al., 2023), which have shown good performance in their respective domains but are not directly tailored to dynamic system identification tasks involving physical priors.

On the other hand, to the best of our knowledge, the specific challenge of aggregated outputs has not been investigated in the existing literature, leaving a notable gap in addressing the challenges posed by this type of observation data. This is a critical issue, as aggregation smooths short-term fluctuations, masks underlying dynamics, and distorts high-frequency components, leading to information loss between observation points.

### 1.3. Main contribution

Motivated by the discussion so far, in this paper, we present a novel approach for the identification of nonlinear systems able to handle non-uniform observation conditions, including missing data, multiple runs, and aggregated measurements. Specifically, we enhance the identification performance beyond the available sequence of input–output observations by integrating off-white models — domain-specific physical principles where some parameters have unknown or uncertain numerical values (Ljung, 2010) — with black-box approximators. By leveraging the known physical structure of the system, we ensure that the model remains interpretable, while the black-box term compensates not only the unknown or unmodeled dynamics but also the discrepancies that cannot be resolved due non-uniform observations. Building on the framework introduced in Donati et al. (2025a), our approach formulates the identification task as a multi-step optimization using first-order methods. Then, unlike most existing techniques (see, e.g., Liu et al. (2013), Raghavan et al. (2006)), a modification on either the cost function (for missing measurements and multiple runs) or the estimation model (for aggregated data) allows the method to handle non-uniform observations. On the other hand, it is important to highlight that the considered framework relies on first-order methods and the ensuing multi-step problem is in general nonconvex. This clearly leads to the possibility of converging to a sub-optimal solution, and implies an evident sensitivity to initial conditions and black-box approximator choice, necessitating careful tuning of the algorithm's hyper-parameters and appropriate functions selection to ensure robustness. Despite these limitations, the efficacy of this approach in the case on uniform observations has been already proven in Donati et al. (2025a), where the results showed how significantly this novel framework outperformed state-of-the-art methods. Here, we focus on demonstrating how such an informed framework proves particularly advantageous also in scenarios with non-uniform measurements, addressing limitations of traditional linear and black-box methods, and providing a more accurate and interpretable representation of the system dynamics. Moreover, we show how this framework can also be employed for case studies involving temporally averaged or accumulated data over fixed time windows, filling the gap in the literature.

The investigation of the flexibility of the framework is integrated with a rigorous analysis of the impact of data loss on parameter estimation through the definition of specific theoretical bounds on the estimation error. In particular, we show the existence of an upper bound on the parametric error in the case of missing measurements. This bound depends on the percentage of missing data and the total number of observations. Analogously, we demonstrate that a similar bound also exists for aggregated observations. In this case, the accuracy of the identified parameters is influenced by the length of the aggregation window. Then, we showcase the effectiveness and robustness of the proposed approach when applied to real-world case studies characterized by challenging dynamics and different types of non-uniform observations. Specifically, we select two different application fields. The first case study involves the identification with a dataset affected by missing measurements. In this case, real data of a continuous stirred-tank reactor provided by the DaISy database (De Moor et al., 1997) – a collection of real-world datasets frequently used to validate system identification techniques, also including scenarios with missing data (see e.g., Markovsky (2013), Varanasi and Jampana (2020)) – are utilized. The second case study targets the challenges of working with aggregated data, focusing on the identification of nonlinear ecologic systems. More specifically, we select a Lotka–Volterra model, which has been widely used to represent not only the standard predator–prey interactions (Wangersky, 1978) but also the nonlinear dynamics of economic agents or market behaviors (Malcai et al., 2002). In both case studies, the results highlight the efficacy of the proposed framework in accurately reconstructing the system dynamics despite the presence of non-uniform data.

### 1.4. Paper outline

The remainder of this paper is structured as follows. In Section 2 the proposed identification framework is briefly introduced. The extension

to multiple runs and missing observations is discussed in Section 3. Section 4 focuses on the case of aggregated observations, which are treated as a combination of the previous scenarios by the introduction of an extended system model. Results on the two practical case studies affected by non-uniform observations are discussed in Section 5 and main conclusions are drawn in Section 6.

**Notation.** Given integers $a, b \in \mathbb{N}$, $a \leq b$, we denote by $[a, b]$ the sequence of integers $\{a, \dots, b\}$. Given an integer $N > 0$ and $N$ vectors $v_i \in \mathbb{R}^n$, $\forall i$, $\forall n$, we denote as $\mathbf{v}_N$ the set of $N$ vectors $\{v_1, \dots, v_N\}$. Given a sequence $\mathbf{n}_N$ of $N$ integers and a sequence of vectors in time $v_k \in \mathbb{R}^n$, $\forall k$, we denote as $\mathbf{v}_{[\mathbf{n}_N]}$ the sequence of vectors at the times contained in $\mathbf{n}_N$, i.e., $\{v_{n_1}, \dots, v_{n_N}\}$. Moreover, given integers $a, b \in \mathbb{N}$, $a < b$, we denote as $\mathbf{v}_{a:b}$ the sequence of vectors from time $a$ to time $b$, i.e., $\{v_a, v_{a+1}, \dots, v_{b-1}, v_b\}$. The $n \times n$ identity matrix is denoted with $\mathbb{I}_n \in \mathbb{R}^{n,n}$, $1_n \in \mathbb{R}^n$ represents a vector of $n$ ones, and $0_n \in \mathbb{R}^n$ represents a vector of $n$ zeros. Similarly, $\mathbb{1}_{n,m} \in \mathbb{R}^{n,m}$ and $\mathbb{0}_{n,m} \in \mathbb{R}^{n,m}$ represent an $n \times m$ matrix of all ones and all zeros, respectively. Given a matrix $A \in \mathbb{R}^{m,n}$, we define its $\ell_p$ norm as $\|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}$.

## 2. Proposed identification approach

In this section, we briefly outline the physics-based system identification approach proposed in Donati et al. (2025a), which combines an off-white model with black-box approximators.

We consider a nonlinear, time-invariant system described by the following mathematical model

$$\begin{aligned} \mathcal{S}: \quad & x_{k+1} = f\left(x_k, u_k; \theta\right) + \Delta(x_k, u_k), \\ & z_k = h\left(x_k; \theta\right), \end{aligned} \tag{1}$$

where $x \in \mathbb{R}^{n_x}$ is the state vector, $\theta \in \mathbb{R}^{n_\theta}$ is the parameter vector, $u \in \mathbb{R}^{n_u}$ is the input, and $z \in \mathbb{R}^{n_z}$ is the output. Functions $f$ and $h$ are known and assumed to be nonlinear, time-invariant, and continuously differentiable. The term $\Delta$, representing unmodeled dynamics, is unknown.

Given a multi-step input/output sequence of length $T$,

$$\begin{aligned} \widetilde{\mathbf{u}}_{0:T-1} &= \{\widetilde{u}_0, \dots, \widetilde{u}_{T-1}\}, \quad \widetilde{u}_k = u_k + \eta_k^u, \\ \widetilde{\mathbf{z}}_{0:T-1} &= \{\widetilde{z}_0, \dots, \widetilde{z}_{T-1}\}, \quad \widetilde{z}_k = z_k + \eta_k^z, \end{aligned} \tag{2}$$

with $\eta_k^u$ and $\eta_k^z$ being the input and output measurement noise, respectively, we seek to estimate the unknown system parameters $\theta$ and initial condition $x_0$ while compensating for the unknown term $\Delta$. To this aim, we define the following estimation model $\mathcal{M}$ to approximate the system $\mathcal{S}$

$$\begin{aligned} \mathcal{M}: \quad & \hat{x}_{k+1} = f(\hat{x}_k, u_k; \hat{\theta}) + \delta(\hat{x}_k, u_k; \omega), \\ & \hat{z}_k = h(\hat{x}_k; \hat{\theta}), \end{aligned} \tag{3}$$

where $\delta$ is a generic approximator of the unknown term $\Delta$, e.g., a linear combination of basis functions from a given dictionary, with parameters $\omega \in \mathbb{R}^{n_\omega}$ to be learned.

For uniform observations, the framework uses the prediction error $e_k \doteq \widetilde{z}_k - \hat{z}_k$ and aims to estimate the optimal values of $\theta$, $x_0$, and $\omega$ by solving the optimization problem

$$\left(\theta^\star, x_0^\star, \omega^\star\right) \doteq \arg\min_{\theta, x_0, \omega} \; C_T(\theta, x_0, \omega; \mathbf{e}_{0:T}, \hat{\mathbf{x}}_{1:T}). \tag{4}$$

Here, $C_T$ is a twice continuously differentiable, multi-step cost function, generally defined as

$$C_T \doteq \frac{1}{T} \sum_{k=0}^{T-1} \mathcal{L}_k\left(\theta, x_0, \omega; e_k, \hat{x}_k\right), \quad \mathcal{L}_k = \|\widetilde{z}_k - \hat{z}_k\|_2^2. \tag{5}$$

Specifically, the optimal estimation of the unknown system parameters $\theta$, initial condition $x_0$, and black-box parameters $\omega$ is achieved by minimizing $C_T$ in (5) over a given horizon[1] $T$. Here, $C_T$ incorporates terms that account for the squared prediction errors and may also include additional regularization or physics-based penalty terms to enforce constraints derived from the system knowledge (see Donati et al. (2025a) for further details). As discussed in Donati et al. (2024), the optimization problem is addressed by using first-order methods. In particular, the gradient of the cost function $C_T$ is computed at each iteration, and the parameters are updated accordingly, converging to a (potentially sub-optimal) solution. Specifically, we adopt the gradient computation framework from (Donati et al., 2025b), which leverages automatic differentiation in the context of system identification to find a solution to (4) (Donati et al., 2025b, Algorithm 1). Within this context, we note that, in multi-step nonlinear identification, recursive prediction can lead to unstable gradients as the horizon or parameter dimension increases. In Donati et al. (2025b), we formally analyze the gradient evolution and provide guarantees for gradient stability using a linear parameter-varying formulation. These results support the use of forward-mode automatic differentiation and justify the robustness of the proposed first-order optimization scheme in practice. In particular, it is shown that the gradient computation scales linearly with both the number of parameters and the prediction horizon, making the proposed first-order approach also suitable for nonlinear systems with higher dimensions. On the other hand, as is common in nonlinear system identification, the resulting optimization problem is generally non-convex, and convergence may depend on initialization and step size (Ghadimi & Lan, 2016).

## 3. Missing observations and multiple runs

In numerous practical scenarios, it is necessary to identify a system from *irregular sampling*, like in the case of missing data, or from multiple runs with unknown initial conditions. In this section, we show how the physics-based approach can be seamlessly extended to include these scenarios.

### 3.1. Missing observations

First, we present the scenario of *missing observations*, which arises when only a limited set of output measurements are collected at non-uniform time steps in a given multi-step setting. Practically, this applies to, e.g., sensor failure, data loss, data cleaning, or limited sampling capabilities.

To formally define this setup, we consider the uniform sequence of observations in (2) and the set of *available time steps* $\boldsymbol{\kappa}_N$, i.e., a set of $N < T$ ordered integers defined as

$$\boldsymbol{\kappa}_N = \{k_1, \dots, k_N\}, \; k_j \in [0, T-1], \; \forall j \in [1, N], \tag{6}$$

where a time index $k_j \in \boldsymbol{\kappa}_N$ is included if at least one output component is available at time $k_j$. Accordingly, the sequence of available measurements is defined as follows

$$\widetilde{\mathbf{z}}_{[\boldsymbol{\kappa}_N]} = \{\widetilde{z}_{k_1}, \dots, \widetilde{z}_{k_N}\}. \tag{7}$$

In this case, the cost function $C_T$ is defined by selecting only the time steps at which measurements are known to be available, i.e.,

$$C_T = \frac{1}{T} \sum_{j=1}^{N} \mathcal{L}_{k_j}, \quad \mathcal{L}_{k_j} = \|\widetilde{z}_{k_j} - \hat{z}_{k_j}\|_2^2, \tag{8}$$

with $k_j$ the $j$th element of $\boldsymbol{\kappa}_N$. Moreover, we note that the special case of partial measurements, i.e., when only a subset of the output components is available at a given time, can be seamlessly incorporated

[1] In our framework, the horizon $T$ corresponds to the available data length; while larger $T$ values generally improve estimation accuracy, they may also increase computational cost (Donati et al., 2025b), requiring a trade-off between information richness and numerical complexity.

into the cost function (8). Specifically, we redefine the loss as $\mathcal{L}_{k_j} = (\widetilde{z}_{k_j} - \hat{z}_{k_j})^\top P_{k_j} (\widetilde{z}_{k_j} - \hat{z}_{k_j})$, where $P_{k_j} = \text{diag}(p_{11}, \dots, p_{n_z n_z})$ is a diagonal matrix such that, for $i \in [1, n_z]$, $p_{ii} = 1$ if the $i$th output component is available at time $k_j$, and $p_{ii} = 0$ otherwise.

