actuators

MDPI

Article

White-, Gray- and Black-Box Nonlinear Identification of a Four-Tank MIMO Plant: An Experimental Comparative Study

Alex Espinal *, David Moreno-Salinas and Dictino Chaos García

Department of Computer Science and Automatic Control, Universidad Nacional de Educación a Distancia (UNED), 28040 Madrid, Spain; dmoreno@dia.uned.es (D.M.-S.); dchaos@dia.uned.es (D.C.G.)
* Correspondence: aespinal@dia.uned.es

Abstract

Identifying nonlinear systems is considered a major challenge. This is because this task requires two main steps: selecting a suitable nonlinear model structure and estimating the model parameters to accurately capture its dynamics. In this article, three nonlinear identification techniques are used for a physical four-tank multi-input multi-output (MIMO) plant. Only experimental data is used in this work for the identification of the plant: dedicated tank-emptying and pump-flow tests, alongside a long experiment in which the plant, in a non-minimum phase configuration, is excited with Pseudo-Random Sequence (PRS) signals. The nonlinear identification techniques used are a white-box model developed entirely from prior knowledge of the plant's dynamics; a gray-box model that combines the white-box model with additional terms obtained from experimental data; and a black-box model represented by a system of differential equations obtained by applying symbolic regression (SR) based on genetic programming (GP). The models were assessed on experimental estimation and validation datasets using the root mean square error (RMSE), the fit percentage (FIT) and the coefficient of determination (R^2). The results show that, on the validation dataset, the gray-box and black-box models achieve similar accuracy (FIT between 80% and 91%), clearly higher than that of the white-box model (FIT between 28% and 83%).

Keywords: four-tank plant; nonlinear identification; white/gray/black-box models; symbolic regression; genetic programming

1. Introduction

Mathematical modeling of industrial processes is a complex problem that engineers face daily [1]. Much of the literature on nonlinear identification describes numerous methods and techniques for identifying mathematical models that represent the dynamics of problems in various areas of science and engineering [2–5]. However, in practice, system identification techniques are generally classified into three paradigms [6] based on the degree of prior physical knowledge incorporated into the mathematical structure:

- White-Box Modeling. This involves developing a model based on scientific relationships that physically represent the process, in which all parameters are known [7].
- Gray-Box Modeling. This approach, also known as hybrid modeling, uses prior knowledge about the process and estimates the unknown parts from experimentally measured data [8].
- Black-Box Modeling. This approach employs nonlinear parametric or non-parametric models, such as artificial neural networks, Nonlinear AutoRegressive eXogenous

Check for updates

Academic Editor: Francisco Beltran-Carbajal

Received: 14 October 2026
Revised: 20 November 2026
Accepted: 25 November 2026
Published: 27 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

(NARX) models, or symbolic expressions, to name but a few, that fit the experimentally obtained process data without explicit reliance on physical laws [9].

According to the existing literature, it is possible to find quite a few works in which modeling and identification are applied to the four-tank plant and other similar plants. Starting from the nonlinear theoretical model derived from physical laws, some authors have identified and linearized the plant around an operating point, as in [10,11]. In [12], the nonlinear theoretical model of a plant with six coupled tanks is linearised around an operating point in non-minimum phase configuration, and subsequently, experiments are carried out with the real plant to validate the model. The obtained results clearly demonstrate that the calculated linear model adequately characterizes the dynamics of the real process and can be used in simulations. In [13], the model of a laboratory-scale coupled tank system is identified, based on physical laws, and the identified model has been validated through real-time experimentation. Likewise, in [14], a comparison is made of two parametric identification methods: the transfer function and the state space. They are used for the identification of a real plant with four coupled tanks, where the system is excited with Pseudo-Random Binary Sequence (PRBS) inputs for the collection of validation data, concluding that the estimated models fit with good precision to the real dynamics of the plant. Genetic algorithms (GAs) have also been used for estimating the parameters of the nonlinear model of the two-tank plant in series [15].

The four-tank plant is also a widely used benchmark for multivariable control; for example, decentralised PI control [16,17], LQR and predictive control [18] or nonlinear MPC [19]. Most of these designs rely on simplified theoretical or linearised models; hence, accurate nonlinear models identified from experimental data, as pursued in this work, are a prerequisite for the model-based nonlinear control of the real plant.

Recent works propose the identification and modeling of the four-tank plant using gray-box modeling approaches. In [20], the identification and modeling of a four-tank plant using gray-box models with minimum and non-minimum phase dynamics is proposed. The input is excited using PRBS signals, and a nonlinear Kalman Filter based on the physical model of the system is used to filter noise from the output data, and the identified model is validated with real-time experimental data. A practical procedure to obtain a nonlinear, configurable gray-box model that can be implemented in a real plant is described in [21], and ref. [22] proposes the nonlinear gray-box identification of a plant with two tanks in series, followed by an adaptive controller that stabilizes the water level against disturbances. Important works that apply gray-box modeling are found in [23], where a detailed description of a method for obtaining a semi-physical model based on the knowledge and basic principles of the processes is made, and in [24], where a general methodology for building gray-box or semi-physical models is presented by combining physical knowledge of the system and black-box modeling.

There are other relevant contributions in the literature that address the identification of the four-tank plant using black-box modeling. For example, some authors apply the parametric Nonlinear AutoRegressive eXogenous (NARX) model to represent the black-box model in open-loop and closed-loop systems, exciting the plant with PRS input signals and using the Forward Regression Orthogonal Least Squares (FROLS) and Simulation Error Minimization with Pruning (SEMP) methods for the selection of the model structure and the estimation of the parameters [25]. In addition, ref. [26] proposes a nonlinear identification method by exciting the plant with PRS signals for the control of two tanks using Nonlinear Autoregressive Moving Average with eXogenous inputs (NARMAX) models. Likewise, in [27], the Output Error State Space (MOESP) algorithm is used to estimate the state-space equations, using input and output data to represent the dynamics of the four-tank plant. Furthermore, ref. [28] proposes the identification and modeling of the four-tank plant

using the black-box modeling approach, exciting the white-box model with different input signals, such as variable step, PRBS and Amplitude-modulated Pseudo-Random Binary Signal (APRBS).

The application of machine learning (ML) in the identification of nonlinear systems has grown rapidly in recent years, with applications in various fields. For example, GA and SR have been widely used for the identification of black-box models applied to ships [29,30]; in gray-box identification and modeling, and a predictive nonlinear controller applied to polymerization processes [31]. They have also been used for the identification of dynamic systems [32], in controller design [33], and in extreme learning models based on SR for the identification of systems [34].

Following this trend, the aim of this work is to develop accurate and high-performing mathematical models using experimentally measured real data from a four-tank plant, employing various nonlinear identification techniques. In the literature, a simple theoretical model is frequently employed, assuming that the actuators, such as pumps and three-way valves, are linear, which is incorrect for real-world implementations of this process. Furthermore, in most cases, identification is performed at a specific operating point, or the models are trained using simulated data. Unlike these works, in this work—particularly in the gray-box model—the tank-emptying dynamics and pump flow rates are identified separately, and the three-way valves are set to a predetermined value. Additionally, plant identification is performed considering several possible operating points and using experimentally measured data.

Therefore, this work presents three nonlinear identification model structures: a white-box model, a gray-box model, and a black-box model. The white-box model is derived from mass balance and Bernoulli's laws. The gray-box model is constructed using the white-box model with additional terms estimated from experimentally measured data to achieve a better fit. The black-box model is a nonlinear system obtained using SR. With this technique, the model does not have a fixed or known structure; that is, the form of the equations is not fixed beforehand. Therefore, the model structure that best fits the training data is defined through the identification process itself. The three resulting models have been compared on a set of experimentally measured data from different operating points, obtained by exciting the pumps of the actual plant with Pseudo-Random Sequence (PRS) signals [35].

The main contributions of this work are summarized as follows:

1. Experimental Comparative Framework: White-, gray- and black-box nonlinear models of a physical four-tank MIMO plant in non-minimum phase configuration are compared on the same experimental PRS dataset covering multiple operating points, including the real imperfections of the plant (pump cavitation and offset pressure transmitters).
2. Gray-Box Formulation with Separately Characterized Nonlinearities: Tank discharge (discharge coefficients, with the measured offset between the pressure tap and the outlet orifice) and pump flow rates (third-order polynomials) are identified with dedicated experiments, which reduces the mean validation RMSE with respect to the white-box model.
3. Continuous-Time Symbolic Black-Box Model: Symbolic regression yields a compact ODE model with 16 coefficients, obtained from normal-operation data only, with an accuracy similar to that of the gray-box model, and whose variability with respect to the random initialization of GP is quantified.
4. Practical Guidelines: The three approaches are compared in terms of accuracy, number of parameters, experimental cost, interpretability, physical consistency and validity range, providing guidance for selecting the modeling approach for similar processes.

The rest of the manuscript is organized as follows. In Section 2, a comprehensive description of the four-tank MIMO plant used to perform the experiments is provided,

and the excitation signal and the dataset used for estimating and validating are described. Section 3 describes the process of identifying the four-tank plant, detailing the construction of the white-box, gray-box, and black-box models. Section 4 presents and compares the experimental results with estimation and validation metrics. Finally, the conclusions and future work are discussed in Section 5.

2. Experimental Framework

As mentioned in Section 1, this work addresses the identification of a multi-input multi-output (MIMO) nonlinear system, such as the four-tank plant, which is commonly used in system identification studies. The four-tank plant is a nonlinear process where various MIMO system phenomena can be illustrated, such as the degree of interaction and minimum and non-minimum phase behavior, which is useful for teaching and research [36]. The plant was proposed by Karl Henrik Johansson [17] of Lund Institute of Technology in Sweden and is widely used in university laboratories. Some of the universities that have similar prototypes in their laboratories have their research published in a number of studies [37–40].

2.1. Plant Description

As can be seen in the Piping and Instrumentation Diagram (P&ID) in Figure 1, the system consists of four interconnected tanks with a square prism shape (Tank 1, Tank 2, Tank 3, and Tank 4) and a reservoir with a rectangular prism shape. The instrumentation consists of four pressure transmitters (PT 1, PT 2, PT 3, and PT 4) and four flow transmitters (FT 1, FT 2, FT 3, and FT 4). The actuators consist of two variable flow pumps (Pump 1 and Pump 2), two three-way valves (Valve 1 and Valve 2), four solenoid valves (v_1, v_2, v_3, and v_4), and one manual valve (v_0).

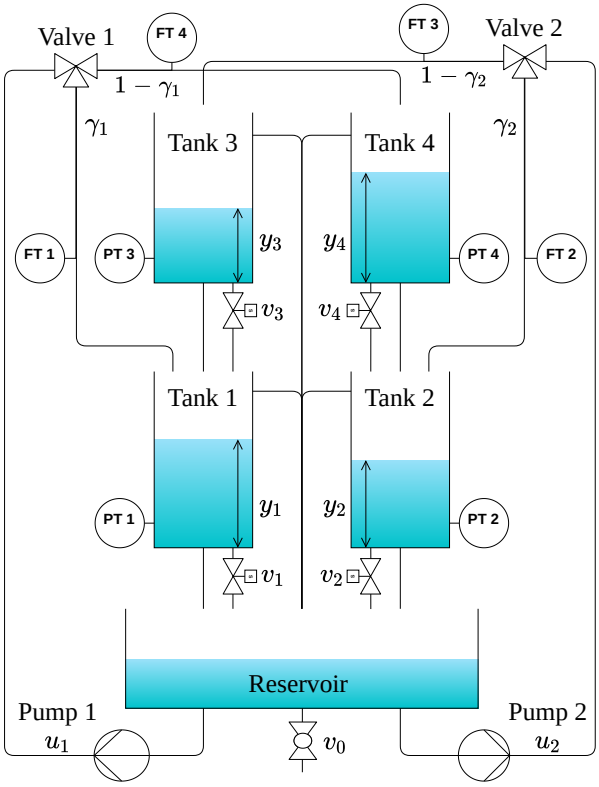


Figure 1. P&ID of the four-tank plant.