Despite the simplicity of accounting for missing measurements in the optimization problem, it is crucial to assess their effect on the estimation error in the identified parameters, relying on the assumption on the local identifiability reported in the follows.

**Assumption 1** (*Local Identifiability*)**.** The system is locally identifiable according to, e.g., Bellman and Åström (1970) and Donati et al. (2025a, Definition 1). In other words, the Hessian of the loss function evaluated in $\theta^\star$ is always positive definite, i.e.,

$$H \doteq \left.\frac{\partial^2 C_T(\theta;\cdot)}{\partial^2 \theta}\right|_{\theta=\theta^\star} \succ 0.$$

In the following theorem, which proof is reported in Appendix A, we demonstrate that there exists an upper bound on the discrepancy between the parameters identified with $N$ available measurements and those obtained with a complete set of $T$ data. This bound grows proportionally to the square root of the percentage of missing measurements and inversely proportional to $\sqrt{T}$.

**Theorem 1** (*Error Bound with Missing Measurements*)**.** *Let $\theta_T^\star$ represents the vector of identified parameters obtained by solving problem (4) using a complete set of $T$ observations. Similarly, let $\theta_N^\star$ denotes the identified parameters when only $0 < N \leq T$ observations are available, due to missing data. Define $p_{miss} = \frac{T-N}{T}$ as the percentage of missing observations and let Assumption 1 hold. Then, the error between the identified parameters under missing measurements and those from the complete dataset satisfies*

$$\|\theta_T^\star - \theta_N^\star\|_2 \leq \sigma_\xi \frac{1}{\sqrt{T}} \sqrt{p_{miss}}, \tag{9}$$

*for some constant $\sigma_\xi \in \mathbb{R}$.*

Theorem 1 establishes a link between the percentage of missing observations, $p_{\text{miss}}$, the multi-step horizon $T$, and the error in the identified parameters. This bound indicates that the worst-case parametric error increases with the square root of the missing data fraction, highlighting the sensitivity of parameter estimation to the gaps in the data. On the other hand, the factor $\frac{1}{\sqrt{T}}$ highlights that being the missing data percentage fixed, the datasets collected over shorter horizons (i.e., lower $T$) are inherently more sensitive to missing data, thus leading to a relatively larger error. At the same time, larger datasets effectively help to mitigate the negative impact of missing entries. The proportionality constant $\sigma_\xi$ depends on the maximum singular value of a matrix whose columns describe how the optimal parameters vary with respect to data weighting (see Appendix A for details). This suggests that systems with certain structural properties (i.e., small $\sigma_\xi$) are more robust to incomplete datasets.

In the following section, we present a numerical examples that showcases the adherence of the retrieved upper bound with a simulated system subject to missing measurements.

*Numerical analysis of the upper bound on missing data*

To support the result of Theorem 1, we demonstrate how missing observations quantitatively affect the parameter identification accuracy by comparing the identification of a second-order linear system with and without missing measurements. The system, presented in Yılmaz et al. (2018), is described by the following transfer function

$$g(z) = \frac{(\theta_1 z + \theta_2)}{(z^2 + \theta_3 z + \theta_4)},$$

and in the canonical companion state-space form as

$$x_{k+1} = \begin{bmatrix} -\theta_3 & -\theta_4 \\ 1 & 0 \end{bmatrix} x_k + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u,$$
$$z_k = \begin{bmatrix} \theta_1 & \theta_2 \end{bmatrix} x_k,$$

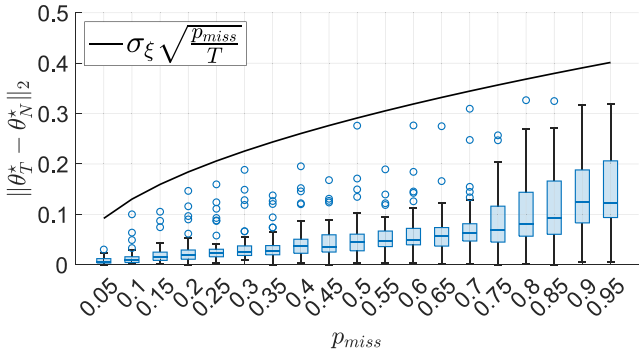


**Fig. 1.** Box-plot illustrating the distribution of parameter estimation errors for varying percentages of missing data ($p_{\text{miss}}$). Each box represents the interquartile range (IQR) of errors, with the median shown as the horizontal line, while the whiskers extend the range of errors to values within $1.5 \times$ IQR ($\approx 3\sigma$) beyond the quartiles. Dots represent errors that fall outside this range. The solid line represents the upper bound trajectory from Theorem 1.

with $\theta_1 = 0.1037$, $\theta_2 = -0.08657$, $\theta_3 = -1.78$, $\theta_4 = 0.9$. First, we excite the system's step response with a perturbation defined by $\mathcal{N}(0, 0.01)$, measured over a horizon $T = 104$. Then, we use this signal within the proposed approach to identify the system parameters, accounting for different percentages of missing measurements, i.e., from $p_{\text{miss}} = 0.05$ to $p_{\text{miss}} = 0.95$. To achieve this goal, we minimize a cost function of the form (8) using a first-order optimization method. The predicted states and outputs are propagated along the horizon $T$, while the gradient is computed via automatic differentiation, relying on the approach outlined in Donati et al. (2024, 2025b).

Fig. 1 shows the relationship between the percentage of missing observations ($p_{\text{miss}}$) and the $\ell_2$ norm of the difference between parameters identified with $p_{\text{miss}}T$ missing data and in the complete-data case (i.e., $p_{\text{miss}} = 0$), respectively. Results are collected from 50 simulations for each percentage of missing data, with each simulation featuring different noise values and initial parameter conditions. We can notice that the error bound scales with the proportion of missing data. Moreover, when overlapping the upper bound derived in Theorem 1 for $\sigma_\xi = 4.2$ to the numerical data, we can observe that this bound well retrace the data behavior, and it appears more conservative at lower values of $p_{\text{miss}}$. Furthermore, we can observe that the shorter boxes reflect estimation errors tightly clustered around the mean, with few outliers approaching the upper bound. This suggests that, for lower percentages of missing data, the identified parameters are more accurate and stable, exhibiting limited variability even in the presence of some data gaps. Conversely, as $p_{\text{miss}}$ increases, the spread of estimation errors around the mean widens, as indicated by the larger boxes. This behavior denotes greater uncertainty and variability in the parameter estimates and it also suggests that a higher amount of missing data makes the identification process less reliable, resulting in a tighter and less conservative upper bound.

To complete the analysis, Fig. 2 shows the root mean square error (RMSE) between predictions and observations for varying percentages of missing data. In this case, the plot reveals, as expected, an increasing RMSE trend with higher $p_{\text{miss}}$ values, thus underscoring the growing discrepancy between predictions and observations as the proportion of missing data rises.

### 3.2. Multiple runs

We now consider the case of *multiple runs*, where identification relies on data from different system trajectories. This is the case, for example, of repeated experiments under varying initial conditions, different inputs, environmental disturbances, or sensor placements. Additionally, multiple runs are also exploited to identify a more robust

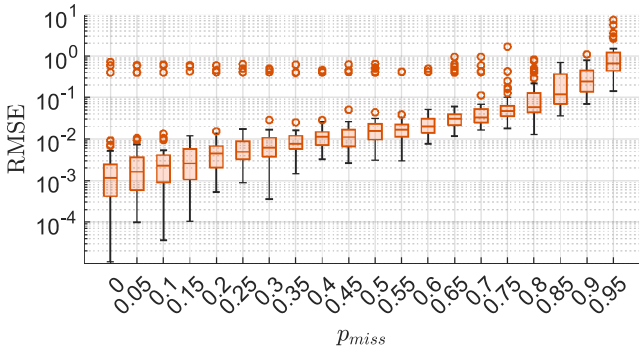


**Fig. 2.** Box-plot illustrating the RMSE for varying percentages of missing data (logarithmic scale).

model, capable of capturing a wider range of system behaviors and generalizing well to unseen scenarios. Alternatively, multiple runs can be defined from a single trajectory by dividing it into smaller segments (see, e.g., the multiple shooting method in Ribeiro et al. (2020)) to obtain specific optimization properties, such as smoothing the cost function and improving numerical stability. In this setup, we consider $M$ different runs of the system $S$ (1), starting from $M$ different initial conditions. For each $i$th experiment, with $i = [1, M]$, the following noise-corrupted sequences of input–output data, each of length[2] $T_r$, are available, i.e.,

$$\begin{aligned}\widetilde{\mathbf{u}}^{(i)}_{[0:T_r-1]} &= \{\widetilde{u}^{(i)}_0, \dots, \widetilde{u}^{(i)}_{T_r-1}\},\\ \widetilde{\mathbf{z}}^{(i)}_{[0:T_r-1]} &= \{\widetilde{z}^{(i)}_0, \dots, \widetilde{z}^{(i)}_{T_r-1}\}.\end{aligned} \tag{10}$$

Hence, the cost function $C_T$ in (4) can be redefined as

$$C_T = \frac{1}{MT_r}\sum_{i=1}^{M}\sum_{k=0}^{T_r-1}\mathcal{L}^{(i)}_k, \quad \mathcal{L}^{(i)}_k = \|\widetilde{z}^{(i)}_k - \hat{z}^{(i)}_k\|_2^2. \tag{11}$$

In the case of multiple runs, the number of decision variables in the optimization problem (4) increases. Indeed, since each run starts from a different initial condition $x^{(i)}_0$, $i \in [1, M]$, the optimization problem (4) must be minimized with respect to all initial conditions $x^{(1)}_0, \dots, x^{(M)}_0$. Moreover, it is worth noting that missing measurements and multiple runs can occur simultaneously, as remarked in the following.

**Remark 1** (*Multiple Runs with Missing Measurements*)**.** Let us consider $M$ runs, each of length $T_r$. Then, we define the set of $N < T_r$ available time steps[3] for the $i$th run as

$$\boldsymbol{\kappa}^{(i)}_N = \{k^{(i)}_1, \dots, k^{(i)}_N\},\ k_j \in [0, T_r - 1],\ \forall j \in [1, N].$$

Accordingly, the available measurements at each run are

$$\widetilde{\mathbf{z}}^{(i)}_{[\boldsymbol{k}^{(i)}_N]} = \{\widetilde{z}^{(i)}_{k^{(i)}_1}, \dots, \widetilde{z}^{(i)}_{k^{(i)}_N}\}.$$

Hence, in the case of multiple runs with missing measurements, the cost function $C_T$ can be defined by combining (8) and (11), i.e.,

$$C_T = \frac{1}{MT_r}\sum_{i=1}^{M}\sum_{j=1}^{N}\mathcal{L}^{(i)}_{k^{(i)}_j}, \quad \mathcal{L}^{(i)}_{k^{(i)}_j} = \|\widetilde{z}^{(i)}_{k^{(i)}_j} - \hat{z}^{(i)}_{k^{(i)}_j}\|_2^2. \tag{12}$$

To enhance the clarity of the proposed method, Algorithm 1 summarizes the key steps of the identification procedure in the presence of missing data and multiple runs.