All instrumentation and actuators are connected to a Programmable Logic Controller (PLC), which executes the programming code to start the plant. A Supervisory Control and Data Acquisition (SCADA) system has been developed for plant monitoring and control. This system runs on a standard personal computer (PC) and is also used to collect experimentally measured data [41]. All the aforementioned components are mounted on a stainless steel frame.

The tanks are located on three levels: the reservoir at the lowest level, which serves as both the source and sink for the liquid circulating in the system; tanks 1 and 2 at the intermediate level (T1 and T2), which drain into the reservoir; and tanks 3 and 4 at the highest level (T3 and T4), which drain into the intermediate tanks. Each tank has two orifices of a specific cross-section at the bottom: one orifice through which the liquid flows freely, and the other that must be opened by the solenoid valve to create disturbances.

The liquid circulation in the system occurs as follows: pump 1 pumps liquid from the reservoir to three-way valve 1, which then distributes it to tanks 1 and 4. Meanwhile, the flow from pump 2 is also distributed by the three-way valve 2, supplying tanks 2 and 3. Since tanks 4 and 2 are connected, the flow from pump 1 indirectly affects the liquid level in tank 2. Similarly, because tanks 3 and 1 are connected, the flow from pump 2 indirectly affects the liquid level in tank 1. Therefore, the four-tank plant is a highly coupled system that is complex to identify.

The system has two inputs, which correspond to the voltages u_1 and u_2 applied to the pumps, and the outputs are the levels y_1, y_2, y_3, and y_4 measured by the pressure transmitters in each tank.

2.2. Plant Setup and Dataset Generation

The experimental data were obtained by exciting with PRS signals the actual four-tank MIMO plant depicted in Figure 2, located in the laboratory of the Department of Computer Science and Automatic Control at UNED. This approach makes it possible to capture the plant dynamics at various operating points without excessively disturbing its normal operation, while also enabling the acquisition, within a reasonable time frame, of the input–output data required to identify suitable plant models, all while keeping the amplitudes of inputs and outputs within user-defined value ranges [42].

The plant is home-built; that is, made by hand, with some imperfections, such as old and heavily used actuators, and not optimally located pressure transmitters. Part of this leads to pumps cavitating, which makes levels slightly unstable and causes some limitations in the measurement of pressure since the transmitters are not located at the bottom of the tanks. Consequently, this creates the need to perform a large number of experiments and makes the process of identifying accurate models more complex and tedious.

Before conducting the experiment, the plant was configured by setting the opening parameters (γ_1, γ_2) of the three-way valves to ensure it is coupled and has a non-minimum phase dynamic. Non-minimum phase dynamics are required due to plant physical limitations, as the location of the pressure transmitters in the tanks does not allow level values to be recorded correctly in another configuration.

The pumping unit uses Grundfos UPE 25-40 180 (Series 2000) electronic circulator pumps with a factory-built intelligent differential-pressure controller driven by a $[0, 10]$ V analogue input [43]. Using the pumps' experimentally measured flow rate at the plant under maximum control voltage, the gain parameters (k_1, k_2) were identified, since the system's overall behaviour is highly nonlinear and exhibits cross-coupling, making the nominal manufacturer data inadequate for an accurate white-box model. These constants provide a direct mapping between an increase in the control signal and the resulting change in the fluid volumetric flow rate under steady-state loading.

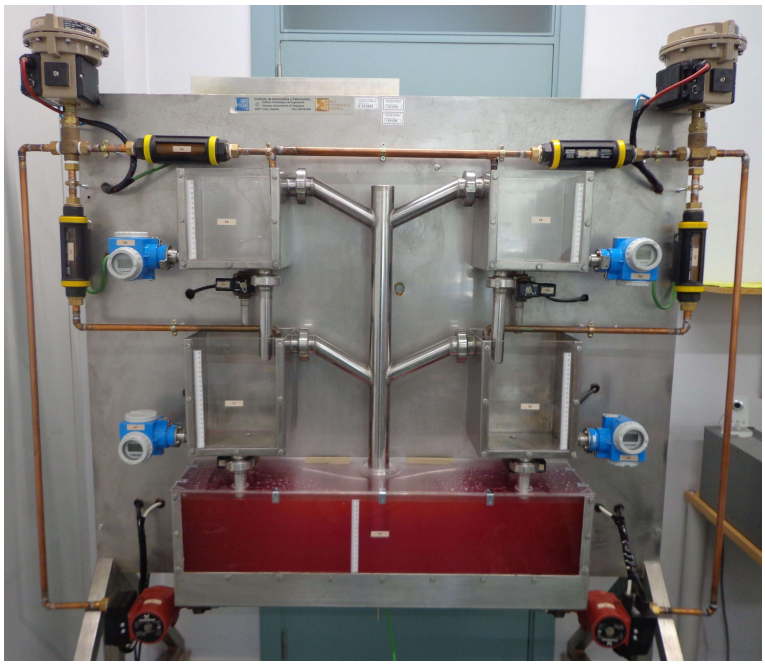


Figure 2. Real four-tank plant.

It should be noted that this linear, single-point calibration of the pump gains (k_1, k_2) at maximum control voltage is an inherent assumption of the white-box model: the pump flow rate is assumed to be proportional to the control voltage over the whole operating range, which, as shown in Section 3.2.2, does not hold for the actual pumps. Table 1 shows the actual plant parameter values.

Table 1. Plant parameter values.

Parameter	Description	Value	Unit
A_1, A_2	Cross area (T1, T2)	390.83	cm^2
A_3, A_4	Cross area (T3, T4)	392.92	cm^2
a_1, a_2	Orifice area (T1, T2)	1.5394	cm^2
a_3, a_4	Orifice area (T3, T4)	1.3273	cm^2
g	Gravity acceleration	981	cm/s^2
k_1, k_2	Pump 1, 2 gain	36, 30	$cm^3/(V \cdot s)$
γ_1, γ_2	Valve 1, 2 gain	0.35, 0.34	[–]

The chosen PRS signal has been designed to excite the pumps with a sampling period of 1 s and a voltage range of $[3, 7]$ V, which guarantees stable plant behavior. The voltage applied to each pump is a multilevel random sequence with amplitudes drawn from a uniform distribution in this range and a hold time of 60 min. The two sequences are shifted by 30 min, so that only one pump changes at a time: each upper tank receives a step every 60 min and each intermediate tank, which is affected by both pumps, every 30 min. At 3 V, the level of the upper tanks is close to the pressure tap, and below this voltage it does not reach the transmitters. The time constants of the tanks are approximately 1–2 min, so every transient is completed well before the next step. The long hold time was chosen to average out the slow level fluctuations caused by pump cavitation and to characterise the steady state at each operating point. Consequently, the input power is concentrated at frequencies below the corner frequencies of the tanks, and the transient information is provided by the step responses (15 in each dataset). The generated dataset has 57,600 input and output samples and is divided into two equal parts: the first 28,800 samples are used for estimation and the remaining 28,800 samples for model validation, as shown in Figure 3.

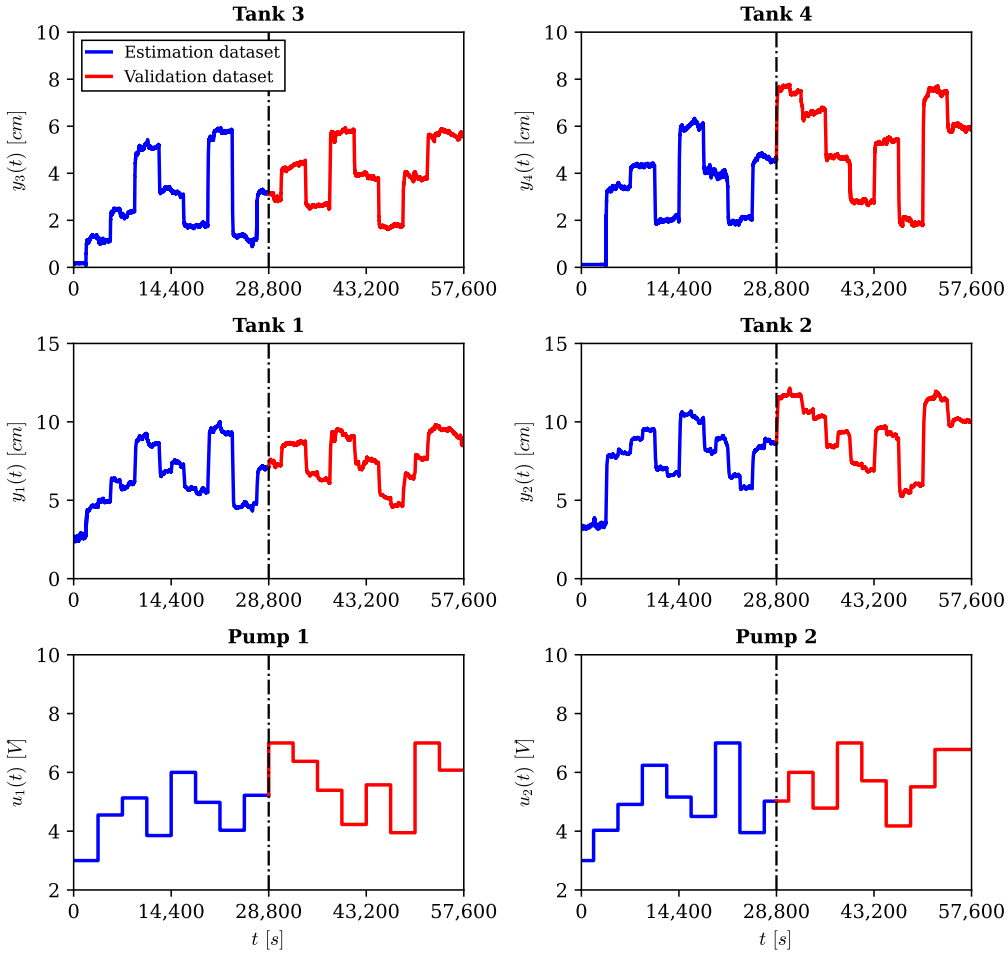


Figure 3. Estimation (blue) and validation (red) datasets. Tank levels are $y_1(t)$ to $y_4(t)$, respectively, and $u_1(t), u_2(t)$ are the voltages applied to the pumps.

Since the amplitudes are random, the two parts do not cover exactly the same voltage levels: u_1 reaches 7 V only in the validation part, so the validation dataset includes operating conditions not visited during estimation.

2.3. Performance Metrics and Simulation Protocol

To evaluate the performance of each model across the estimation (training) and the validation datasets, three standard goodness-of-fit metrics are calculated for each tank level y_i ($i = 1, 2, 3, 4$):

- Root Mean Square Error (RMSE): This represents the standard deviation of the residuals, or prediction errors, and describes how these residuals are distributed.

$$\text{RMSE}_i = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(\hat{y}_i(k) - y_i(k))^2} \tag{1}$$

- Fit Percentage (FIT): This refers to the percentage of output variation captured by the model.

$$\text{FIT}_i = \left(1 - \frac{\sqrt{\sum_{k=1}^{N}(\hat{y}_i(k) - y_i(k))^2}}{\sqrt{\sum_{k=1}^{N}(\bar{y}_i - y_i(k))^2}}\right) \times 100 \tag{2}$$

- Coefficient of Determination $\left(R^2\right)$: The proportion of total variance explained by the model predictions.

$$R^2{}_i = 1 - \frac{\sum_{k=1}^{N}(\hat{y}_i(k) - y_i(k))^2}{\sum_{k=1}^{N}(\bar{y}_i - y_i(k))^2} \tag{3}$$

where N is the number of samples, y_i is the measured output, $\hat{y}_i$ is the simulated output, and $\bar{y}_i$ is the mean of the measured output.

All models are assessed in free-run (open-loop) simulation. Specifically, the ODE systems (4), (12) and (16) are integrated by using only the measured pump voltages $u_1(t)$ and $u_2(t)$ as inputs (kept piecewise constant between sampling instants), and the measured tank levels are never fed back into the models. The integration is performed with the ode45 solver of MATLAB R2026b, with relative and absolute tolerances of 10^{-6} and 10^{-8}, and the resulting simulated levels are sampled every 1 s to evaluate (1)–(3). For each dataset, the simulation starts from the measured levels at the first available sample ($k = 1$ for estimation and $k = 28{,}801$ for validation). Since the transmitters read zero when the level is below their pressure tap, negative simulated levels are set to zero before computing the metrics. The only exception is the selection stage of the black-box model (Section 3.3.3), in which each equation is evaluated separately with the measured levels of the upper tanks as inputs. Finally, note that even though the black-box model is identified by fitting level derivatives in Section 3.3, it is still evaluated in free-run simulation.

3. Nonlinear Identification Methods

In this section, the nonlinear identification procedure of the four-tank plant is carried out, applying white-box, gray-box, and black-box modeling.

3.1. White-Box Modeling

The nonlinear set of ordinary differential equations describing the four-tank system is derived from mass conservation and Bernoulli's principle [44]. By expressing the potential and kinetic energy terms, one obtains the fluid velocity as $v = \sqrt{2gy}$ and the volumetric flow rate q, leaving a tank with an outlet area a as $q = av = a\sqrt{2gy}$, where y is the level of the tank.

The corresponding white-box model of the four interconnected tanks is derived from first principles under idealized assumptions: the liquid is treated as an inviscid (perfect) fluid, the actuators are assumed to behave linearly, and the constrictions of water flow through the valves, any flow losses due to pipe friction or discharge through the tank orifices, and the vortex generation at the water outlet of the orifice are neglected.

The resulting model is given in (4).

$$\begin{aligned}
\frac{dy_1(t)}{dt} &= -\frac{a_1}{A_1}\sqrt{2gy_1(t)} + \frac{a_3}{A_1}\sqrt{2gy_3(t)} + \frac{\gamma_1 k_1}{A_1}u_1(t) \\
\frac{dy_2(t)}{dt} &= -\frac{a_2}{A_2}\sqrt{2gy_2(t)} + \frac{a_4}{A_2}\sqrt{2gy_4(t)} + \frac{\gamma_2 k_2}{A_2}u_2(t) \\
\frac{dy_3(t)}{dt} &= -\frac{a_3}{A_3}\sqrt{2gy_3(t)} + \frac{(1-\gamma_2)k_2}{A_3}u_2(t) \\
\frac{dy_4(t)}{dt} &= -\frac{a_4}{A_4}\sqrt{2gy_4(t)} + \frac{(1-\gamma_1)k_1}{A_4}u_1(t)
\end{aligned} \tag{4}$$

The flow rate k_1u_1 from pump 1 is split into flow rates $\gamma_1 k_1 u_1$ to tank 1 and $(1-\gamma_1)k_1u_1$ to tank 4. Similarly, the flow rate k_2u_2 from pump 2 can be split into flow rates $\gamma_2 k_2 u_2$ to tank 2 and $(1-\gamma_2)k_2u_2$ to tank 3. The parameters γ_1 and γ_2 are used to adjust the valve openings before an experiment.

The plant exhibits non-minimum phase behavior when the gain values of valves γ_1 and γ_2 are at $0 < \gamma_1 + \gamma_2 < 1$, and minimum phase behavior when the values of γ_1 and γ_2 are at $1 < \gamma_1 + \gamma_2 < 2$.

3.2. Gray-Box Modeling

The procedure for constructing the gray-box model of the plant involves combining the white-box model with additional terms adjusted based on experimentally measured data. Therefore, the gray-box model will consist of known and unknown parameters. The unknown parameters will need to be estimated to fit the measured data. Thus, the model configuration is considered in two parts:

- Tank-Emptying Dynamics. Tank discharge describes how a tank empties, focusing on the non-constant flow rate that decreases as the fluid level drops.
- Pump Flow Dynamics. This refers to the flow rate entering the plant, that is, the flow rate that the pumps deliver from the reservoir to the tanks.

3.2.1. Tank Emptying Dynamics

When a tank is drained through an orifice that is free of leaks, and the fluid is water, the theoretical outflow velocity v_T can be expressed using principles like Torricelli's law, which is derived from Bernoulli's Equation [45]:

$$v_T = \sqrt{2gy} \tag{5}$$

However, in practice, the actual velocity v_R of the fluid through the orifice is less than the theoretical velocity v_T due to friction losses. This difference is corrected by the velocity coefficient C_v [46], in the form

$$v_R = C_v\sqrt{2gy} \tag{6}$$

Furthermore, the area of the contracted flow is smaller than the area of the orifice. The ratio between the area of the contracted flow a_o and the area of the orifice a is called the contraction coefficient C_c:

$$C_c = \frac{a_o}{a} \tag{7}$$

Therefore, since the flow rate can be represented by $Q = av_R$, the flow rate exiting through the orifice is given by

$$Q = aC_vC_c\sqrt{2gy} \tag{8}$$

The discharge coefficient C_d, a dimensionless factor correcting for energy losses and flow inefficiencies in real-world applications [47], results from the combination of the velocity and contraction coefficients. For sharp-edged orifices, the discharge coefficient is about 0.61, and it increases with the rounding of the inlet edge, up to about 0.80 when the rounding radius reaches 10% of the orifice diameter [48,49]. Additionally, the measured level y is not the hydraulic head above the outlet orifice: the actual head is $y + \beta$, where β is the vertical distance between the pressure tap of the level transmitter and the outlet orifice, which is located at the end of a short pipe below the tank bottom. This offset is a geometric constant of the plant, $\beta = 4$ cm for the four tanks, and it is therefore measured and not estimated. A measured level $y = 0$ thus does not correspond to an empty tank but to a water level at the height of the transmitter, below which water still flows through the orifice. Therefore, the non-zero discharge predicted at $y = 0$ is physically correct. The outflow of each tank is given by

$$Q = aC_d\sqrt{2g(y + \beta)} \tag{9}$$

Finally, the physical model of tank emptying is obtained by combining (9) with the mass balance of the tank, as shown in (10). Since the outflow must vanish when the water level reaches the orifice ($y = -\beta$), the max function is used to keep the model physically consistent for levels below the transmitter.

$$A\frac{\mathrm{d}y}{\mathrm{d}t} = -aC_d\sqrt{2g\max(y+\beta, 0)} \tag{10}$$

In a series of experiments to estimate the discharge coefficients C_d, the tanks were emptied from their maximum levels. Measurements were taken within $[0, 18]$ cm for the intermediate tanks (T1, T2) and $[0, 14]$ cm for the upper tanks (T3, T4). With β fixed at its measured value, C_d is the only unknown parameter of (10), and it was estimated for each tank from the recorded level data via least squares, as illustrated in Figure 4.

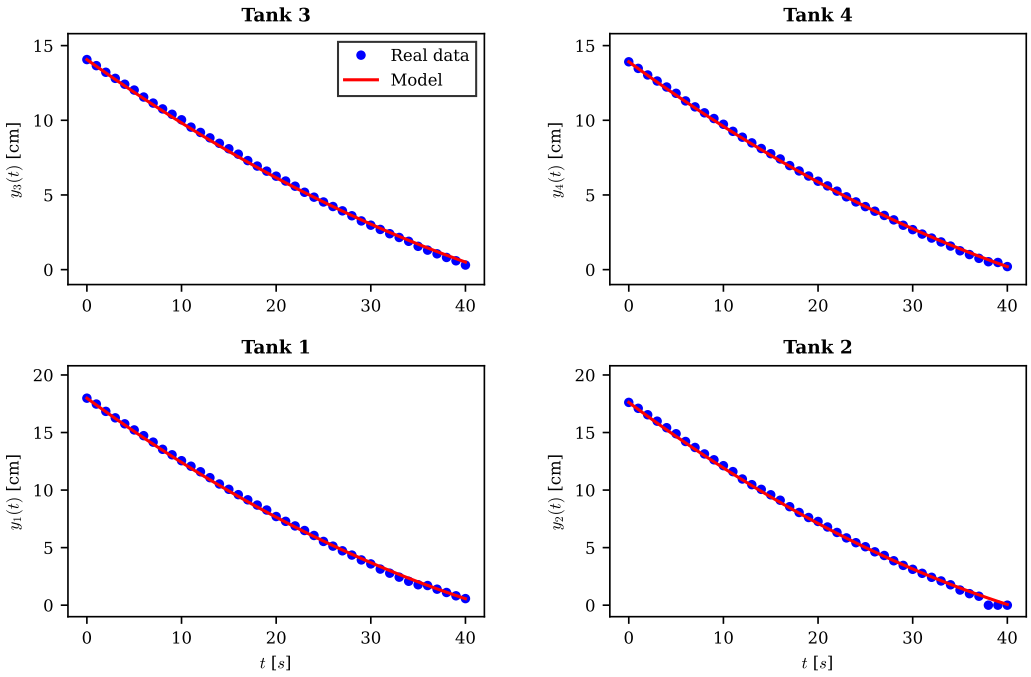


Figure 4. Real emptying curves (blue) versus their estimates (red). Tank discharge levels are $y_1(t)$ to $y_4(t)$, respectively.

Table 2 presents the values of the discharge coefficients C_{d_i} estimated from the experimentally measured data and the coefficients of determination $R^2{}_i$ of each fit, with $i = 1, 2, 3, 4$ representing the plant tanks.

Table 2. Tank-emptying model parameter values and coefficients of determination R^2 of each fit.

Parameter	T1	T2	T3	T4	Unit
C_{d_i}	0.7317	0.7552	0.7110	0.7296	[−]
$R^2{}_i$	0.9993	0.9991	0.9992	0.9996	[−]

The estimated values lie between those of a sharp-edged and a well-rounded orifice, which is consistent with the slightly rounded inlet edges of the plant outlets.

3.2.2. Pump Flow Dynamics

Based on the flow rates delivered by the pumps from the reservoir and the configuration of the three-way valves, it is possible to determine the inlet flow dynamics for each tank and, consequently, obtain the associated characteristic curve. To individually estimate the

inlet flow characteristic curves for each tank, the three-way valves were set according to the configuration shown in Table 1. Subsequently, a stepped voltage signal was applied to pump 1 in the range of $[2, 7]$ V, with 1 V increments, maintaining each level until the inlet flow rate to tanks 1 and 4 reached steady-state. The experiment was repeated in the same manner with pump 2, which feeds tanks 2 and 3.

Once the corresponding flow data were acquired, the steady-state mean flow rate was calculated for each applied voltage step, accounting for the independent operation of the pumps. Finally, the parameters of a third-order polynomial (11) were estimated using the least squares method to represent the mean flow rate.

$$F_i(u_j) = p_{1i}u_j^3 + p_{2i}u_j^2 + p_{3i}u_j \tag{11}$$

where F_i is the flow rate entering tank i, u_j is the voltage applied to pump j, which is the pump that feeds tank i, see (13), and (p_{1i}, p_{2i}, p_{3i}) are the parameters to be estimated. Figure 5 shows the flow rate curves of the tanks for different voltage levels and their estimates.