[2] For simplicity, we assume that each trajectory has the same length $T_r$. However, this non-restrictive assumption can be easily relaxed to accommodate sequences of different lengths.

[3] Without loss of generality, we use the same $N$ and $T_r$ for each run.

**Algorithm 1** Identification with multiple runs and missing measurements

**Inputs:** Non-uniform dataset with missing data (7), and multiple runs (10). Time indices with available outputs $\boldsymbol{\kappa}^{(i)}_N$ (6) for each run. Initial guesses for $\theta$, $\omega$, and $x^{(i)}_0$.
**while** not converged **do**
  **for** each run $i = 1$ to $M$ **do**
    Simulate model over $T_r$ using current $(\theta, \omega, x^{(i)}_0)$.
    Compute error $e_k$ at times $k \in \boldsymbol{\kappa}^{(i)}_N$.
  **end for**
  Compute total cost function over all runs (12).
  Update $\theta$, $\omega$, $x^{(i)}_0$ using gradient descent (Donati et al., 2025b).
**end while**
**Output:** Estimated parameters $\theta^\star$, $\omega^\star$, and $x^{(i)\star}_0$.

## 4. Aggregated observations

The case of aggregated — cumulative or averaged — observations occurs when over a given time window only gathered information of multiple individual measurements is accessible. This is the case, for example, of monitoring changes of some quantities over extended periods rather than capturing short-term data for practical reasons. In this context, it is important to differentiate two concepts. On one side, we have the so-called running averages, where each measurement is computed as the average of a set of preceding measurements, including the current one. On the other hand, *periodic averaging*, i.e., averaged observations considered as the mean of a fixed set of measurements over a specified time window, without, in general, overlapping between consecutive windows. While the proposed framework can accommodate running averages as a special case, the primary focus will be on *periodic averaging*.

While missing measurements and multiple runs can be handled with minor adjustments to the cost function (5), aggregated measurements require differently refining the initial formulation. Specifically, let us consider $T$ observations and $M$ (possibly consequent) time windows of length $T_r \doteq T/M$ for system (1). Then, the sequence of available measurements can be defined as

$$\begin{aligned}\widetilde{\mathbf{Z}}_M &= \{\widetilde{Z}^{(1)}_{T_r}, \dots, \widetilde{Z}^{(M)}_{T_r}\}, \quad \widetilde{Z}^{(i)}_{T_r} = Z^{(i)}_{T_r} + \eta^Z_i,\\ Z^{(i)}_{T_r} &\doteq \alpha\sum_{k=0}^{T_r-1} z^{(i)}_k, \quad \forall i \in [1, M],\end{aligned} \tag{13}$$

with $\eta^Z_i$ the measurement noise related to the $i$th cumulative observation, and the parameter $\alpha$ defined according to the type of data, i.e., $\alpha = 1$ for cumulative measurements and $\alpha = \frac{1}{T_r}$ when considering averaged measurements.

This non-standard representation of the observations compresses multiple individual measurements into a single data point, concealing short-term dynamics and making it difficult to directly apply standard identification techniques. To address this challenge, we propose an extended system model that reinterprets the problem as one involving both missing measurements and multiple runs, as detailed next.

First, we define the following extended system,

$$\bar{S}: \quad x_{k+1} = f\left(x_k, u_k; \theta\right) + \Delta(x_k, u_k), \tag{14a}$$

$$c_{k+1} = c_k + h\left(x_k; \theta\right), \tag{14b}$$

$$\bar{z}_k = \alpha c_k, \tag{14c}$$

described by the extended state vector $\bar{x} = [x_k^\top, c_k^\top]^\top$, with $c \in \mathbb{R}^{n_z}$ the cumulative state, which aggregates the outputs $h(x_k; \theta)$ over time, and the new output $\bar{z}_k \in \mathbb{R}^{n_z}$. The extended system configuration is depicted in Fig. 3.

The following theorem, which proof is reported in Appendix B, describes how the case of aggregated measurements can be handled by

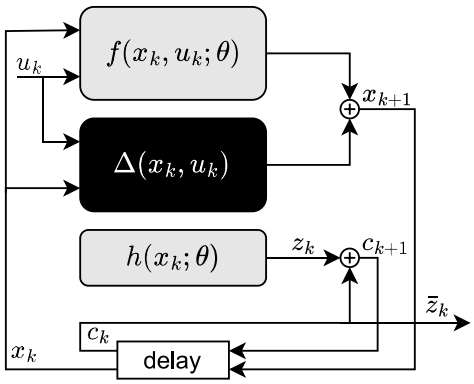


**Fig. 3.** Extended system configuration.

exploiting the tools previously defined for missing measurements and multiple runs, by properly integrating the extended system $\bar{S}$ (14).

**Theorem 2** (*Systems Equivalence*). *Let us consider $M$ aggregated (cumulative or averaged) observations defined by (13) for the system (1), collected from $M$ (possibly consecutive) time windows of length $T_r$. Let $x_0^{(i)}$ be the initial condition of the $i$th time window. Then, let us consider $M$ multiple runs of length $T_r + 1$ of the system (14), having initial conditions $[x_0^{(i)\top}, 0_{n_z}^\top]^\top$. For each run, let us consider missing measurements, as detailed in Section 3.1, with a vector of available time steps $\boldsymbol{\kappa}_1 = \{T_r\}$. The resulting sequence of available $M$ observations for the extended system in (14), i.e., $\bar{\mathbf{z}}_{[\boldsymbol{\kappa}_1]}^{(i)}$ with $i \in [1, M]$, corresponds to the aggregated observations defined by (13) for system (1). That is*

$$\bar{\mathbf{z}}_{[\boldsymbol{\kappa}_1]}^{(i)} = Z_{T_r}^{(i)}, \quad \forall i \in [1, M]. \tag{15}$$

Now, let us define the extended estimation model as

$$\begin{aligned} \bar{\mathcal{M}}: \quad & \hat{x}_{k+1} = f(\hat{x}_k, u_k; \hat{\theta}) + \delta(\hat{x}_k, u_k; \omega), \\ & \hat{c}_{k+1} = \hat{c}_k + h(\hat{x}_k; \hat{\theta}), \\ & \hat{\bar{z}}_k = \alpha c_k. \end{aligned} \tag{16}$$

From Theorem 2 and Remark 1, it follows that, considering $M$ multiple runs and $\boldsymbol{\kappa}_1 = \{T_r\}$, the cost function $C_T$ can be redefined as

$$C_T = \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{1} \mathcal{L}_{k_j}^{(i)} = \frac{1}{M} \sum_{i=1}^{M} \mathcal{L}_{T_r}^{(i)}, \tag{17}$$

with $\mathcal{L}_{T_r}^{(i)} = \| \widetilde{Z}_{T_r}^{(i)} - \hat{\bar{z}}_{T_r}^{(i)} \|_2^2$ and $\hat{c}_0^{(i)} = 0_{n_z}$ for all $i \in [1, M]$.

The proposed method streamlines both analysis and implementation by treating aggregated measurements in the same way as missing data and multiple runs, as discussed in Remark 1. Once the data are represented within this framework, previously established algorithms can be applied directly for identification, avoiding the need for non-standard formulations to handle data aggregation in the cost function. A practical implementation of the proposed framework for aggregated observations is provided in Algorithm 2.

Next, we need to assess the contribution of the aggregated observations to the estimation error of the identified parameters. Similar to the case of missing data, in the following theorem, we demonstrate the existence of an upper bound on the discrepancy between the parameters identified with aggregated data (with a window of length $T_r < T$) and those from a complete dataset. This bound increases proportionally to $\sqrt{T_r}$.

**Theorem 3** (*Error Bound with Aggregated Observations*). *Let $\theta_T^\star$ be the vector of identified parameters obtained as the solution to the optimization problem (4) when a complete set of $T$ observations is available. Similarly, let $\theta_{T_r}^\star$ represent the vector of identified parameters obtained when the observations are aggregated over a window of length $T_r$. Let Assumption 1 hold. Then, the error between the identified parameters under aggregated observations and those obtained from the complete dataset satisfies*

$$\| \theta_T^\star - \theta_{T_r}^\star \|_2 \leq L_\theta \beta_{T_r}, \tag{18}$$

*for some constant $L_\theta \in \mathbb{R}$, where $\beta_{T_r}$ is bounded and depends on $\sqrt{T_r}$ as follows*

$$\sqrt{T_r} - 1 \leq \beta_{T_r} \leq \sqrt{T_r} + 1. \tag{19}$$

**Algorithm 2** Identification with aggregated observations

**Input:** Aggregated outputs $\widetilde{Z}_{T_r}^{(i)}$ (13) for $M$ time windows of length $T_r$, input sequences for each window, scaling factor $\alpha$. Initial guesses for $\theta$, $\omega$, and $x_0^{(i)}$.
**while** not converged **do**
  **for** each window $i = 1$ to $M$ **do**
    Simulate extended model (16) over horizon $T_r$.
    Use $\hat{z}_{T_r}$ as predicted cumulative output.
    Compute error $e_{T_r}^{(i)} = \widetilde{Z}_{T_r}^{(i)} - \hat{\bar{z}}_{T_r}^{(i)}$.
  **end for**
  Compute total cost (17).
  Update $\theta$, $\omega$, and $x_0^{(i)}$ using gradient descent (Donati et al., 2025b).
**end while**
**Output:** Estimated parameters $\theta^\star$, $\omega^\star$, and $x_0^{(i)\star}$.

Theorem 3, which proof is provided in Appendix C, implies that larger aggregation windows (or fewer measurements), can lead to greater deviations in the identified parameters. Similar to the case with missing measurements, this behavior emphasizes the effect of non-uniform observation on the parameter estimation accuracy: while aggregating data may be more practical in some scenarios, the resulting effect can mask short-term dynamics, leading to less accurate identification.

## 5. Case studies

In this section, we present two case studies to demonstrate the efficacy of the proposed framework in handling missing and aggregated observations, respectively. We remark that in both examples the optimization is carried out using a first-order method, propagating the predictions while the gradient is computed through automatic differentiation (Donati et al., 2024, 2025b).

### 5.1. Identification with missing measurements

To explore the practical use of the proposed identification method for handling the case of missing data, we consider the continuous stirred-tank reactor (CSTR) described in Morningred et al. (1992). Specifically, we aim to identify the dynamical models for the CSTR relying on a real dataset with different rates of missing observation extracted from the DaISy benchmarks collection (De Moor et al., 1997).