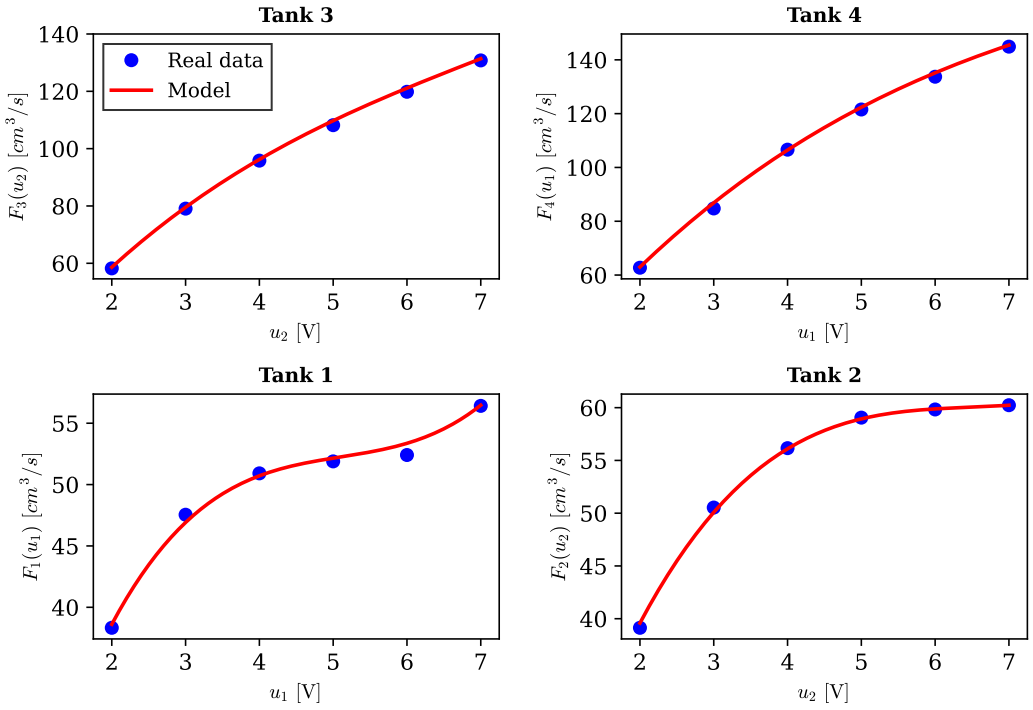


Figure 5. Real flow rate curves (blue) versus their estimates (red). Tanks flow rates are F_1 to F_4, respectively.

Table 3 shows the parameter values of the models that describe the flow rate curves, which were estimated from the experimentally measured data together with the coefficient of determination $R^2{}_i$ of each fit.

Table 3. Parameter values of the flow rate models and the coefficients of determination R^2 of each fit.

Parameter	T1	T2	T3	T4	Unit
p_{1i}	0.3553	0.2154	0.1723	0.0988	$cm^3/(V^3 \cdot s)$
p_{2i}	−5.4457	−4.1761	−3.6591	−3.0318	$cm^3/(V^2 \cdot s)$
p_{3i}	28.7770	27.2837	35.9387	37.1617	$cm^3/(V \cdot s)$
$R^2{}_i$	0.9922	0.9987	0.9986	0.9984	[−]

3.2.3. Final Gray-Box Model

After determining the emptying dynamics and identifying the equation that represents the characteristic curve of the flow rates in the tanks from the experimentally measured data, the initial structure (white-box model) is configured with new parameters, resulting in the gray-box model shown in (12).

$$\begin{aligned}
\frac{\mathrm{d}y_1(t)}{\mathrm{d}t} &= -\frac{a_1}{A_1}C_{d_1}\sqrt{2g\cdot\max(y_1(t)+\beta,0)} + \frac{a_3}{A_1}C_{d_3}\sqrt{2g\cdot\max(y_3(t)+\beta,0)} \\
&\quad + \frac{1}{A_1}F_1(u_1) \\
\frac{\mathrm{d}y_2(t)}{\mathrm{d}t} &= -\frac{a_2}{A_2}C_{d_2}\sqrt{2g\cdot\max(y_2(t)+\beta,0)} + \frac{a_4}{A_2}C_{d_4}\sqrt{2g\cdot\max(y_4(t)+\beta,0)} \\
&\quad + \frac{1}{A_2}F_2(u_2) \\
\frac{\mathrm{d}y_3(t)}{\mathrm{d}t} &= -\frac{a_3}{A_3}C_{d_3}\sqrt{2g\cdot\max(y_3(t)+\beta,0)} + \frac{1}{A_3}F_3(u_2) \\
\frac{\mathrm{d}y_4(t)}{\mathrm{d}t} &= -\frac{a_4}{A_4}C_{d_4}\sqrt{2g\cdot\max(y_4(t)+\beta,0)} + \frac{1}{A_4}F_4(u_1)
\end{aligned} \tag{12}$$

As explained in Section 3.2.1, negative values of y_i in the simulation correspond to levels between the outlet orifice and the pressure tap, which the transmitters read as zero. The flow rates of the pumps are given by

$$\begin{aligned}
F_1(u_1) &= p_{11}u_1^3 + p_{21}u_1^2 + p_{31}u_1 \\
F_2(u_2) &= p_{12}u_2^3 + p_{22}u_2^2 + p_{32}u_2 \\
F_3(u_2) &= p_{13}u_2^3 + p_{23}u_2^2 + p_{33}u_2 \\
F_4(u_1) &= p_{14}u_1^3 + p_{24}u_1^2 + p_{34}u_1
\end{aligned} \tag{13}$$

3.3. Black-Box Modeling

To obtain a simple black-box model that fits the training data accurately, SR based on GP is used without needing to specify the model structure beforehand, as the appropriate structure is identified simultaneously with the model parameter values. The black-box models are obtained in two stages. In the first stage, the structure of each equation in (14) is determined: the multigene expression whose genes are combined linearly to predict the level derivative $\dot{y}_i$. In this stage, the gene weights are computed by ordinary least squares on the numerically differentiated levels, that is, by minimising an equation error. In the second stage, the parameters of the selected structure are refined in simulation (Section 3.3.4).

3.3.1. Problem Formulation

GP is a specific type of evolutionary computation in which computer programs are developed to solve a given problem [50]. The SR via GP seeks a mathematical expression that describes the relationship between multiple input variables and a single output variable, given a dataset of observed values for all variables [34]. That is, SR seeks a function f with $y = f(x_i)$, where x_i are the input variables and y is the output variable. The function f is represented as a mathematical expression in terms of elementary functions and operators.

Therefore, the goal is to use SR to find the expressions f_i of a system of ordinary differential equations (ODEs) in the following form:

$$\begin{aligned} \dot{y}_1(t) &= f_1(y_1(t), y_3(t), u_1(t)) \\ \dot{y}_2(t) &= f_2(y_2(t), y_4(t), u_2(t)) \\ \dot{y}_3(t) &= f_3(y_3(t), u_2(t)) \\ \dot{y}_4(t) &= f_4(y_4(t), u_1(t)) \end{aligned} \tag{14}$$

where the input data $(y_1(t), y_2(t), y_3(t), y_4(t))$ are the tank levels and the voltages applied to the pumps $(u_1(t), u_2(t))$. The output data $(\dot{y}_1(t), \dot{y}_2(t), \dot{y}_3(t), \dot{y}_4(t))$ are the time derivatives of the tank levels, whose values are obtained using centered derivatives, where points to the left and right are used, resulting in a closer approximation to the true value than forward or backward methods by averaging their results [51]. The formula for calculating the centered derivative of the level data is shown below:

$$\dot{y}_i(t) = \frac{y_i(t+h) - y_i(t-h)}{2h} \tag{15}$$

where $y_i(t)$ are the values of the tank levels measured experimentally at time t, with $i = 1, 2, 3, 4$ representing the plant tanks, and $h = 1$ s is the sampling time.

The derivatives were computed from the raw level measurements, without prior filtering. The measurement noise, estimated in steady-state windows, has a standard deviation of approximately 0.009–0.012 cm depending on the tank, which yields a noise standard deviation of about 0.006–0.009 cm s^{-1} in (15), of the same order as the derivative itself. Its effect on the SR fit is limited because it enters the regressand, where it is averaged out by the least squares estimation of the gene weights, and because the Pareto-based selection penalises complex expressions that could fit the noise [4]. Moreover, the models are selected and evaluated in free-run simulation (Section 2.3), where derivative noise plays no role.

3.3.2. GPTIPS Configuration

To implement SR and derive the corresponding black-box model, GPTIPS 2.0 (Genetic Programming Toolbox for the Identification of Physical Systems) for SR in MATLAB, developed by [52], is used. GPTIPS implements multigene symbolic regression (MSR), which involves one or more genes, each of which is a genetic programming tree. Genes are incrementally acquired by individuals to enhance fitness. The overall model is a weighted linear combination of each gene.

Before initiating the model search, it is necessary to define a series of parameters in GPTIPS. These parameters were defined after extensive testing to achieve a set of simple and accurate models. Table 4 presents the parameters and their values so that the genetic programming algorithm can begin the model search for each tank level.

The population size is 100 models over 500 generations in 2 runs, whose final populations are merged. No time limit is imposed, so every run stops at the maximum number of generations and the result depends only on the random seed.

A Pareto tournament size of 20 models is chosen for the competition, with a probability of 0.3 to encourage less complex models and a recombination fraction of 0.3 so that approximately 1/3 of the models are copied in the next generation.

Table 4. GPTIPS configuration for the intermediate tanks (T1, T2) and the upper tanks (T3, T4).

Parameter	T1, T2	T3, T4
Run control		
Population size	100	100
Number of generations	500	500
Number of runs	2	2
Timeout	None	None
Selection		
Pareto tournament size	20	20
Pareto probability	0.3	0.3
Recombination fraction	0.3	0.3
Genes and nodes		
Maximum number of genes	3	2
Number of inputs	3	2
Maximum depth of trees	3	3
ERC probability	0.3	0.3
Function nodes	plus, minus, sqrt, square	

The maximum number of genes for the intermediate tanks (T1, T2) is 3, and 2 for the upper tanks (T3, T4), resulting in 3 and 2 genetic programming trees, respectively, with a maximum depth of 3. There are 3 input variables for T1 and T2 and 2 for T3 and T4, and a bias term is included in every model. The function nodes are addition, subtraction, square root and square. The parameter values are lower for T3 and T4 because they exhibit simpler dynamics than T1 and T2 as they are not disturbed. The ephemeral random constant (ERC) probability is a parameter that sets the probability that a leaf node of a tree is a numerical constant rather than an input variable. More operations are available, but the goal is to obtain a model that accurately represents the real system in the simplest way possible.

The parameter values have been chosen to define models that are not overly complex, but that are capable of accurately fitting the training data, and whose search space contains the main terms of the physical model (square roots of the levels and polynomial terms of the inputs) to make it a more general model. Greater tree depth and a larger number of genes can produce models with many nodes, leading to overfitting of the training data. The biases and coefficients of the models are automatically determined using a least squares procedure for each multigene individual.

3.3.3. Structure Selection

Since genetic programming is a stochastic search, the first stage was repeated in $N_r = 20$ independent runs for each tank, each initialised with a different random seed (1–20) and with the same configuration (Table 4). The complexity of a candidate model is measured with the expressional complexity [53] used by GPTIPS [52], defined as the sum, over all the nodes of each gene tree, of the number of nodes in the subtree rooted at that node, summed over all the genes of the model. Unlike the plain node count, this measure penalises deep, nested expressions more than shallow ones with the same number of nodes, and therefore favours flat, interpretable structures.

During the evolution, complexity is taken into account through Pareto tournament selection, in which individuals compete on both their fitness and their expressional complexity. At the end of each run, the Pareto front of non-dominated models is obtained, that is, the models for which no other model is simultaneously more accurate and less complex. Because the fitness is computed on the level derivative, whereas the models are used to predict the level, every model on the front was simulated in free run on the estimation dataset. In each run, the model selected is the least complex one whose simulation RMSE is

within 5% of the RMSE of the most accurate model of the front. The final model of each tank was then chosen among the 20 selected models.