#### 5.1.1. System description

The continuous stirred-tank reactor, depicted in Fig. 4, is governed by an exothermic process with irreversible reaction, where the product concentration is controlled by regulating the coolant flow. This system has been widely investigated and it is recognized as a highly challenging benchmark for nonlinear process modeling, optimization, and control (see, e.g., Gopaluni (2008), Lu and Huang (2014), Morningred et al. (1992) and references therein). From a system identification perspective, its inherent nonlinear dynamics and sensitivity to operating conditions make it an ideal test-bench for validating identification

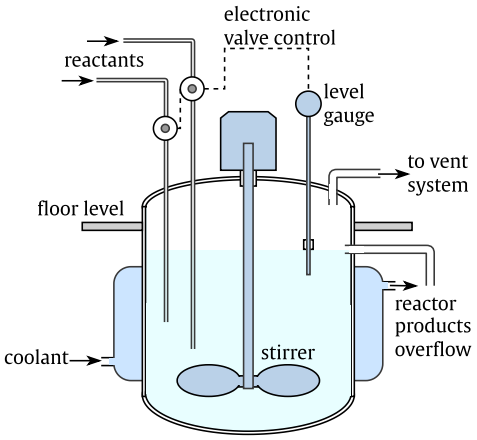


**Fig. 4.** Schematic illustration of a CSTR system.

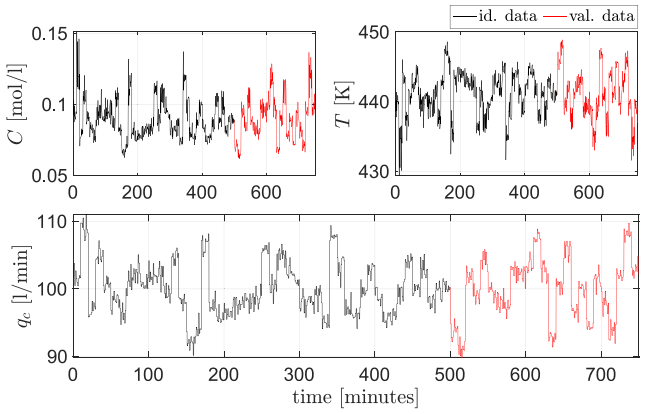


**Fig. 5.** Identification (black) and validation (red) input–output measurements related to the CSTR system. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 1**
Nominal CSTR parameter values.

| Name | Description | Value | Unit |
|---|---|---|---|
| $C_A$ | product concentration | $x_1$ | [mol/l] |
| $T$ | reactor temperature | $x_2$ | [K] |
| $q_c$ | coolant flow rate | $u$ | [l/min] |
| $q$ | process flow rate | 100 | [l/min] |
| $C_0$ | feed concentration | 1 | [mol/l] |
| $T_0$ | feed temperature | 350 | [K] |
| $T_{c0}$ | inlet coolant temp | 350 | [K] |
| $V$ | CSTR volume | 100 | [l] |
| $hA$ | heat transfer term | $7 \cdot 10^5$ | [cal/min/K] |
| $k_0$ | reaction rate constant | $7.2 \cdot 10^{10}$ | [min$^{-1}$] |
| $\frac{E}{R}$ | activation energy term | $1 \cdot 10^4$ | [K] |
| $\Delta H$ | heat of reaction | $-2 \cdot 10^5$ | [cal/mol] |
| $\rho, \rho_c$ | liquid densities | $1 \cdot 10^3$ | g/l |
| $C_p, C_{pc}$ | specific heats | 1 | [cal/g/K] |
| $\Delta t$ | sampling time | 0.1 | [min] |

strategies. This process has been studied in the literature also in the case of missing data conditions. Some works, such as (Deng & Huang, 2012; Gopaluni et al., 2009; Yang et al., 2018), focus on the same system but rely on different datasets, while others, such as (Demeester, 2020; Liu et al., 2013; Yuan et al., 2023), specifically investigate the same dataset used in this paper.

As described in Morningred et al. (1992), the CSTR system is governed by the following discretized dynamical first-principle equations

$$\begin{aligned} C_{k+1} &= C_k + \Delta t \left[ \frac{q}{V} (C_0 - C_k) - k_0 C_k e^{-\frac{E}{RT_k}} \right], \\ T_{k+1} &= T_k + \Delta t \left[ \frac{q}{V} (T_0 - T_k) - \frac{(-\Delta H)\, k_0}{\rho C_p} C_k e^{-\frac{E}{RT_k}} \right. \\ &\quad \left. + \frac{\rho_c C_{pc}}{\rho C_p V} q_{c,k} \left( 1 - e^{-\frac{hA}{q_{c,k} \rho C_p}} \right) (T_{c0} - T_k) \right], \end{aligned} \tag{20}$$

where the product concentration $C_k$ and the reactor temperature $T_k$ are the state variables, whereas the coolant flow rate $q_{c,k}$ is the input. Moreover, in this system, the outputs coincide with the states, i.e., $\tilde{z}_k = [C_k, T_k]^\top$. The goal is to identify the following vector of parameters $\theta = \left[k_0, \frac{(-\Delta H)k_0}{\rho C_p}, hA\right]^\top$, as in Gopaluni (2008), Gopaluni et al. (2009). The nominal parameter values used in the simulations and their physical description are reported in Table 1. The input–output dataset for this process is illustrated in Fig. 5. It includes 7500 samples, 5000 allocated for the identification task (black line) and the remaining 2500 (red line) reserved for validation.

In the following simulations, we numerically verify Assumption 1 by approximating the Hessian using both the Gauss–Newton method (see, e.g., Björck (1996)), i.e., $H(\theta) \approx J^\top J$ where $J$ is the Jacobian of the residual vector with respect to the parameters and finite-difference perturbations of the gradient. In both cases, the approximate Hessians are positive definite. The Hessian exhibits a condition number $\approx 10^8$, primarily due to the large variation in the scale of the estimated parameters, which ranged from $10^5$ to over $10^{13}$. While such disparities in parameter magnitudes can lead to ill-conditioned Hessians, the optimization process is stable and convergent over al the simulations, confirming the practical validity of Assumption 1.

### 5.1.2. Identification results

Identification is achieved by minimizing a cost function of the form (8) over the identification data, generating different versions of the original dataset with missing measurement rates ranging from $p_{\text{miss}} = 0\%$ to $p_{\text{miss}} = 75\%$. For each rate of missing data, the results are collected over 200 simulations, each one employing different initial values for the parameters to be identified. Specifically, the estimated parameters are initialized randomly, with each initial value $\widehat{\theta}_{i,0}$ selected within a ball around the nominal parameter value $\theta_i$ and a radius of 30% of $\theta_i$. That is, $\widehat{\theta}_{i,0}$ is chosen such that $\widehat{\theta}_{i,0} \in [\theta_i - 0.3\theta_i, \theta_i + 0.3\theta_i]$. The states initial conditions to be estimated are initialized at $\hat{x}_{0,0} = \tilde{z}_0$.

The black-box compensation term $\delta(\cdot)$ is introduced into the dynamical model to handle the process nonlinearities and to guarantee adaptation to unmodeled variations in system parameters. Indeed, the CSTR system is typically subject to changes in the environmental and operational conditions, and it may experience fluctuations that the basic physics-based model (20) alone cannot capture accurately (Deng & Huang, 2012; Morningred et al., 1992). In this example, $\delta$ is defined as a linear combination of sigmoid, softplus, hyperbolic tangent, trigonometric, and polynomial functions. Then, a regularization term is introduced in the cost function (8) to promote a sparse black-box component. This is done by minimizing an approximation of the $\ell_1$-norm of the black-box weights $\omega$ (see Donati et al. (2025a) for further details). The results presented next are computed on the validation dataset.

Fig. 6 depicts the effect of missing observations on the model fitness scores, which are computed for the $i$th output as

$$\text{fit}_{\%}^{(i)} = 100 \left( 1 - \frac{\sum_{k=0}^{T-1} (\hat{z}_k^{(i)} - \tilde{z}_k^{(i)})^2}{\sum_{k=0}^{T-1} (\hat{z}_k^{(i)} - \frac{1}{T} \sum_{k=0}^{T-1} \tilde{z}_k^{(i)})^2} \right).$$

The box plot highlights the distribution of the model fitness across 200 simulations for each output and for each level of data loss. The global fitness (yellow boxes) represents the average between the two outputs' fitness, i.e., $\text{fit}_{\%}^{(1)}$ (green boxes) for the first output and $\text{fit}_{\%}^{(2)}$ (red boxes) for the second one. Hence, we have $\text{fit}_{\%} = (\text{fit}_{\%}^{(1)} + \text{fit}_{\%}^{(2)})/2$.

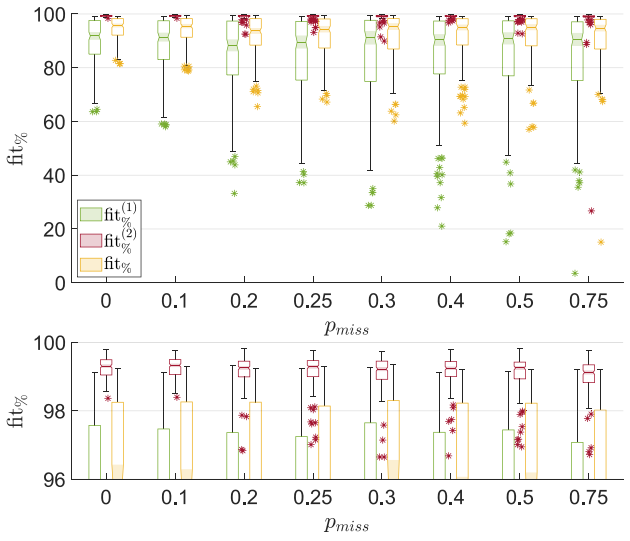


**Fig. 6.** Box-plot illustrating the distribution of the fitness scores for varying percentages of missing data. The bottom plot presents a zoom on the fitness scores for the second output specifically. Each box shows the IQR, with the median as a horizontal line and whiskers extending to 1.5×IQR. Asterisks represent fitness scores that fall outside this range. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 2**
Global fitness scores.

| $p_{\text{miss}}$ | $\text{fit}_{\%}$ (mean $\pm 1\sigma$) | Nuc-SId | PF-NSId |
|---|---|---|---|
| 0 | 94.3 ± 4.84 | 84.7 | 89.0 |
| 10 | 93.7 ± 5.47 | 86.2 | 88.0 |
| 20 | 91.9 ± 7.49 | 85.3 | 87.0 |
| 25 | 91.8 ± 7.61 | / | / |
| 30 | 91.4 ± 8.62 | 85.4 | / |
| 40 | 91.8 ± 8.40 | 85.2 | / |
| 50 | 92.1 ± 8.28 | 83.7 | / |
| 75 | 91.2 ± 9.70 | / | / |

These results demonstrate the robustness of the proposed approach in the case of missing data. Indeed, as the percentage of missing data increases, the average fitness scores remain consistently high, with a gradual decrease only at higher levels of data loss. Moreover, the narrow IQR across simulations for lower values of $p_{\text{miss}}$ indicates low variability in the identified models. As $p_{\text{miss}}$ increases, the IQR becomes wider, reflecting the expected growth in variability due to reduced information availability. Within this context, we remark that robustness does not imply invariance to missing data but rather refers to the ability of the framework to maintain high average performance and reconstruct meaningful system behavior despite increasing $p_{\text{miss}}$. This behavior is indeed consistent with the theoretical error bounds, which grow with $\sqrt{p_{\text{miss}}}$: as the number of available measurements decreases, errors in parameter estimation naturally increase (see Theorem 1), leading to greater variability in prediction quality. Notably, the second output retains better fitness overall, as highlighted in the zoomed section, suggesting a higher resilience of this output to missing observations.

These outcomes are also reflected in Table 2, which presents the global fitness scores, compared to the results obtained in Liu et al. (2013) (Nuc-SId) and Gopaluni et al. (2009) (PF-NSId) for the same amount of missing data. In this case, the results confirm the ability of the approach in maintaining high global fitness scores across varying levels of missing data, outperforming benchmark results obtained with linear (Nuc-SId) and black-box based (PF-SId) identification methods.

Then, the true and predicted trajectories of the CSTR system under four different percentages of missing observations $p_{\text{miss}}$ are compared in Fig. 7, where we have also reported the ±1 standard deviation bands around the mean trajectories. The results highlight the ability of the proposed method to approximate the system's dynamics and confirm its inherent robustness to substantial missing data.

**Table 3**
RMSE scores (mean $\pm 1\sigma$).

| $p_{\text{miss}}$ | $\text{RMSE}_C \times 10^3$ | $\text{RMSE}_T$ |
|---|---|---|
| 0 | 4.54 ± 2.54 | 0.29 ± 0.06 |
| 10 | 4.88 ± 2.85 | 0.29 ± 0.06 |
| 20 | 5.61 ± 3.81 | 0.31 ± 0.09 |
| 25 | 5.80 ± 3.97 | 0.31 ± 0.10 |
| 30 | 6.09 ± 4.76 | 0.32 ± 0.10 |
| 40 | 5.94 ± 4.96 | 0.31 ± 0.07 |
| 50 | 5.76 ± 4.79 | 0.31 ± 0.09 |
| 75 | 6.39 ± 7.60 | 0.34 ± 0.16 |

**Table 4**
RRSE scores.

| $p_{\text{miss}}$ | RRSE (mean $\pm 1\sigma$) | ODE-RSSM | NSM-SId (mean $\pm 1\sigma$) |
|---|---|---|---|
| 0 | 0.0787 ± 0.0161 | 0.0659 | 0.0220 ± 0.005 |
| 10 | 0.0775 ± 0.0152 | / | / |
| 20 | 0.0804 ± 0.0173 | / | / |
| 25 | 0.0785 ± 0.0161 | / | / |
| 30 | 0.0839 ± 0.0192 | / | / |
| 40 | 0.0825 ± 0.0166 | / | / |
| 50 | 0.0834 ± 0.0187 | 0.1336 | 0.0920 ± 0.0140 |
| 75 | 0.0864 ± 0.0249 | 0.2595 | / |

Table 3 collects the RMSE scores for the two outputs. The reported values demonstrate the ability of the proposed framework to maintain low the prediction errors across varying levels of missing data for both the outputs, while the relatively small standard deviations indicate stable performance, even under significant data loss. Then, Table 4 shows a comparison with (Yuan et al., 2023) (ODE-RSSM), and Demeester (2020) (NSM-SId) in terms of the relative root squared error (RRSE), calculated as in Demeester (2020),

$$\text{RRSE} = \frac{1}{n_z}\sum_{i=1}^{n_z}\sqrt{\left(\frac{\sum_{k=0}^{T-1}(\hat{z}_k^{(i)} - \tilde{z}_k^{(i)})^2}{\sum_{k=0}^{T-1}(\hat{z}_k^{(i)} - \frac{1}{T}\sum_{k=0}^{T-1}\tilde{z}_k^{(i)})^2}\right)}.$$

Here, an RRSE of 1 indicates performance equivalent to predicting the mean. For consistency with (Yuan et al., 2023), and Demeester (2020), the outputs and predictions were normalized in this phase by subtracting their mean and dividing by their standard deviation before calculation. In this case, the results reported in Table 4 highlight how the proposed approach is able to provide RRSE values comparable with the ones obtained with the black-box-based benchmark. In particular, better performance are obtained under moderate levels of missing data. Moreover, compared to black-box methods, the proposed approach offers improved interpretability of the identified parameters.

Last, the parameters identified for varying percentages of missing data are reported in Table 5. Here, the results reflect the level of accuracy and consistency of the parameter estimation achievable under different levels of data loss, demonstrating the ability of the framework to accurately recover interpretable system parameters, maintaining relative reliability across different data loss scenarios. However, it is also important to highlight how the identified parameter may differ from the nominal one, as in the case of $\theta_1$. In the considered case study, this discrepancy may be caused by the variations due to environmental and operational conditions, as it commonly happens in real systems.

Summarizing, the application of the proposed approach to the CSTR system in the presence of missing measurements demonstrates its robustness in handling real-world scenarios with substantial missing data. Despite the inherent challenges, the combination of the physics-based model and the black-box component effectively compensates for missing data, accurately identifying the system parameters. Moreover, the obtained results align with those reported in the literature, and a

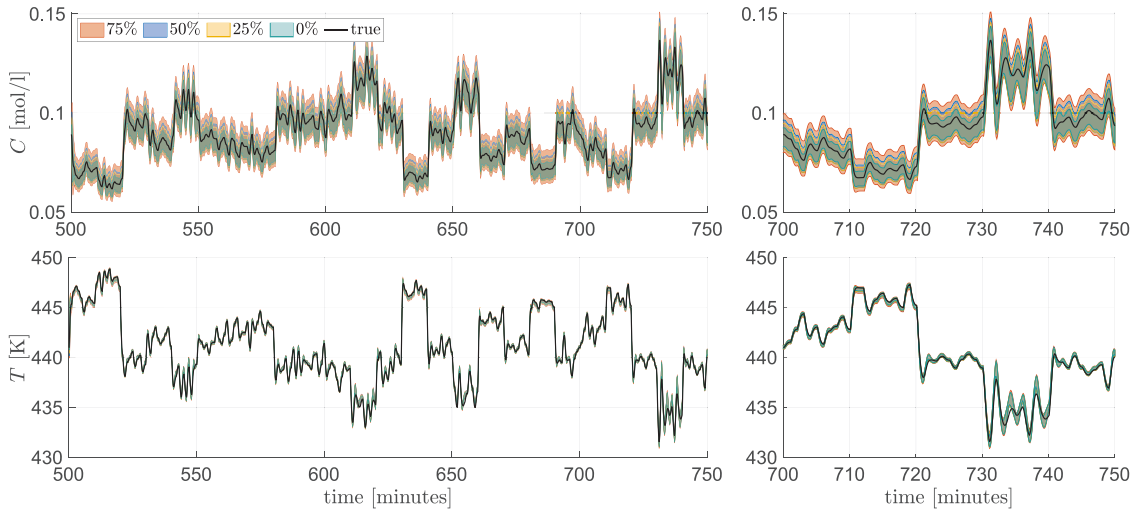


**Fig. 7.** Comparison of true and predicted trajectories for the CSTR system under varying percentages of missing data (indicated in the legend), represented with ±1 standard deviation bands around the mean trajectories. The plots on the right provide a zoomed-in view of the interval [700, 750] minutes to highlight detailed behavior. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 5**
Identified parameters (mean ± 1$\sigma$).

| $p_{\text{miss}}$ | $\theta_1 \times 10^{-10}$ | $\theta_2 \times 10^{-13}$ | $\theta_3 \times 10^{-5}$ |
|---|---|---|---|
| 0 | 7.64 ± 0.40 | −1.44 ± 0.13 | 7.00 ± 0.57 |
| 10 | 7.64 ± 0.37 | −1.45 ± 0.12 | 6.93 ± 0.50 |
| 20 | 7.64 ± 0.40 | −1.44 ± 0.13 | 6.97 ± 0.55 |
| 25 | 7.64 ± 0.43 | −1.44 ± 0.12 | 6.88 ± 0.55 |
| 30 | 7.60 ± 0.43 | −1.42 ± 0.14 | 6.99 ± 0.51 |
| 40 | 7.58 ± 0.40 | −1.43 ± 0.13 | 7.01 ± 0.52 |
| 50 | 7.59 ± 0.46 | −1.42 ± 0.13 | 6.91 ± 0.55 |
| 75 | 7.54 ± 0.42 | −1.41 ± 0.14 | 6.92 ± 0.49 |
| Nominal | 7.20 | −1.44 | 7.00 |

comparison with methods applied to the same benchmark showcases competitive performance particularly under moderate data loss, highlighting the framework's reliability in practical process modeling and its adaptability to real-world conditions.

### 5.2. Identification with averaged observations

In this section, we focus on validating the efficacy of the proposed framework in the case of aggregated observations. In particular, we aim to identify a generic Lotka–Volterra model, which has been largely used to describe the dynamics of a variety of real-world systems.

#### 5.2.1. System description and motivations

The Lotka–Volterra model consists of a set of nonlinear equations commonly used to describe the dynamics of systems involving different interacting species. Specifically, this model describes how the population of the different species varies over time. This model is well-known for describing the dynamics of biological systems, as the interaction between predator and prey populations (Wangersky, 1978). However, its application extends also beyond the ecological domain, as for instance in the economic context, where it is used to represent the wealth of individual investors or the market capitalization of companies (Malcai et al., 2002).

In both ecological and economical framework, the use of averaged measurements is a common practice. For example, significant biological species fluctuations may occur over the year, and sampling on a specific date or during a short period might yield a distorted view of the population's typical behavior (Wangersky, 1978). On the other hand, high-frequency data might be unavailable in the economic context, and only averaged values over extended periods can be observed (Givoly & Palmon, 1982). These averages reflect general trends while concealing short-term variations. Consequently, using aggregated or averaged data within the proposed framework enables the identification of models that fit the available data well, while still capturing detailed dynamics and adapting to the limited resolution of the observations.

#### 5.2.2. Dynamical model

The discretized Lotka–Volterra model with $n_x = 2$ states $x_k = [x_{1,k}, x_{2,k}]^\top$ and parameters $\theta = [\theta_1, \theta_2, \theta_3, \theta_4]^\top$ is given by

$$\begin{aligned} x_{1,k+1} &= x_{1,k} + \theta_1 x_{1,k} - \theta_2 x_{1,k} x_{2,k} + \Delta_1(x_k), \\ x_{2,k+1} &= x_{2,k} - \theta_3 x_{2,k} + \theta_4 x_{1,k} x_{2,k} + \Delta_2(x_k), \\ z_k &= x_k, \end{aligned} \tag{21}$$

where $x_1$ is the population density of prey, $x_2$ is the population density of the predator, $\theta_1$ and $\theta_2$ are the prey's parameters, describing the maximum per capita growth rate and the effect of predators on the prey growth rate, respectively, $\theta_3$ and $\theta_4$ are the predator's parameters, describing the per capita death rate and the effect of prey on the predator's growth rate, respectively. All parameters are positive and real. On the other hand, $\Delta_1$ and $\Delta_2$ represent unmodeled dynamics that capture external factors affecting the populations beyond the basic predator–prey interaction. In the considered case study we have $\theta = [0.13, 0.02, 0.12, 0.02]$, while $\Delta_1$ and $\Delta_2$ are quadratic terms that may represent, e.g., the intraspecific competition within each population, implying that the growth of each population is influenced not only by the interaction between predator and prey but also by the density-dependent effects within each population. In particular, we considered

$$\Delta_1(x_k) = \mu 10^{-4} x_{1,k}^2, \quad \Delta_2(x_k) = -\mu 5 \cdot 10^{-4} x_{2,k}^2,$$

where $\mu > 0$ is a tuning parameter to control the size of the unmodeled terms. In this case study, $\mu = 10$ is selected to introduce a meaningful but interpretable model mismatch, preserving the structure of the physical model while sufficiently challenging the black-box compensator.[4] Clearly, the unmodeled dynamics, being unknown, cannot be incorporated into the physical model. Instead, it must be compensated by the black-box approximator $\delta$. Notice that, although the unmodeled dynamics seem relatively small, the impact on the population dynamics

[4] Larger values of $\mu$ were found to excessively distort the system dynamics, compromising interpretability and realism at the base of the proposed case study.

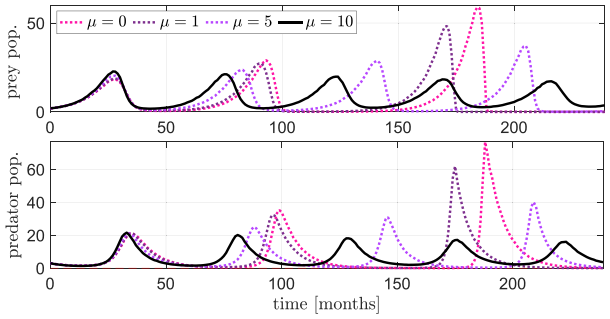


**Fig. 8.** Populations evolution for different values of $\mu$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

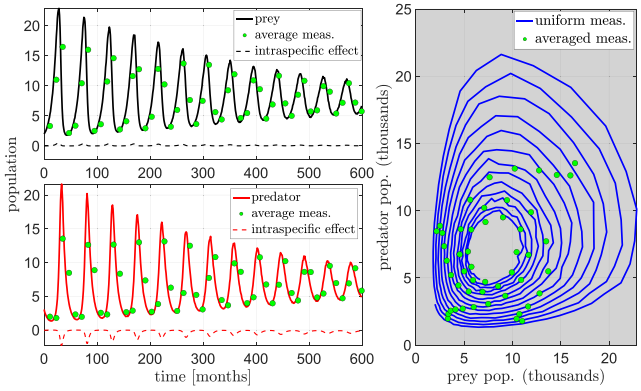


**Fig. 9.** Monthly prey (black lines) and predator (red lines) populations evolutions with yearly average measurements (green circles). Dashed lines represent the unmodeled dynamics. On the right, the phase plot of the system with uniform measurements (blue lines) and average measurements is shown. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

is relevant. This is highlighted in Fig. 8, where the evolution of predator and prey populations is represented over a period of 20 years for different values of $\mu$. In this case, it is evident the importance of efficiently compensating for unmodeled dynamics in order to accurately capture the system's behavior.

As in the previous case study, we conduct a numerical verification of Assumption 1 using the Gauss–Newton approximation and the finite-difference methods. The resulting Hessians are positive definite, with condition numbers around $10^3$, indicating good numerical conditioning.