Since GP does not constrain the structure, a physical plausibility check was applied only at this point. Among the models whose simulation RMSE is within 8% of the best of the 20 runs, a tolerance wider than within each run because the runs explore different structures, the simplest one with a square-root dependence on the pump voltage or on the levels and an outflow term in the level of the tank was adopted (2.3, 6.4, 7.7 and 1.1% above the best run for tanks 1 to 4). In the runs that produced these models, the selected model of tanks 1 and 4 is also the most accurate model of the corresponding front, with an expressional complexity of 20 and 14, respectively. For tanks 2 and 3, the selected model is clearly simpler than the most accurate one (complexity of 12 versus 17 and of 10 versus 23), at the cost of an increase of only 0.8 and 1.0 % in the estimation RMSE. These complexities, like those of Table 5, are computed on the unsimplified gene trees, which is the measure used during the search and the selection. In the models of tanks 1 and 4, the genes $y_1 + (u_1 - y_1)$ and $y_4 + y_4$ reduce to u_1 and $2y_4$, so the complexity of the four simplified expressions in (16) is 10, 12, 10 and 10.

Table 5 summarises the mean and standard deviation of the free-run RMSE and FIT obtained by the model selected in each run, together with the range of the validation FIT and the mean expressional complexity. Each equation is simulated with the measured levels of the other tanks as inputs, the estimation indices are computed over the estimation dataset and the validation indices are computed over the validation dataset. The results are repeatable: for y_1, y_3 and y_4, the standard deviation of the validation FIT is at most 0.3 percentage points and the difference between the best and the worst run does not exceed 1.2 points. Tank 2 shows the largest dispersion (74.0 ± 2.5%, range 67.0–75.6%) and also the highest mean complexity. In all cases, the validation indices are similar to the estimation ones, so there is no evidence of overfitting.

Table 5. Statistics of the symbolic-regression models over $N_r = 20$ independent runs (random seeds 1–20). Values are mean ± standard deviation of the free-run RMSE and FIT of the model selected in each run, min–max is the range of the validation FIT over the runs, and complexity is the mean expressional complexity of the selected models.

	Estimation		Validation			
Tank Level	**RMSE [cm]**	**FIT [%]**	**RMSE [cm]**	**FIT [%]**	**Min–Max [%]**	**Complexity**
y_1	0.446 ± 0.012	76.7 ± 0.6	0.343 ± 0.003	76.6 ± 0.2	75.9–76.9	17.1
y_2	0.515 ± 0.062	74.5 ± 3.0	0.503 ± 0.048	74.0 ± 2.5	67.0–75.6	24.2
y_3	0.288 ± 0.013	83.2 ± 0.7	0.204 ± 0.004	84.9 ± 0.3	84.0–85.2	17.6
y_4	0.187 ± 0.003	89.2 ± 0.2	0.185 ± 0.002	90.3 ± 0.1	90.1–90.5	16.8

3.3.4. Parameter Refinement

Estimating the gene weights by equation error makes the evolutionary search computationally tractable since thousands of candidate structures must be evaluated, but it has two drawbacks. First, the equation error contains not only the amplified measurement noise but also the slow level fluctuations caused by pump cavitation (Section 2.2). During the long intervals of constant pump voltage, these fluctuations add variance to the level regressors without providing information on the dynamics, so the least squares estimates of the gene weights tend to be attenuated and the resulting models have slower dynamics than the plant. Second, minimising the error in the derivative does not guarantee a good free-run simulation of the levels, which is the intended use of the models and the criterion used to compare them with the white-box and gray-box models.

To overcome these drawbacks, in the second stage, the structure selected by symbolic regression was kept fixed, and all its parameters (the gene weights and the numerical

constants inside the genes) were re-estimated by minimising the free-run simulation error of the levels on the estimation dataset. Tanks 3 and 4 were refined first; for tanks 1 and 2, the level of the upper tank was taken from the simulation of the refined model of tank 3 or 4, so that the four equations are refined and evaluated in free run of the complete model, driven only by the pump voltages. This output-error estimation was carried out by nonlinear least squares, using the GP parameters as initial values and a continuous-time formulation integrated with MATLAB's ode45 solver (relative tolerance 10^{-6}, absolute tolerance 10^{-8}). Table 6 compares the final model of each tank before and after this refinement.

Table 6. Two-stage identification of the black-box models: free-run performance of the final model of each tank with the gene weights estimated from the numerical derivatives (SR) and after re-estimating all its parameters by minimising the free-run simulation error of the levels (Refined). Both columns correspond to the free-run simulation of the complete model (16), driven only by the pump voltages.

	Estimation FIT [%]		**Validation FIT [%]**	
Tank Level	**SR**	**Refined**	**SR**	**Refined**
y_1	71.7	83.4	74.8	82.2
y_2	70.5	80.0	71.1	82.6
y_3	82.5	85.2	85.2	88.1
y_4	89.3	90.5	90.4	91.4

Table 6 shows that the refinement improves the FIT of all four tanks, by 1.2–11.7 percentage points on the estimation data and by 1.0–11.5 percentage points on the validation data, while the structure and complexity of the models remain unchanged. The largest improvements are obtained for tanks 1 and 2 (y_1: 74.8 → 82.2%; y_2: 71.1 → 82.6% in validation), whose equations depend on two levels, both affected by the slow fluctuations described above, and which accumulate the errors of the upper tanks in the free-run simulation. This confirms that symbolic regression on numerical derivatives is effective in identifying the model structure, but the final parameters should be estimated in free-run simulation of the complete model.

3.3.5. Final Black-Box Model

The resulting black-box model is illustrated in (16).

$$\begin{aligned}
\dot{y}_1(t) &= d_{1,0} + d_{1,1}\, y_1(t) + d_{1,2}\, u_1(t) + d_{1,3}\left(\sqrt{y_1(t)} + y_3(t)\right) \\
\dot{y}_2(t) &= d_{2,0} + d_{2,1}\left(\sqrt{u_2(t)} - y_2(t)\right) + d_{2,2}\, y_4(t) + d_{2,3}\sqrt{y_4(t)} \\
\dot{y}_3(t) &= d_{3,0} + d_{3,1}\, y_3(t) + d_{3,2}\sqrt{u_2(t) + c_3} \\
\dot{y}_4(t) &= d_{4,0} + d_{4,1}\, 2y_4(t) + d_{4,2}\sqrt{u_1(t) + c_4}
\end{aligned} \tag{16}$$

Model (16) follows the multigene form used by GPTIPS [52], in which the output is the sum of a bias term $d_{i,0}$ and the genes weighted by $d_{i,j}$, where i denotes the tank and j the gene. A gene may contain more than one term, in which case its terms share the same weight: this is the case of $d_{1,3}$, which multiplies both $\sqrt{y_1}$ and y_3, and of $d_{2,1}$, which multiplies both $\sqrt{u_2}$ and y_2. Similarly, the gene $y_4 + y_4$ of tank 4 is written as $2y_4$, so that $d_{4,1}$ is the weight found by GP. The constants c_3 and c_4 inside the square roots are ERCs, numerical leaf nodes that GP inserts in the gene trees with the ERC probability of Table 4, and whose values are generated randomly during the evolution rather than estimated by least squares. Both the weight sharing and the ERCs are a consequence of the structure found by GP and were preserved in the refinement stage, in which the structure is kept fixed and only the values of $d_{i,j}$ and c_i are re-estimated. Table 7 lists the refined values.

Table 7. Parameters of the black-box model (16) after the refinement stage. $d_{i,0}$ is the bias term, $d_{i,j}$ is the weight of gene j, and c_i is the ERC inside the square root.

Parameter	T1	T2	T3	T4	Unit
$d_{i,0}$	0.01784	0.02544	−0.29475	−0.28902	$\mathrm{cm\,s^{-1}}$
$d_{i,1}$	−0.02181	0.02170	−0.01648	−0.00809	Varies [1]
$d_{i,2}$	0.00486	0.01437	0.13433	0.14570	Varies [1]
$d_{i,3}$	0.01843	0.02380	–	–	Varies [1]
c_i	–	–	1.49185	0.95991	V

[1] The unit of a gene weight depends on the terms of its gene. The values are valid for levels in cm, voltages in V and time in s.

The bias terms $d_{i,0}$ in (16) are fitting intercepts and have no direct physical meaning; together with the input terms, they represent the net inflow over the identified operating range. The same applies to the constants c_i. Consequently, model (16) is valid only within the identified input range $u_1, u_2 \in [3, 7]$ V and the corresponding level ranges, and it must not be used to extrapolate to the discharge regime ($u = 0$ V).

4. Results

This section presents the results of the estimation and validation of the fitted white-box, gray-box, and black-box models.

4.1. White-Box Model

Figure 6 shows the plant's response (blue) along with that of the white-box model (red). The figure shows the tank levels 1 to 4 on the left side for the white-box model estimation and on the right side for the white-box model validation, respectively. On the left side (T3, T4), the model attempts to follow the data, but there are clear differences. These differences are due to the fact that the white-box model was constructed ideally, assuming no losses during tank emptying and a linear pump flow rate, and neglecting other plant imperfections, as explained in Sections 2.2 and 3.1. Similarly, the white-box model attempts to track the levels of T1 and T2, but because these tanks are connected, the differences between the plant's response and the model's are more pronounced. Furthermore, the tank levels 1 to 4 on the right side are illustrated for the validation of the white-box model, where the observed differences are very similar to the model's estimated response.

4.2. Gray-Box Model

The results of the gray-box model are shown in Figure 7, where the plant's response (blue) is presented alongside that of the gray-box model (red). This demonstrates a greater fit than the white-box model, as the gray-box model was constructed by combining parts of the white-box model's physical equations with the nonlinear behavior of the pump flow rates. The figure shows the tank levels 1 to 4 on the left side for the gray-box model estimation. Note how accurately the model fits the actual plant behavior in T3 and T4. The slight difference observed is due to the pumps experiencing cavitation losses, causing a drop in flow rate and, consequently, a decrease or increase in the stationary levels in the tanks. The results for the levels in T1 and T2 are also very similar; however, the main difference in this case is due to the instability of the levels caused by pump cavitation, which is introduced by the coupling of the tanks. Similarly, the results of the gray-box model validation are illustrated on the right side in Figure 7, where the model's high predictive capacity is evident for the levels of T3 and T4, as well as for the levels of T1 and T2.