### 5.2.3. Identification results

In the proposed example, we simulate the evolution of the predator and prey populations based on the Lotka–Volterra model (21) over 75 years (900 months) with initial condition $x_0 = [2, 3]^\top$. The data are generated for both populations at *monthly intervals*, capturing the interaction dynamics described by the model. The first 600 months are used as identification data, while the subsequent 300 months are used for validation. In the considered scenario, the measures averaged over time windows of $T_r = 12$ months are exploited, leading to $M = 50$ identification measurements for each population. Fig. 9 illustrates the monthly population evolution and the yearly average evolution of the prey and predator populations over the entire 50-year identification period. This figure highlights how data averaging captures the overall trend while masking finer details and short-term interactions, which can pose challenges for accurately identifying the system's underlying dynamics.

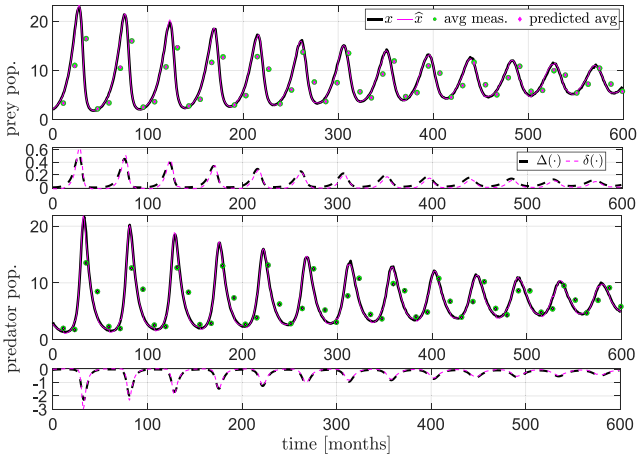


**Fig. 10.** Comparison between the true evolution of the populations (black lines) and the one predicted (purple lines) using the model identified from averaged measurements on the identification data. Green markers represent the averaged measurements, while purple markers represent the reconstructed averages. Purple dashed lines indicate the unmodeled dynamics predicted by the black-box term $\delta(\cdot)$, compared with the true one, $\Delta(\cdot)$ (black dashed line). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

First, we exploit the results of Theorem 2 to identify the system by employing an extended model of the form

$$x_{1,k+1} = x_{1,k} + \theta_1 x_{1,k} - \theta_2 x_{1,k} x_{2,k} + \delta_1(x_k),$$
$$x_{2,k+1} = x_{2,k} - \theta_3 x_{2,k} + \theta_4 x_{1,k} x_{2,k} + \delta_2(x_k),$$
$$c_{k+1} = c_k + x_k,\ \bar{z}_k = \frac{1}{T_r} c_k,$$

to estimate the underlying parameters of the predator–prey system from averaged observations. Then, the identification task is performed by minimizing a cost function of the form (17) considering $M$ multiple runs and $\boldsymbol{\kappa}_1 = \{T_r\}$.

The estimated parameters are initialized randomly as $\widehat{\theta}_0 = \theta + \mathcal{N}(0.02, \sigma_\theta)$, with $\sigma_\theta = 0.05$. Analogously, the initial conditions of the states are initialized at $\hat{x}_{0,0} = \bar{z}_0$. Also in this case, the black-box term $\delta$ is defined as a linear combination of selected basis functions, i.e., sigmoid, softplus, hyperbolic tangent, and trigonometric functions. Additionally, the cost function $C_T$ incorporates physical penalties and regularization terms to enforce specific properties. Specifically, the positivity of $\widehat{\theta}$ is ensured using an exponential barrier function, while the sparsity of the black-box component $\delta$ is promoted through an $\ell_1$-norm approximation applied to the black-box weights $\omega$ (Donati et al., 2025a).

Figs. 10 and 11 showcase the predictions from the identified model compared with the averaged observations and the population behavior for identification and validation data, respectively. The results highlight how the proposed approach is able to successfully reconstructs the predator and prey dynamics based on the available averaged data and demonstrate the accuracy of the identified model even when facing aggregated measurements.

The accuracy of the method is also reflected in to adherence of the identified parameters and the state initial condition, i.e., $\widehat{\theta} = [0.1318, 0.0204, 0.1214, 0.0198]^\top$ and $\hat{x}_0 = [2.0842, 2.7412]^\top$, with the ground truth values used in the simulation, i.e., $\theta = [0.13, 0.02, 0.12, 0.02]$ and $x_0 = [2, 3]^\top$.

Next, we extend the preliminary analysis by considering different sizes of averaging windows. Hence, in addition to the $T_r = 12$-month window (yielding $M = 50$ averaged data points), we fixed the total amount of data ($T = 900$) and the starting value of the estimated parameters and initial condition. Then, we analyze the identification outcomes using larger windows $T_r$, which imply a lower number of

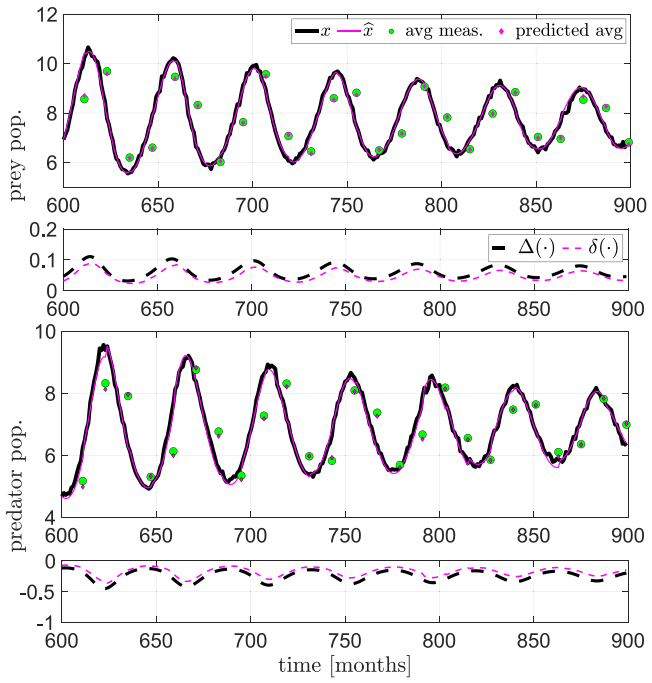


**Fig. 11.** True population evolution (black lines) with the averaged measurements (green markers) and predicted population evolution (purple lines) with the reconstructed averages (purple markers) using the model identified from averaged measurements on the validation data. Black and purple dashed lines represent $\Delta(\cdot)$ and $\delta(\cdot)$, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 6**
Effect of the averaging window size $T_r$ on identification performance.

| $T_r$ | $M_{\text{tr}}/M_{\text{val}}$ | $\text{RMSE}_{\text{tr}}$ | $\text{RMSE}_{\text{val}}$ | $\|\theta-\widehat{\theta}\|_2$ | $\|x_0-\widehat{x}_0\|_2$ |
|---|---|---|---|---|---|
| 12 | 50/25 | 0.2737 | 0.3436 | 0.0023 | 0.2721 |
| 15 | 40/20 | 0.2183 | 0.2846 | 0.0029 | 0.2047 |
| 20 | 30/15 | 0.4742 | 0.5486 | 0.0123 | 0.5718 |
| 24 | 25/13 | 0.8443 | 0.8502 | 0.0217 | 0.9885 |
| 40 | 15/8 | 1.4328 | 1.4242 | 0.0442 | 1.1534 |
| 50 | 12/6 | 1.3346 | 1.3539 | 0.0589 | 0.9157 |

available average measurements. Table 6 summarizes the results for each simulated case, including the prediction accuracy measured in terms of the root mean square error,[5] and parametric error for both $\theta$ and $x_0$, for each window size.

From the reported results, we can observe that, being $T = MT_r$ fixed, an increase in $T_r$ results in fewer available measurements ($M$), which generally leads to less accurate estimates of the parameters and initial conditions, as indicated by increasing error values. Furthermore, a larger $T_r$ not only decreases the data available for identification but also enhances the smoothing effect on short-term dynamics, due to averaging over more values. Consequently, this implies a reduced accuracy, as reflected in the observed trends. Last, it is worth noting that the similar error values observed for $T_r = 12$ and $T_r = 15$ suggest a range where the performance remains relatively stable, indicating the presence of an "optimal" averaging window size, beyond which performance begins to degrade. This aspect will be explored in future analysis.

[5] The RMSE has been computed considering the entire set of observations and predictions over the horizon.

## 6. Conclusions

In this paper, we presented an identification framework able to handle non-uniform observations. The proposed physics-based approach, defined by a proper combination of off-white models and black-box approximators, is designed to address real-world scenarios where measurements may be missing, aggregated, or collected across multiple experimental runs. Specifically, for the case of missing measurements and multiple runs, we show how a minimal modification of the cost function allows to handle this class of non-uniform data. On the other hand, for aggregated — cumulative or averaged — measurements, we show how the identification problem can be handled by relying on an extended system model that reinterprets this problem as one combining missing measurements and multiple runs. In this way, we showcase the high flexibility of the approach to deal with non-uniform data.

Moreover, a theoretical analysis has been carried out to assess the effect of non-uniform observations on the estimation accuracy. We demonstrate that for missing measurement there exists an upper bound on the parametric error that depends on the percentage of missing data, the length of the observation horizon, and the resulting error. Analogously, we prove that a similar bound also exists for the case of aggregated observations, highlighting how the length of the aggregation window impacts the accuracy of parameter estimation.

The effectiveness of the proposed approach is demonstrated through two case studies, involving two different types of non-uniform observations. First, we assess the performance of the approach over a scenario with missing measurements involving the identification of a continuous stirred-tank reactor. Then, we test the proposed framework in the presence of aggregated data for a Lotka–Volterra system. In both case studies, we demonstrate the ability of the proposed approach to accurately reconstruct system dynamics under non-uniform observations.

## CRediT authorship contribution statement

**Cesare Donati:** Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Martina Mammarella:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Fabrizio Dabbene:** Writing – review & editing, Validation, Supervision, Project administration, Formal analysis, Conceptualization. **Carlo Novara:** Writing – review & editing, Validation, Supervision, Project administration, Formal analysis, Conceptualization. **Constantino Lagoa:** Writing – review & editing, Validation, Supervision, Project administration, Formal analysis, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgment

C. Donati acknowledges support from PRIN project TECHIE "A control and network-based approach for fostering the adoption of new technologies in the ecological transition", Cod. 2022KPHA24 CUP: D53D23001320006.

## Appendix A. Proof to Theorem 1

Let us consider a cost function of the form

$$C(\theta,\gamma)=\gamma^\top\zeta_T, \tag{22}$$

where $\gamma\in\mathbb{R}^T$ is a generic vector of real coefficient and $\zeta_T\in\mathbb{R}^T$ is a vector containing the squared $\ell_2$-norm of the prediction error for each element, i.e., $\zeta_{T,k}=\|\widetilde{z}_k-\widehat{z}_k\|_2^2$. Now, since $C$ depends on $\gamma$, we notice that also a minimizer of this cost is a function of $\gamma$. Hence, denoting with $\theta^\star=[\theta_1^\star,\dots,\theta_{n_\theta}^\star]$ the solution obtained by minimizing a cost function of the type (22), we have also that $\theta^\star\equiv\theta^\star(\gamma)$. Then, given $\gamma_1\neq\gamma_2$, we define the associated cost functions according to (22), i.e.,

$$C_1(\theta,\gamma)=\gamma_1^\top\zeta_T, \tag{23a}$$

$$C_2(\theta,\gamma)=\gamma_2^\top\zeta_T, \tag{23b}$$

Therefore, considering the solutions of (23a) and (23b), and applying the mean value theorem, for each parameter $\theta_i^\star(\cdot)$ there exists a vector $\breve{\gamma}^{(i)}=(1-a)\gamma_1+a\gamma_2$, with $a\in[0,1]$, such that

$$\theta_i^\star(\gamma_1)-\theta_i^\star(\gamma_2)=\xi(\breve{\gamma}^{(i)})^\top(\gamma_1-\gamma_2), \tag{24}$$

with $\xi(\gamma)\doteq\frac{\partial\theta_i^\star(\gamma)}{\partial\gamma}\in\mathbb{R}^T$.

Then, to compute $\frac{\partial\theta^\star(\cdot)}{\partial\gamma}\in\mathbb{R}^{n_\theta,T}$ we rely on the implicit differentiation technique. First, we consider that the minimizer $\theta^\star$ is a solution of $\frac{\partial C_T(\theta,\gamma)}{\partial\theta}=0$. It follows that

$$\frac{\mathrm{d}}{\mathrm{d}\gamma}\frac{\partial C_T(\theta^\star,\gamma)}{\partial\theta}=0, \tag{25}$$

where $\frac{\mathrm{d}}{\mathrm{d}\gamma}$ denotes the Jacobian with respect to $\gamma$, given by

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\gamma}\frac{\partial C_T(\theta^\star,\gamma)}{\partial\theta}&=\frac{\partial^2 C_T(\theta^\star,\gamma)}{\partial\gamma\partial\theta}+\frac{\partial^2 C_T(\theta^\star,\gamma)}{\partial^2\theta}\frac{\partial\theta^\star(\gamma)}{\partial\gamma}\\&=G(\gamma)+H(\gamma)\frac{\partial\theta^\star(\gamma)}{\partial\gamma},\end{aligned} \tag{26}$$

where $H$ is the Hessian matrix of the cost function with respect to $\theta$, and $G$ reflects the influence of missing data on the system's behavior. Therefore, considering (25), (26), and knowing that the Hessian $H$ is invertible according to Assumption 1, we obtain

$$\frac{\partial\theta^\star(\gamma)}{\partial\gamma}=-H^{-1}G(\gamma).$$

This implies that, for the $i$th parameter $\theta_i^\star$, the vector $\xi(\breve{\gamma}^{(i)})^\top$ in (24) corresponds to the $i$th row of the matrix $-H^{-1}G(\breve{\gamma}^{(i)})\in\mathbb{R}^{n_\theta,T}$. Thus, according to (24), we have

$$\theta^\star(\gamma_1)-\theta^\star(\gamma_2)=\Xi^\top(\gamma_1-\gamma_2),$$

with $\Xi=\left[\xi(\breve{\gamma}^{(1)}),\dots,\xi(\breve{\gamma}^{(n_\theta)})\right]\in\mathbb{R}^{T,n_\theta}$, that yields

$$\begin{aligned}\|\theta^\star(\gamma_1)-\theta^\star(\gamma_2)\|_2&=\|\Xi^\top(\gamma_1-\gamma_2)\|_2\\&\leq\|\Xi^\top\|_2\|(\gamma_1-\gamma_2)\|_2.\end{aligned} \tag{27}$$

Let us now consider the case of missing measurements and a set of available time-steps $\boldsymbol{\kappa}_N$ (6). Specifically, let us define the coefficient vectors in (23a) and (23b) as

$$\gamma_1=\gamma_T\doteq\left[\frac{1}{T},\dots,\frac{1}{T}\right]^\top \tag{28a}$$

$$\gamma_2=\gamma_N=\left[\gamma_{N,i}\right],\ \gamma_{N,i}\doteq\begin{cases}0 & \text{if } i\notin\boldsymbol{\kappa}_N,\\ \frac{1}{T} & \text{otherwise.}\end{cases} \tag{28b}$$

Then, for the coefficient vector $\gamma_T$, the associated cost function $C(\theta,\gamma_T)$ of the form (22) is equivalent to the cost function in (5). Indeed, we have

$$\begin{aligned}C(\theta,\gamma_T)=\gamma_T^\top\zeta_T&=\left[\frac{1}{T},\dots,\frac{1}{T}\right]\begin{bmatrix}\|\widetilde{z}_0-\widehat{z}_0\|_2^2\\ \vdots\\ \|\widetilde{z}_{T-1}-\widehat{z}_{T-1}\|_2^2\end{bmatrix}\\&=\sum_{k=0}^{T-1}\frac{1}{T}\|\widetilde{z}_k-\widehat{z}_k\|_2^2=\frac{1}{T}\sum_{k=0}^{T-1}\|\widetilde{z}_k-\widehat{z}_k\|_2^2.\end{aligned} \tag{29}$$

Analogously, for the coefficient vector $\gamma_N$,[6] defined in (28b) the associated cost function $C(\theta,\gamma_N)$ of the form (22) is equivalent to one given by (8), i.e.,

$$C(\theta,\gamma_N)=\gamma_N^\top\zeta_T=\frac{1}{T}\sum_{j=0}^{N}\|\widetilde{z}_{k_j}-\widehat{z}_{k_j}\|_2^2, \tag{30}$$

where $k_j$ is the $j$th element of $\boldsymbol{\kappa}_N$. Hence, according to (29) and (30), it follows that (27) holds for the solutions $\theta_T^\star$ and $\theta_N^\star$ of (5) and (8), respectively. Moreover, defining $\sigma_\xi$ as the maximum singular value of the matrix $\Xi^\top$, i.e., $\sigma_\xi=\|\Xi^\top\|_2\doteq\sigma_{max}(\Xi^\top)$, we obtain

$$\|\theta^\star(\gamma_T)-\theta^\star(\gamma_N)\|_2=\|\theta_T^\star-\theta_N^\star\|_2\leq\sigma_\xi\|\gamma_T-\gamma_N\|_2. \tag{31}$$

Then, observing that

$$\|(\gamma_T-\gamma_N)\|_2=\sqrt{\frac{T-N}{T^2}}=\frac{1}{\sqrt{T}}\sqrt{p_{\text{miss}}}, \tag{32}$$

and combining (31) with (32), we yield (9), concluding the proof.

## Appendix B. Proof to Theorem 2

First, we consider the extended system (14) for a generic run[7] $i$ of length $T_r+1$ with missing measurements defined by $\boldsymbol{\kappa}_1=T_r$. From (7) we have

$$\bar{\mathbf{z}}_{[\boldsymbol{\kappa}_1]}=\{\bar{z}_{k_1}\}=\bar{z}_{T_r}. \tag{33}$$

Moreover, from (14b)–(14c), we obtain

$$\bar{z}_{T_r}=\alpha c_{T_r}=\alpha c_{T_r-1}+\alpha h(x_{T_r-1};\theta). \tag{34}$$

Then, iterating (14b) backward, we obtain

$$\begin{aligned}c_{T_r-1}&=c_{T_r-2}+h(x_{T_r-2};\theta)\\&=c_{T_r-3}+h(x_{T_r-3};\theta)+h(x_{T_r-2};\theta)\\&\ \ \vdots\\&=c_0+h(x_0;\theta)+\cdots+h(x_{T_r-2};\theta).\end{aligned} \tag{35}$$

Now, from (34) and (35) it follows that

$$\bar{z}_{T_r}=\alpha h(x_0;\theta)+\cdots+\alpha h(x_{T_r-1};\theta),$$

having $c_0=0_{n_z}$ by definition. Moreover, from (1) we have $z_k=h(x_k;\theta)$, which implies that

$$\bar{z}_{T_r}=\alpha z_0+\cdots+\alpha z_{T_r-1}=\alpha\sum_{k=0}^{T_r-1}z_k. \tag{36}$$

Finally, considering (13) and (33), we have that (36) implies (15), which concludes the proof.

## Appendix C. Proof to Theorem 3

For simplicity and without loss of generality, let us consider the case of $n_z=1$. The extension to the general case is straightforward. Given the prediction error $e_k=\widetilde{z}_k-\widehat{z}_k$, let us consider the vector

$$\varepsilon=\left[e_0,\dots,e_{T-1}\right]^\top\in\mathbb{R}^T.$$

Moreover, consider the following cost function

$$C(\Gamma,\theta)=p\|\Gamma\varepsilon\|_2^2,$$

with $p\in\mathbb{R}$ a generic constant, and $\Gamma\in\mathbb{R}^{T,T}$ an aggregation matrix of coefficients. Given that $C$ depends on $\Gamma$, any minimizer of the cost

[6] Notice that when $N=0$, i.e., $p_{\text{miss}}=1$, the system becomes non-identifiable, as $C_T(\theta,\cdot)=0$ for all $\theta$, thereby violating Assumption 1.

[7] The superscript $(i)$ is omitted for clarity.

function $C$ will also be a function of $\Gamma$. Therefore, $\theta^\star$, i.e., the solution obtained by minimizing $C$, can be explicitly expressed as

$$\theta^\star \equiv \theta^\star(\Gamma). \tag{37}$$

Thus, considering the cost function $C(\Gamma,\theta)$ in (5) when $T$ uniform measurements are available, we have

$$C_T = \frac{1}{T}\sum_{k=0}^{T-1}\|e_k\|_2^2 = \frac{1}{T}\|\Gamma_T\varepsilon\|_2^2 \doteq C(\Gamma_T,\theta), \tag{38}$$

with $\Gamma_T = \mathbb{I}_T$. Similarly, considering $M$ aggregated measurements with windows length $T_r$, as defined in (13), and applying (36), we can rewrite (17) as

$$\begin{aligned} C_T &= \frac{1}{M}\sum_{i=1}^{M}\left\|\alpha\sum_{k=0}^{T_r-1}\tilde{z}_k^{(i)} - \alpha\sum_{k=0}^{T_r-1}\hat{z}_k^{(i)}\right\|_2^2 \\ &= \frac{\alpha^2}{M}\sum_{i=1}^{M}\left\|\sum_{k=0}^{T_r-1}(\tilde{z}_k^{(i)} - \hat{z}_k^{(i)})\right\|_2^2 \\ &= \frac{\alpha^2}{M}\left\|\Gamma_{T_r}\varepsilon\right\|_2^2 \doteq C(\Gamma_{T_r},\theta), \end{aligned} \tag{39}$$

with $\tilde{z}_k^{(i)} = z_k^{(i)} + \frac{\eta_i^Z}{T_r}$, $\Gamma_{T_r} = \left[g_1,\dots,g_T\right]^\top$, and

$$g_j = \begin{cases} \left[0_{(j-1)T_r}^\top, 1_{T_r}^\top, 0_{T-jT_r}^\top\right]^\top & \text{if } j \le M, \\ 0_T & \text{otherwise.} \end{cases} \tag{40}$$

Thus, being $C(\Gamma,\theta)$ twice continuously differentiable and the system identifiable according to Assumption 1, by applying the implicit function theorem (Krantz & Parks, 2002) to the gradient function $\nabla_\theta C(\Gamma,\theta(\Gamma)) = \frac{\partial C(\Gamma,\theta(\Gamma))}{\partial\theta}$ it follows that $\theta(\Gamma)$ is continuously differentiable, and consequently also Lipschitz continuous, with respect to $\Gamma$. Now, let us consider $\theta_T^\star$ and $\theta_{T_r}^\star$, i.e., the minimizers of (5) and (17), respectively. According to (37)–(39), we have $\theta_T^\star = \theta^\star(\Gamma_T)$ and $\theta_{T_r}^\star = \theta^\star(\Gamma_{T_r})$. Therefore, from Lipschitz continuity it follows that

$$\|\theta_T^\star - \theta_{T_r}^\star\|_2 \le L_\theta\|\Gamma_T - \Gamma_{T_r}\|_2, \tag{41}$$

for some Lipschitz constant $L_\theta$. Then, defining $\beta_{T_r}$ as the maximum singular value of the matrix $\Gamma_T - \Gamma_{T_r}$, we have

$$\|\Gamma_T - \Gamma_{T_r}\|_2 = \sigma_{max}(\Gamma_T - \Gamma_{T_r}) = \beta_{T_r},$$

which, combined with (41), yields (18). Thus, applying triangle inequality, the following relation holds, i.e.,

$$\begin{aligned} \beta_{T_r} = \|\Gamma_T - \Gamma_{T_r}\|_2 &\le \|\Gamma_T\|_2 + \| - \Gamma_{T_r}\|_2 \\ &= \|\Gamma_T\|_2 + \|\Gamma_{T_r}\|_2 \\ &= \sigma_{max}(\Gamma_T) + \sigma_{max}(\Gamma_{T_r}). \end{aligned} \tag{42}$$