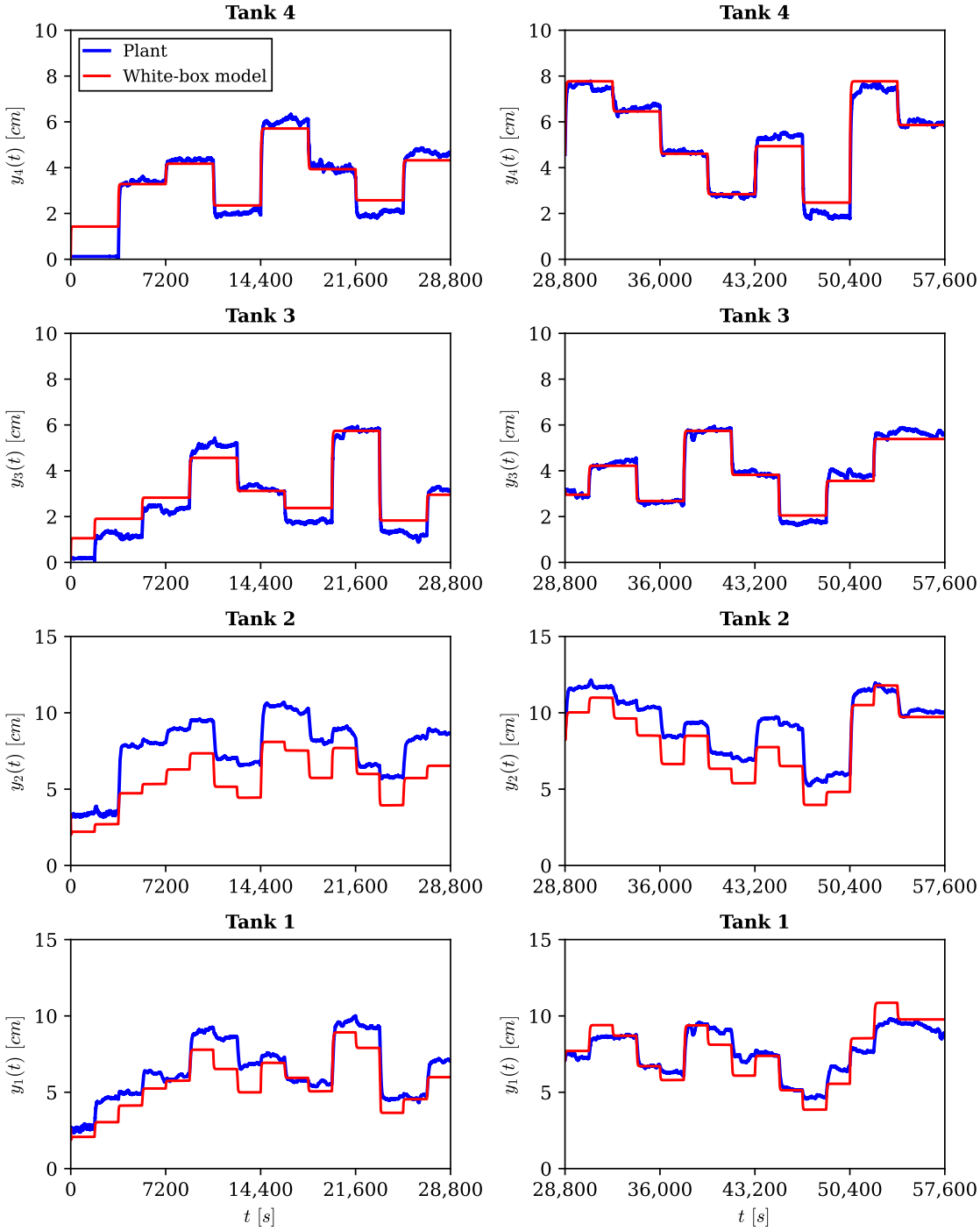


Figure 6. Plant response (blue) versus white-box model (red). Tank levels 1 to 4 on the left side for model estimation and on the right side for model validation, respectively.

4.3. Black-Box Model

The plant's response (blue) with respect to the black-box model (red) is shown in Figure 8. Using SR based on GP, a simple and accurate black-box model has been obtained, with better performance than the white-box model and an accuracy similar to that of the gray-box model, as can be noticed by comparing Figure 8 with Figures 6 and 7. Figure 8 illustrates the tank levels 1 to 4 on the left side for model estimation. Similarly, the tank levels 1 to 4 can be observed on the right side for the validation of the model.

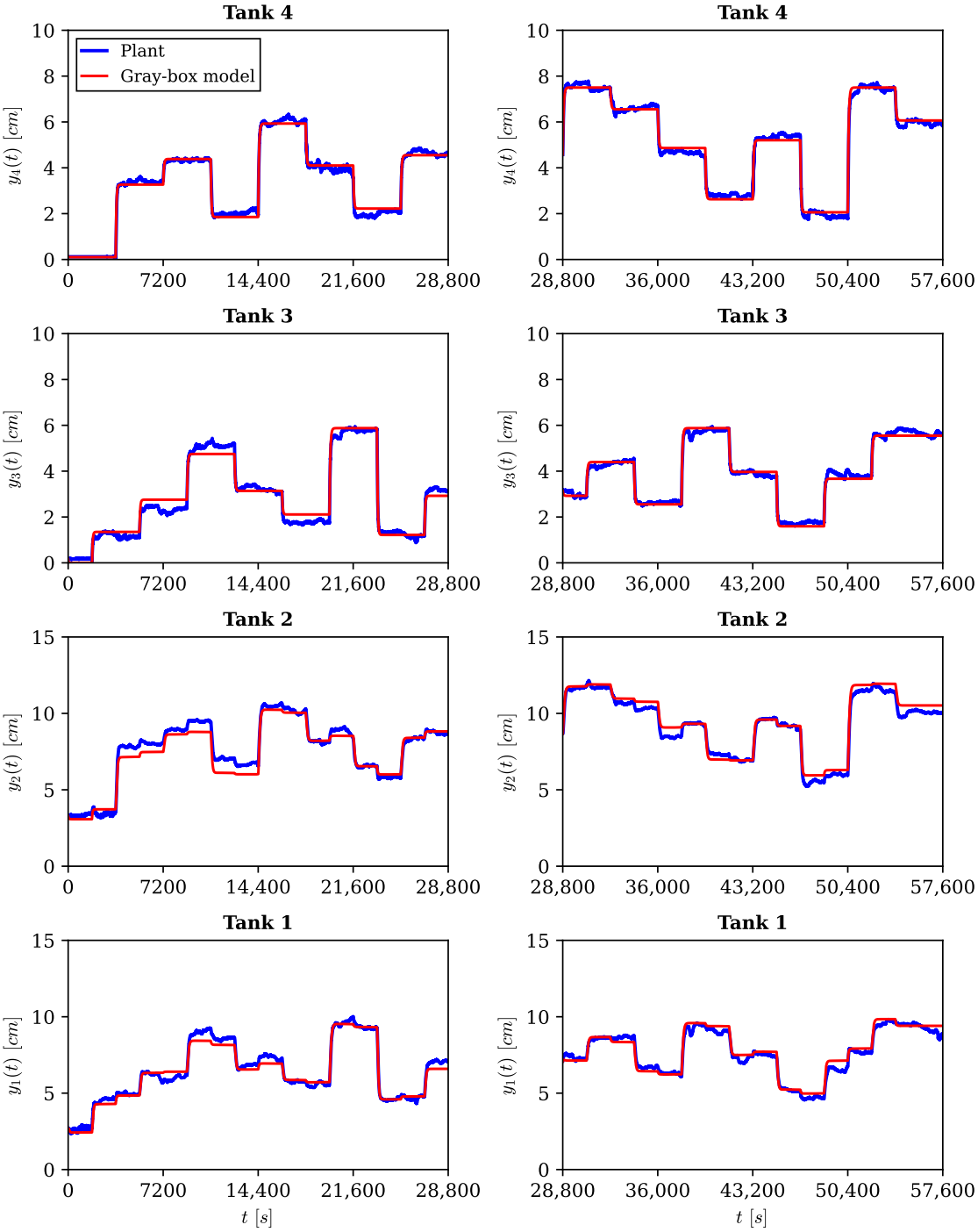


Figure 7. Plant response (blue) versus gray-box model (red). Tank levels 1 to 4 on the left side for model estimation and on the right side for model validation, respectively.

4.4. Goodness-of-Fit Performance Metrics

While the performance of the resulting models is shown graphically in the figures above, a more objective comparison is obtained by assessing their quality using the RMSE, the FIT and the R^2, as described in Section 2.3. Tables 8 and 9 present the quality metrics of the models obtained during estimation and validation.

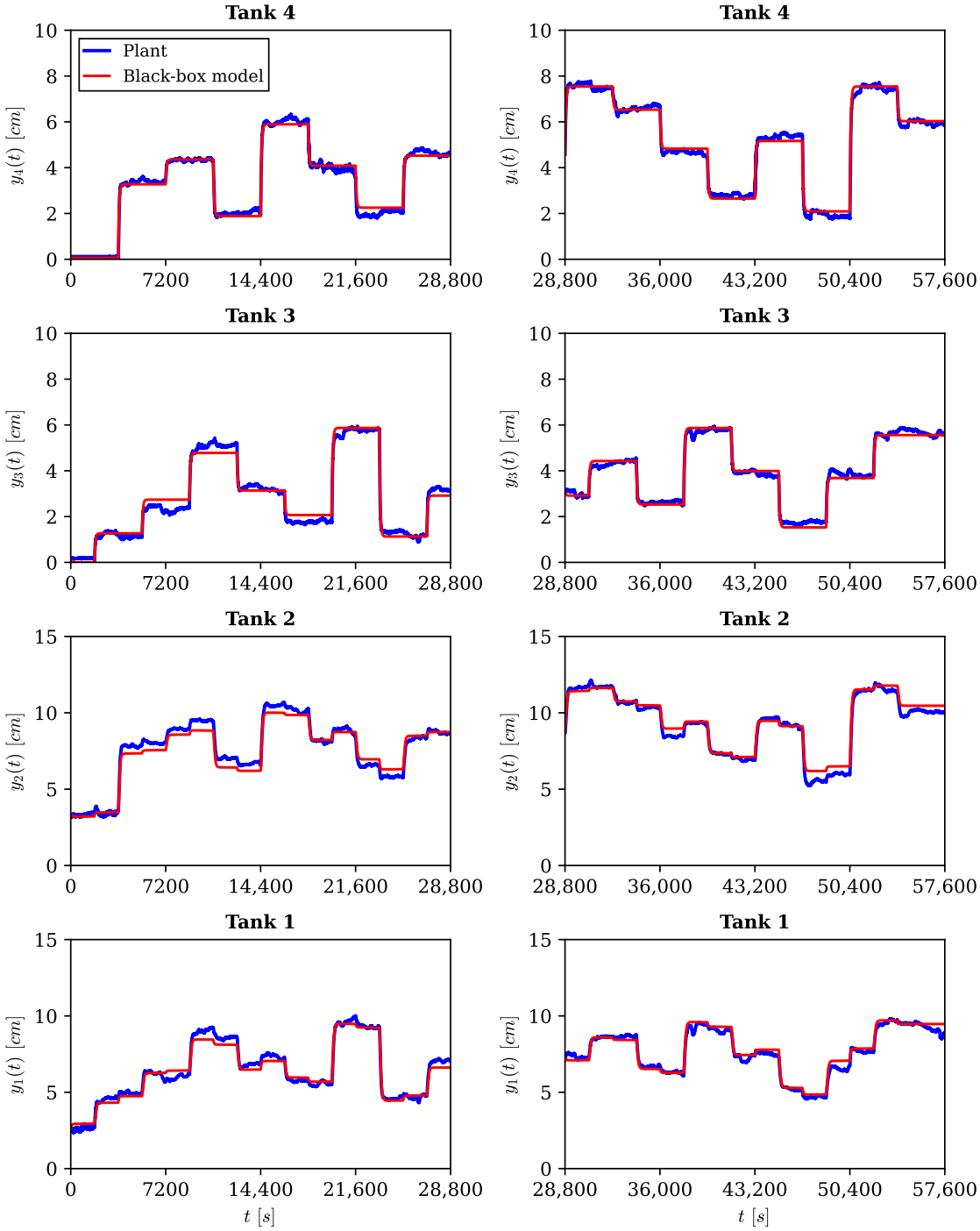


Figure 8. Plant response (blue) versus black-box model (red). Tank levels 1 to 4 on the left side for model estimation and on the right side for model validation, respectively.

Tables 8 and 9 show that the white-box model is clearly the least accurate, especially in the intermediate tanks, where the errors of the upper tanks and of the linear pump model accumulate. The gray-box and black-box models achieve similar accuracy in both datasets, and neither loses accuracy from estimation to validation. Their advantages and limitations are discussed in Section 4.5.

Table 8. Estimation dataset performance metrics. The best value of each metric per tank level is highlighted in green and shown in bold.

Model	Tank Level	RMSE [cm]	FIT [%]	R^2 [–]
White-box	y_1	1.0805	43.48	0.6806
	y_2	2.1429	−5.90	−0.1214
	y_3	0.5378	68.69	0.9020
	y_4	0.5630	67.57	0.8948
Gray-box	y_1	**0.3117**	**83.70**	**0.9734**
	y_2	0.4472	77.90	0.9512
	y_3	0.2716	84.19	0.9750
	y_4	**0.1591**	**90.84**	**0.9916**
Black-box	y_1	0.3168	83.43	0.9725
	y_2	**0.4045**	**80.01**	**0.9600**
	y_3	**0.2548**	**85.17**	**0.9780**
	y_4	0.1647	90.51	0.9910

Table 9. Validation dataset performance metrics. The best value of each metric per tank level is highlighted in green and shown in bold.