Similarly, applying the reverse triangle inequality

$$\begin{aligned} \beta_{T_r} = \|\Gamma_T - \Gamma_{T_r}\|_2 &\ge |\|\Gamma_T\|_2 - \|\Gamma_{T_r}\|_2| \\ &= |\sigma_{max}(\Gamma_T) - \sigma_{max}(\Gamma_{T_r})|. \end{aligned} \tag{43}$$

Here, it is easy to verify that $\sigma_{max}(\Gamma_T) = \sigma_{max}(\mathbb{I}_T) = 1$. Moreover, the following relation holds for all $T_r$, i.e.,

$$\sigma_{max}(\Gamma_{T_r}) = \sqrt{T_r}. \tag{44}$$

In particular, we have that $\sigma_{max}(\Gamma_{T_r}) = \sqrt{\lambda_{max}(\Gamma_{T_r}\Gamma_{T_r}^\top)}$, where

$$\Gamma_{T_r}\Gamma_{T_r}^\top = \begin{bmatrix} g_1^\top \\ \vdots \\ g_T^\top \end{bmatrix}\begin{bmatrix} g_1,\dots,g_T \end{bmatrix} = \begin{bmatrix} g_1^\top g_1 & \cdots & g_1^\top g_T \\ \vdots & \ddots & \vdots \\ g_T^\top g_1 & \cdots & g_T^\top g_T \end{bmatrix}.$$

Here, according to (40), it is easy to verify that $g_{i,j} = 0$, $\forall i > M$, $\forall j$, and

$$g_i^\top g_j = \begin{cases} 1_{T_r}^\top 1_{T_r} = T_r & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

It follows that

$$\Gamma_{T_r}\Gamma_{T_r}^\top = \begin{bmatrix} T_r\mathbb{I}_M & \mathbb{0}_{M,T-M} \\ \mathbb{0}_{T-M,M} & \mathbb{0}_{T-M,T-M} \end{bmatrix},$$

is a diagonal matrix, where $\lambda_1 = \cdots = \lambda_M = T_r$, $\lambda_{M+1} = \cdots = \lambda_T = 0$, and $\sigma_{max}(\Gamma_{T_r}) = \sqrt{T_r}$, proving the statement (44). Thus, from (42)–(44), we have

$$|1 - \sqrt{T_r}| \le \beta_{T_r} \le \sqrt{T_r} + 1,$$

which leads to (19) having $T_r \ge 1$, concluding the proof. Notice that $\beta_{T_r} \approx \sqrt{T_r}$ for large $T_r$.

## References


Aguero, J. C., Goodwin, G. C., & Yuz, J. I. (2007). System identification using quantized data. In *2007 46th IEEE conference on decision and control* (pp. 4263–4268). IEEE, http://dx.doi.org/10.1109/CDC.2007.4434350.

Amin, A., & Mourshed, M. (2024). Weather and climate data for energy applications. *Renewable and Sustainable Energy Reviews*, *192*, Article 114247. http://dx.doi.org/10.1016/j.rser.2023.114247.

Bellman, R., & Åström, K. J. (1970). On structural identifiability. *Mathematical Biosciences*, *7*(3–4), 329–339. http://dx.doi.org/10.1016/0025-5564(70)90132-X.

Björck, Å. (1996). Numerical methods for least squares problems. Society for Industrial and Applied Mathematics, http://dx.doi.org/10.1137/1.9781611971484.

De Moor, B., De Gersem, P., De Schutter, B., Favoreel, W., et al. (1997). DAISY: A database for identification of systems. *Journal A*, *38*(4), 4–5.

Demeester, T. (2020). System identification with time-aware neural sequence models. In *Proceedings of the AAAI conference on artificial intelligence*: *vol. 34*, (pp. 3757–3764). http://dx.doi.org/10.1609/aaai.v34i04.5786.

Deng, J., & Huang, B. (2012). Identification of nonlinear parameter varying systems with missing output data. *AIChE Journal*, *58*(11), 3454–3467. http://dx.doi.org/10.1002/aic.13735.

Donati, C., Mammarella, M., Dabbene, F., Novara, C., & Lagoa, C. (2024). One-shot backpropagation for multi-step prediction in physics-based system identification. *20th IFAC Symposium on System Identification (SYSID)*, http://dx.doi.org/10.1016/j.ifacol.2024.08.540.

Donati, C., Mammarella, M., Dabbene, F., Novara, C., & Lagoa, C. (2025). Combining off-white and sparse black models in multi-step physics-based systems identification. *Automatica*, Article 112409. http://dx.doi.org/10.1016/j.automatica.2025.112409.

Donati, C., Mammarella, M., Dabbene, F., Novara, C., & Lagoa, C. (2025). A scalable, gradient-stable approach to multi-step, nonlinear system identification using first-order methods. *6th IFAC Workshop on Linear Parameter Varying Systems*, http://dx.doi.org/10.48550/arXiv.2410.03544, (in press).

Fan, J., Chow, T. W., & Qin, S. J. (2021). Kernel-based statistical process monitoring and fault detection in the presence of missing data. *IEEE Transactions on Industrial Informatics*, *18*(7), 4477–4487. http://dx.doi.org/10.1109/TII.2021.3119377.

Ghadimi, S., & Lan, G. (2016). Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, *156*(1), 59–99. http://dx.doi.org/10.1007/s10107-015-0871-8.

Givoly, D., & Palmon, D. (1982). Timeliness of annual earnings announcements: Some empirical evidence. *Accounting Review*, 486–508.

Gopaluni, R. (2008). A particle filter approach to identification of nonlinear processes under missing observations. *Canadian Journal of Chemical Engineering*, *86*(6), 1081–1092. http://dx.doi.org/10.1002/cjce.20113.

Gopaluni, R. B. (2010). Nonlinear system identification under missing observations: The case of unknown model structure. *Journal of Process Control*, *20*(3), 314–324. http://dx.doi.org/10.1016/j.jprocont.2009.12.008.

Gopaluni, R. B., Schön, T. B., & Wills, A. G. (2009). Particle filter approach to nonlinear system identification under missing observations with a real application. *IFAC Proceedings Volumes*, *42*(10), 810–815. http://dx.doi.org/10.3182/20090706-3-FR-2004.00134.

Grossmann, C., Jones, C. N., & Morari, M. (2009). System identification via nuclear norm regularization for simulated moving bed processes from incomplete data sets. In *Proceedings of the 48h IEEE conference on decision and control (CDC) held jointly with 2009 28th Chinese control conference* (pp. 4692–4697). IEEE, http://dx.doi.org/10.1109/CDC.2009.5400711.

Huntley, H. S., & Hakim, G. J. (2010). Assimilation of time-averaged observations in a quasi-geostrophic atmospheric jet model. *Climate Dynamics*, *35*, 995–1009. http://dx.doi.org/10.1007/s00382-009-0714-5.

Isaksson, A. J. (1993). Identification of ARX-models subject to missing data. *IEEE Transactions on Automatic Control*, *38*(5), 813–819. http://dx.doi.org/10.1109/9.277253.

Kidwell, S. M., & Tomasovych, A. (2013). Implications of time-averaged death assemblages for ecology and conservation biology. *Annual Review of Ecology, Evolution, and Systematics*, *44*(1), 539–563. http://dx.doi.org/10.1146/annurev-ecolsys-110512-135838.

Krantz, S. G., & Parks, H. R. (2002). *The implicit function theorem: History, theory, and applications*. Springer Science & Business Media, http://dx.doi.org/10.1007/978-1-4614-5981-1.

Liu, Z., Hansson, A., & Vandenberghe, L. (2013). Nuclear norm system identification with missing inputs and outputs. *Systems & Control Letters*, *62*(8), 605–612. http://dx.doi.org/10.1016/j.sysconle.2013.04.005.

Ljung, L. (2010). Perspectives on system identification. *Annual Reviews in Control*, *34*(1), 1–12. http://dx.doi.org/10.1016/j.arcontrol.2009.12.001.

Lu, Y., & Huang, B. (2014). Robust multiple-model LPV approach to nonlinear process identification using mixture t distributions. *Journal of Process Control*, *24*(9), 1472–1488. http://dx.doi.org/10.1016/j.jprocont.2014.06.018.

Malcai, O., Biham, O., Richmond, P., & Solomon, S. (2002). Theoretical analysis and simulations of the generalized Lotka-Volterra model. *Physical Review E*, *66*(3), Article 031102. http://dx.doi.org/10.1103/PhysRevE.66.031102.

Markovsky, I. (2013). A software package for system identification in the behavioral setting. *Control Engineering Practice*, *21*(10), 1422–1436. http://dx.doi.org/10.1016/j.conengprac.2013.06.010.

Morningred, J. D., Paden, B. E., Seborg, D. E., & Mellichamp, D. A. (1992). An adaptive nonlinear predictive controller. *Chemical Engineering Science*, *47*(4), 755–762. http://dx.doi.org/10.1016/0009-2509(92)80266-F.

Nakagawa, S., & Freckleton, R. P. (2011). Model averaging, missing data and multiple imputation: A case study for behavioural ecology. *Behavioral Ecology and Sociobiology*, *65*, 103–116. http://dx.doi.org/10.1007/s00265-010-1044-7.

Pillonetto, G., Aravkin, A., Gedon, D., Ljung, L., Ribeiro, A. H., & Schön, T. B. (2025). Deep networks for system identification: a survey. *Automatica*, *171*, Article 111907. http://dx.doi.org/10.1016/j.automatica.2024.111907.

Raghavan, H., Tangirala, A. K., Gopaluni, R. B., & Shah, S. L. (2006). Identification of chemical processes with irregular output sampling. *Control Engineering Practice*, *14*(5), 467–480. http://dx.doi.org/10.1016/j.conengprac.2005.01.015.

Ribeiro, A. H., Tiels, K., Umenberger, J., Schön, T. B., & Aguirre, L. A. (2020). On the smoothness of nonlinear system identification. *Automatica*, *121*, Article 109158. http://dx.doi.org/10.1016/j.automatica.2020.109158.

Sleem, O. M., & Lagoa, C. M. (2024). Parsimonious system identification from fragmented quantised measurements. *International Journal of Control*, *97*(8), 1770–1779. http://dx.doi.org/10.1080/00207179.2023.2228943.

Varanasi, S. K., & Jampana, P. (2020). Nuclear norm subspace identification of continuous time state–space models with missing outputs. *Control Engineering Practice*, *95*, Article 104239. http://dx.doi.org/10.1016/j.conengprac.2019.104239.

Wangersky, P. J. (1978). Lotka-Volterra population models. *Annual Review of Ecology and Systematics*, *9*, 189–218. http://dx.doi.org/10.1146/annurev.es.09.110178.001201.

Yang, X., Liu, X., & Yin, S. (2018). Robust identification of nonlinear systems with missing observations: The case of state-space model structure. *IEEE Transactions on Industrial Informatics*, *15*(5), 2763–2774. http://dx.doi.org/10.1109/TII.2018.2871194.

Yılmaz, B., Bekiroglu, K., Lagoa, C., & Sznaier, M. (2018). A randomized algorithm for parsimonious model identification. *IEEE Transactions on Automatic Control*, *63*(2), 532–539. http://dx.doi.org/10.1109/TAC.2017.2723959.

Yu, Y., Li, V. O., Lam, J. C., & Chan, K. (2023). GCN-ST-MDIR: Graph convolutional network-based spatial-temporal missing air pollution data pattern identification and recovery. *IEEE Transactions on Big Data*, *9*(5), 1347–1364. http://dx.doi.org/10.1109/TBDATA.2023.3277710.

Yuan, Z., Ban, X., Zhang, Z., Li, X., & Dai, H.-N. (2023). ODE-RSSM: learning stochastic recurrent state space model from irregularly sampled data. In *Proceedings of the AAAI conference on artificial intelligence*: *vol. 37*, (pp. 11060–11068). http://dx.doi.org/10.1609/aaai.v37i9.26310.