Model	Tank Level	RMSE [cm]	FIT [%]	R^2 [–]
White-box	y_1	0.7298	50.15	0.7515
	y_2	1.3853	28.30	0.4859
	y_3	0.2249	83.38	0.9724
	y_4	0.3285	82.85	0.9706
Gray-box	y_1	0.2946	79.88	0.9595
	y_2	0.3630	81.21	0.9647
	y_3	**0.1512**	**88.82**	**0.9875**
	y_4	0.1662	91.32	0.9925
Black-box	y_1	**0.2607**	**82.19**	**0.9683**
	y_2	**0.3359**	**82.61**	**0.9698**
	y_3	0.1615	88.07	0.9858
	y_4	**0.1652**	**91.37**	**0.9926**

It should be noted that the white-box and gray-box models do not use the PRS estimation data: their parameters are obtained from physical knowledge and from short dedicated experiments, so their accuracy does not depend on the amount of PRS data. In contrast, the black-box model relies entirely on the estimation dataset and is only valid within the operating region covered by it.

The effect of the amount of identification data on the accuracy of the three approaches has not been studied, since it requires additional experiments of different durations. This study will be addressed in future work.

4.5. Physical Consistency, Validity Range and Suitability for Control

The white-box (4) and gray-box (12) models share the same structure, derived from mass balances and Torricelli's law. In both models, the pump flows are zero at $u = 0$ V and strictly increasing with the voltage (linear in (4) and the cubic curves (13) in (12)), and the outflow of each tank increases with its level. Therefore, every constant input leads to a unique and stable equilibrium, the steady-state levels increase with the pump voltages, and the tanks empty when the pumps stop. The outflow vanishes at $y_i = 0$ in the white-box model and at $y_i = -\beta$ in the gray-box model. Both models are thus physically consistent over the whole input range; they differ in their accuracy, since the white-box model uses

nominal parameters and linear pumps, whereas the gray-box model uses the identified discharge coefficients and pump curves.

Within the identified range, $u_1, u_2 \in [3, 7]$ V, the black-box model (16) is also physically consistent. Each equation has a unique and stable equilibrium, and its steady-state gains are positive: the levels of tanks 3 and 4 increase with the voltage of their pump, and those of tanks 1 and 2 increase with both the voltage of their pump and the level of the upper tank that drains into them, as in the plant. The levels of tanks 1 and 2 cannot become negative, since the right-hand sides of $\dot{y}_1$ and $\dot{y}_2$ are positive at zero level. For tanks 3 and 4, the equilibrium level approaches zero at the lower end of the range and becomes slightly negative below $u_2 \approx 3.3$ V and $u_1 \approx 3.0$ V, respectively; at these voltages, the tanks are practically empty in the experiments, and the simulated levels are saturated at zero. Outside the identified range, the fitting intercepts dominate and the model loses its physical meaning: for $u \to 0$, it predicts residual levels in tanks 1 and 2 (about 2.0 and 1.2 cm) instead of emptying, and the levels of tanks 3 and 4 are kept at zero only by the saturation of the simulation. Above 7 V, the model remains monotonic, but its predictions are extrapolations not supported by data.

For control design, the three models are smooth for positive levels and can be linearised around any operating point. The white-box model allows an analytic design, but its low accuracy limits it to preliminary analysis. In the black-box model, the steady-state gains have the physically expected sign and the inputs enter each equation through a strictly increasing, and therefore invertible, function (linear in tank 1 and a square root in tanks 2–4), so it can be used in nonlinear model-based control, such as feedback linearisation or nonlinear model predictive control, provided that the input constraints $u_1, u_2 \in [3, 7]$ V are enforced. When the controller may drive the plant outside this range, such as during start-up or emptying, the gray-box model is preferable. Table 10 summarises the advantages and limitations of each model.

According to Table 10, the three models differ in structure, identification effort and practical use. The gray-box and black-box models achieve similar accuracy on the validation dataset (mean RMSE of 0.244 cm and 0.231 cm, and mean FIT of 85.3% and 86.1%, respectively). In validation, the black-box model is slightly more accurate in y_1, y_2 and y_4 and the gray-box model in y_3, with differences of at most 2.3 percentage points; on the estimation dataset, each model is more accurate in two tanks and the mean FIT is practically the same (84.2% versus 84.8%). Both models keep their accuracy from the estimation to the validation dataset, whereas the white-box model remains far behind (61.2% in validation). Since the differences in accuracy are small, the choice depends on the other aspects in Table 10. Both models have the same number of fitted parameters (16) and a negligible computational cost. All the parameters of the gray-box model have physical meaning, and the model remains physically consistent outside the identified range, including emptying and start-up; however, it requires prior knowledge of the plant geometry and dedicated experiments (tank-emptying and pump tests). The black-box model is obtained from normal-operation data with no structural assumption beyond the choice of inputs in (14) and the plausibility check used to select the final model among the GP runs. Its terms reproduce the couplings of the plant with the expected signs, but its parameters have no direct physical meaning. Hence, the gray-box model is preferable when extrapolation, interpretability or physical consistency are required, for example for control design over the whole operating range, whereas the black-box model is a convenient alternative when dedicated experiments are not feasible and the plant operates within a known envelope with enforced input constraints.

Table 10. Comparison of the three models in terms of structure, identification and use.

	White-Box	Gray-Box	Black-Box
Model structure			
Modeling principle	Mass balance, nominal parameters	Mass balance, identified C_{d_i}, pump curves	Symbolic regression (GP)
Prior knowledge	Geometry, valve gains, linear pumps	Geometry and model structure	Candidate inputs; plausibility check in the final selection
Fitted parameters	2 (k_1, k_2)	16 (C_{d_i}, pump curves)	16 (gene weights, biases, ERCs)
Physical meaning	All	All	Partial (couplings and signs)
Identification			
Experiments	One flow test	Emptying and pump tests	Estimation dataset (8 h)
Procedure	Direct calculation	Least squares per test	GP and output-error refinement
Identification time	≈2 min	≈17 h of tests	≈30 min per tank (20 runs in parallel and refinement)
Use of the PRS dataset	Validation only (both halves)	Validation only (both halves)	First half (8 h) for estimation, second half (8 h) for validation
Use of the model			
Evaluation cost	4 square roots	4 square roots, 4 cubic polynomials	5 square roots
Outside $u \in [3,7]$ V	Physically consistent	Physically consistent	Not consistent
At low levels	Outflow vanishes at $y_i = 0$	Outflow vanishes at $y_i = -\beta$	Saturation at zero (T3, T4)
Control design	Analytic, low accuracy	Whole range	Valid within [3,7] V
Recommended use	Preliminary analysis, teaching	Extrapolation, interpretability	Known operating envelope

5. Conclusions

In this work, white-box, gray-box and black-box models of a real four-tank plant in non-minimum phase configuration have been identified and compared on the same experimental dataset, obtained by exciting the plant with PRS signals over several operating points. The white-box model, based on nominal parameters and linear pumps, is the simplest one, but it is clearly the least accurate (mean validation FIT of 61.2%). The gray-box model keeps the physical structure and replaces the ideal assumptions with discharge coefficients and third-order pump curves identified in dedicated experiments, which raises the mean validation FIT to 85.3%. The black-box model, an ODE system whose structure is found by SR with GP and whose parameters are then refined in simulation, reaches a similar accuracy (86.1%) from normal-operation data only, with results that are repeatable over 20 independent GP runs. Since both models are similarly accurate, the choice depends on their use: the gray-box model is physically consistent over the whole operating range and is preferable when extrapolation or interpretability is required, whereas the black-box model needs no dedicated experiments but is valid only within the identified input range.

It should be noted that the models have been validated on a single experimental dataset, obtained with the same type of excitation signal and in the same plant configuration as the estimation data. Therefore, the results demonstrate a good fit on the validation dataset within the identified operating range, but they do not establish robustness against different excitation signals, disturbances or operating conditions, which will be addressed in future work.

As additional future work, a larger reservoir will be installed, and the pumps will be readjusted to prevent cavitation, thus ensuring more stable levels in the tanks. Pressure transmitters will be positioned at the bottom of the tanks to improve level readings, and other imperfections in the plant will be corrected. This will result in a more complete and efficient dataset for future experiments.

With the plant adjusted, new experiments will be conducted in both minimum and non-minimum phase configurations. During the model estimation stage, the plant will be excited with PRS signals, ensuring that steady-state operation is reached at each step. For the model validation phase, the plant will be excited using a combination of APRBS and multisine signals, covering both transient and steady-state regimes.

It would be highly relevant to identify the three-way valves and integrate them into the gray-box model, so that they are no longer considered a fixed plant parameter but rather a continuous input variable. This would allow for real-time adjustment of the flow fraction directed to the tanks. With this modification, it will be possible to change the plant configuration between minimum and non-minimum phases in real-time. Experiments of different durations will also be used to study the effect of the amount of data.

Finally, the identified models will be used to design model-based nonlinear controllers for the real plant, and other nonlinear identification techniques, such as NARX models or neural networks, could be added to the comparison. In addition, the physical knowledge used here only to select the final black-box model will be incorporated into the search itself, by means of physically informed symbolic regression in which the function set or the structure of the genes includes terms such as the square root of the hydraulic head, in order to improve the extrapolation of the black-box model at low levels.

Author Contributions: Conceptualization, A.E., D.M.-S. and D.C.G.; methodology, A.E., D.M.-S. and D.C.G.; software, A.E.; validation, A.E., D.M.-S. and D.C.G.; formal analysis, A.E.; investigation, A.E.; resources, A.E.; data curation, A.E.; writing—original draft preparation, A.E.; writing—review and editing, A.E., D.M.-S. and D.C.G.; visualization, A.E.; supervision, D.M.-S. and D.C.G.; project administration, D.M.-S. and D.C.G.; funding acquisition, D.M.-S. and D.C.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was carried out with funding received from the Ministry of Science and Innovation through project reference PID2023-146540OB-C41 (NEMO4EXOAMIC).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset and GPTIPS configuration settings supporting the findings of this study have been deposited in e-cienciaDatos (Consorcio Madroño) and are publicly accessible at https://doi.org/10.21950/DG8EOO.

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

References

1. Schoukens, J.; Ljung, L. Nonlinear System Identification: A User-Oriented Road Map. *IEEE Control Syst. Mag.* **2019**, *39*, 28–99. [CrossRef]
2. Barbosa, B.H.G.; Aguirre, L.A.; Martinez, C.B.; Braga, A.P. Black and Gray-Box Identification of a Hydraulic Pumping System. *IEEE Trans. Control Syst. Technol.* **2011**, *19*, 398–406. [CrossRef]
3. Uçak, K.; Öke, G. Modeling of quadruple tank system using support vector regression. In Proceedings of the 2011 International Symposium on Innovations in Intelligent Systems and Applications, Istanbul, Turkey, 15–18 June 2011; pp. 234–240. [CrossRef]
4. Cohen, B.; Beykal, B.; Bollas, G. Dynamic System Identification from Scarce and Noisy Data Using Symbolic Regression. In Proceedings of the 2023 62nd IEEE Conference on Decision and Control (CDC), Singapore, 13–15 December 2023; pp. 3670–3675. [CrossRef]
5. Borjas, S.; Garcia, C. Subspace Identification for Industrial Processes. *Trends Comput. Appl. Math.* **2011**, *12*, 183–194. [CrossRef]

6. Ljung, L. Black-box models from input-output measurements. In Proceedings of the IMTC 2001. 18th IEEE Instrumentation and Measurement Technology Conference. Rediscovering Measurement in the Age of Informatics (Cat. No. 01CH 37188), Budapest, Hungary, 21–23 May 2001; Volume 1, pp. 138–146. [CrossRef]
7. Giordano, G.; Sjöberg, J. Black-and white-box approaches for cascaded tanks benchmark system identification. *Mech. Syst. Signal Process.* **2018**, *108*, 387–397. [CrossRef]
8. Bohlin, T. Prospects and Problems. In *Practical Grey-Box Process Identification: Theory and Applications*; Advances in Industrial Control; Springer Science & Business Media: London, UK, 2006; Chapter 1, pp. 10–12. [CrossRef]
9. Ljung, L. Models for Time-varying and Nonlinear Systems. In *System Identification: Theory for the User*, 2nd ed.; Pearson Education: Upper Saddle River, NJ, USA, 1998; Chapter 5, pp. 148–152.
10. Azam, S.N.M.; Jørgensen, J.B. Modeling and simulation of a modified quadruple tank system. In Proceedings of the 2015 IEEE International Conference on Control System, Computing and Engineering (ICCSCE), Penang, Malaysia, 27–29 November 2015; pp. 365–370. [CrossRef]
11. Jayaprakash, J.; SenthilRajan, T.; Harish Babu, T. Analysis of Modelling Methods of Quadruple Tank System. *Int. J. Adv. Res. Electr. Electron. Instrum. Eng.* **2014**, *3*, 11552–11565. [CrossRef]
12. Maxim, A.; Ionescu, C.; De Keyser, R. Modelling and identification of a coupled sextuple water tank system. In Proceedings of the 2016 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR), Cluj-Napoca, Romania, 19–21 May 2016; pp. 1–6. [CrossRef]
13. Nath, U.M.; Dey, C.; Mudi, R.K. Model identification of coupled-tank system—MIMO process. In Proceedings of the 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, India, 22–24 February 2017; pp. 1–6. [CrossRef]
14. Milasi, R.M.; Huang, B. On Identification of a Coupled Four Tank System. In Proceedings of the 2021 16th International Conference on Computer Science & Education (ICCSE), Lancaster, UK, 17–21 August 2021; pp. 697–701. [CrossRef]
15. Mohamed, G.; Hossam, S.; Ahmed, A.H.; El-maged, A. Parameter estimation and PI control for a water coupled tank system. *J. Adv. Eng. Trends* **2020**, *38*, 147–159. [CrossRef]
16. Johansson, K.H.; Horch, A.; Wijk, O.; Hansson, A. Teaching multivariable control using the quadruple-tank process. In Proceedings of the 38th IEEE Conference on Decision and Control (Cat. No. 99CH36304), Phoenix, AZ, USA, 7–10 December 1999; Volume 1, pp. 807–812. [CrossRef]
17. Johansson, K.H. The quadruple-tank process: A multivariable laboratory process with an adjustable zero. *IEEE Trans. Control Syst. Technol.* **2000**, *8*, 456–465. [CrossRef]
18. Chamorro, W.; Rueda, P.V. Regulador cuadrático lineal y control predictivo aplicados en un sistema de cuatro tanques: Una comparación de rendimiento. *Enfoque UTE* **2018**, *9*, 204–216. [CrossRef]
19. Oyehan, I.A.; Osunleke, A.S.; Ajani, O.O. Closed-loop stability of a non-minimum phase quadruple tank system using a nonlinear model predictive controller with EKF. *ChemEngineering* **2023**, *7*, 74. [CrossRef]
20. Harsha, S.; Vinod, D.; SaiKrishna, P. Identification and Multi-Objective $\mathcal{H}_\infty$ Control Design for a Quadruple Tank System. In Proceedings of the 2019 IEEE 7th Conference on Systems, Process and Control (ICSPC), Melaka, Malaysia, 13–14 December 2019; pp. 70–75. [CrossRef]
21. Garrido, J.; Garrido-Jurado, S.; Vázquez, F. Grey-Box Modeling and Decoupling Control of a Lab Setup of the Quadruple-Tank System. *Actuators* **2024**, *13*, 87. [CrossRef]
22. Rodriguez, R.; Paredes, V.; Noa, G.; Trucios, L. Grey box identification and adaptive control in a water level system. *IOP Conf. Ser. Mater. Sci. Eng.* **2019**, *707*, 012003. [CrossRef]
23. Alvarez, H.; Lamanna, R.; Vega, P.; Revollar, S. Metodología para la Obtención de Modelos Semifísicos de Base Fenomenológica Aplicada a una Sulfitadora de Jugo de Caña de Azúcar. *Rev. Iberoam. Automática Informática Ind. RIAI* **2009**, *6*, 10–20. [CrossRef]
24. Oussar, Y.; Dreyfus, G. How to be a gray box: Dynamic semi-physical modeling. *Neural Netw.* **2001**, *14*, 1161–1172. [CrossRef] [PubMed]
25. Araújo, I.; Cavalcante, G.; Lúcio, Y.; Araújo, F. Nonlinear system identification of a MIMO quadruple tanks system using NARX model. *Prz. Elektrotechniczny* **2019**, *95*, 66–72. [CrossRef]
26. Stefanoiu, D.; Culita, J.; Voinea, A.C.; Voinea, V. Nonlinear Identification for Control by Using NARMAX Models. *Mathematics* **2024**, *12*, 2252. [CrossRef]
27. da Silva, M.F.H.; da Silva, R.S.; Leite, G.R.; Cordeiro, T.D.; de Araújo, Í.B. Subspace Methods for Dynamic Systems Identification with MOESP. In Proceedings of the Congresso Brasileiro de Automática-CBA, Rio de Janeiro, Brazil, 15–18 October 2024; Volume 4, p. 6. [CrossRef] [PubMed]
28. Subramanian, S.; Chidhambaram, G.B.; Dhandapani, S. Modeling and Validation of a Four-Tank System for Level Control Process Using Black Box and White Box Model Approaches. *IEEJ Trans. Electr. Electron. Eng.* **2021**, *16*, 282–294. [CrossRef]
29. Moreno-Salinas, D.; Besada-Portas, E.; López-Orozco, J.; Chaos, D.; de la Cruz, J.; Aranda, J. Symbolic regression for marine vehicles identification. *IFAC-PapersOnLine* **2015**, *48*, 210–216. [CrossRef]

30. Moreno, R.; Moreno-Salinas, D.; Aranda, J. Black-Box Marine Vehicle Identification with Regression Techniques for Random Manoeuvres. *Electronics* **2019**, *8*, 492. [CrossRef]
31. Scheffold, L.; Finkler, T.; Piechottka, U. Gray-Box system modeling using symbolic regression and nonlinear model predictive control of a semibatch polymerization. *Comput. Chem. Eng.* **2021**, *146*, 107204. [CrossRef]
32. Kronberger, G.; Kammerer, L.; Kommenda, M. Identification of dynamical systems using symbolic regression. In *Proceedings of the Computer Aided Systems Theory—EUROCAST 2019*; Moreno-Díaz, R., Pichler, F., Quesada-Arencibia, A., Eds.; Springer: Cham, Switzerland, 2020; pp. 370–377. [CrossRef]
33. Danai, K.; La Cava, W.G. Controller design by symbolic regression. *Mech. Syst. Signal Process.* **2021**, *151*, 107348. [CrossRef]
34. Köktürk-Güzel, B.E.; Beyhan, S. Symbolic regression based extreme learning machine models for system identification. *Neural Process. Lett.* **2021**, *53*, 1565–1578. [CrossRef]
35. Wang, Z.; Soares, C.G.; Zou, Z. Optimal design of excitation signal for identification of nonlinear ship manoeuvring model. *Ocean Eng.* **2020**, *196*, 106778. [CrossRef]
36. Shneiderman, D.; Palmor, Z. Properties and control of the quadruple-tank process with multivariable dead-times. *J. Process Control* **2010**, *20*, 18–28. [CrossRef]
37. Alvarado, I.; Limon, D.; Garcia-Gabin, W.; Alamo, T.; Camacho, E.F. An educational plant based on the quadruple-tank process. *IFAC Proc. Vol.* **2006**, *39*, 82–87. [CrossRef]
38. Rusli, E.; Ang, S.; Braatz, R.D. A Quadruple-tank process control experiment. *Chem. Eng. Educ.* **2004**, *38*, 174–181.
39. Castelo, G.; Garrido, J.; Vázquez, F. Ajuste, Configuración y Control de Cuatro Tanques Acoplados. In Proceedings of the XXIX Jornadas de Automática, Tarragona, Spain, 3–5 September 2008. (In Spanish)
40. Gouta, H.; Hadj Said, S.; M'sahli, F. Model-based Predictive and Backstepping controllers for a state coupled four-tank system with bounded control inputs: A comparative study. *J. Frankl. Inst.* **2015**, *352*, 4864–4889. [CrossRef]
41. Espinal, A.; Sánchez Moreno, J.; González González, S. Automatización de una Maqueta Industrial de Cuatro Tanques Mediante PLC y SCADA. Master's Thesis, Universidad Complutense de Madrid (UCM) y Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain, 2023.
42. Díaz Martínez, J.M.; Muñoz Mansilla, R. Diseño de experimentos y tratamiento de datos. In *Apuntes de la Asignatura: Identificación de Sistemas*; Lecture Notes; Universidad Nacional de Educación a Distancia (UNED): Madrid, Spain, 2022. (In Spanish)
43. GRUNDFOS, Bjerringbro, Denmark. MAGNA, UPE: Series 2000 Circulator Pumps. Grundfos Data Booklet. 2005. Available online: https://www.grundfos.com/es (accessed on 13 October 2026).
44. Bernal Romero, J.C.; Rodríguez Guzmán, H.E. Diseño y Simulación de Controladores Multivariables para el Sistema de los Cuatro Tanques de Johansson y Estudio del Comportamiento del Sistema en lazo Cerrado a Ataques Ciber-Físicos. Bachelor's Thesis, Universidad Distrital Francisco José de Caldas, Bogotá, Colombia, 2018.
45. Mott, R.L.; Untener, J.A. Flow of fluids and Bernoulli's equation. In *Applied Fluid Mechanics*, 7th ed.; Pearson: Boston, MA, USA, 2015; Chapter 6, pp. 133–169.
46. Lienhard V, J.H.; Lienhard, J.H., IV. Velocity Coefficients For Free Jets From Sharp-Edged Orifices. *J. Fluids Eng.* **1984**, *106*, 13–17. [CrossRef]
47. Sotelo Avila, G. Orificios y compuertas. In *Hidráulica General*; Caniem (Series); Editorial Limusa: Ciudad de México, Mexico, 1994; Volume 1, Chapter 6, pp. 203–204.
48. Albertson, M.L.; Barton, J.R.; Simons, D.B. Flow Measurement. In *Fluid Mechanics for Engineers*; Prentice-Hall Civil Engineering and Engineering Mechanics Series; Prentice-Hall, Inc.: Englewood Cliffs, NJ, USA, 1960; Chapter 11.
49. Trout, T.J. Orifice Plates for Furrow Flow Measurement. Part I: Calibration. *Trans. ASAE* **1986**, *29*, 103–106. [CrossRef]
50. Koza, J.R. Genetic programming as a means for programming computers by natural selection. *Stat. Comput.* **1994**, *4*, 87–112. [CrossRef]
51. Chapra, S.C.; Canale, R.P. Errores de truncamiento y la serie de Taylor. In *Métodos Numéricos para Ingenieros*, 7th ed.; McGraw-Hill Interamericana: Ciudad de México, Mexico, 2015; Chapter 4, pp. 90–93.
52. Searson, D.P. GPTIPS 2: An Open-Source Software Platform for Symbolic Data Mining. In *Handbook of Genetic Programming Applications*; Springer International Publishing: Cham, Switzerland, 2015; Chapter 7, pp. 551–573. [CrossRef]
53. Smits, G.F.; Kotanchek, M. Pareto-Front Exploitation in Symbolic Regression. In *Genetic Programming Theory and Practice II*; O'Reilly, U.M., Yu, T., Riolo, R., Worzel, B., Eds.; Genetic Programming; Springer: Boston, MA, USA, 2005; Volume 8, pp. 283–299. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.