

Article

Data-Driven Linear Representations of Forced Nonlinear MIMO Systems via Hankel Dynamic Mode Decomposition with Lifting

Marcos Villarreal-Esquivel ¹, Juan Francisco Durán-Siguenza ² and Luis Ismael Minchala ^{1,*}¹ Department of Electrical Engineering, Electronics, and Telecommunications, Universidad de Cuenca, Cuenca 010101, Ecuador; marcos.villarreal@ucuenca.edu.ec² Department of Industrial Engineering, Università di Trento, Via Sommarive, 38123 Trento, Italy; juan.duransiguenza@unitn.it

* Correspondence: ismael.minchala@ucuenca.edu.ec

Abstract

Modeling forced nonlinear multivariable dynamical systems remains challenging, particularly when first-principles models are unavailable or strong nonlinear couplings are present. In recent years, data-driven approaches grounded in the Koopman operator theory have gained attention for their ability to represent nonlinear dynamics via linear evolution in appropriately lifted spaces. This work presents a data-driven modeling framework for forced nonlinear multiple-input multiple-output (MIMO) systems based on Hankel Dynamic Mode Decomposition with control and lifting functions (HDMDc+Lift). The proposed methodology exploits Hankel matrices to encode temporal correlations and employs lifting functions to approximate the Koopman operator's action on observable functions. As a result, an augmented-order linear state-space model is identified exclusively from input–output data, without relying on explicit knowledge of the system's governing equations. The effectiveness of the proposed approach is demonstrated using operational data from a real multivariable tank system that was not used during the identification stage. The identified model achieves a coefficient of determination exceeding 0.87 in multi-step prediction tasks. Furthermore, spectral analysis of the resulting linear operator reveals that the dominant dynamical modes of the physical system are accurately captured. At the same time, additional modes associated with nonlinear interactions are also identified. These results highlight the HDMDc+Lift framework's ability to provide accurate and interpretable linear representations of forced nonlinear MIMO dynamics.

Keywords: DMDc; data-driven modeling; HDMDc; Koopman operator; lifting; nonlinear systems; multitank system

MSC: 37M10



Academic Editor: Jonathan Blackledge

Received: 6 January 2026

Revised: 4 February 2026

Accepted: 5 February 2026

Published: 11 February 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Modeling dynamical systems is a fundamental problem in control engineering [1,2], as it underpins system analysis, prediction, supervision, and the synthesis of control strategies for industrial processes. Classical modeling approaches are typically derived from first principles, in which the system's behavior is described by differential equations derived from physical laws [3]; while effective in many applications, this paradigm becomes limited in the presence of strong nonlinearities and complex coupling effects, or when accurate physical descriptions are unavailable or impractical to obtain, motivating the development of data-driven modeling approaches [4].

In many contemporary industrial systems, increasing process complexity leads to unmodeled dynamics, parametric uncertainty, and time-varying operating conditions, which significantly complicate the development of accurate first-principles models [5]. As system dimensionality grows, purely physics-based representations may exhibit limited accuracy or reduced validity, particularly in large-scale nonlinear MIMO systems [6].

In this context, data-driven modeling has emerged as a compelling alternative, enabled by advances in sensing, computation, and system identification methodologies [7,8]. Unlike first-principles approaches, data-driven methods construct approximate models directly from input–output data, making them well suited for complex systems where physical modeling is infeasible or insufficiently accurate.

Among data-driven approaches, operator-theoretic methods based on the Koopman operator have gained considerable attention in recent years. The Koopman operator provides a linear infinite-dimensional representation of nonlinear dynamical systems by the evolution of observable functions of the state [9,10]. This linearity property has enabled the development of numerical techniques that approximate the Koopman operator using measured data, allowing nonlinear dynamics to be analyzed through linear models in well chosen feature spaces.

Of the many methods for estimating the Koopman operator, Dynamic Mode Decomposition (DMD) stands out as one of the most prominent computational approaches for approximating the operator from measurements [11,12]. DMD identifies dominant spatiotemporal modes and their associated spectral properties, offering insight into the underlying system dynamics. However, the classical DMD formulation is restricted to autonomous systems, limiting its direct applicability to systems subject to external forcing or control inputs [12].

To address this limitation, Dynamic Mode Decomposition with control (DMDc) was introduced, explicitly incorporating control inputs into the identification process [13]. This extension enables the separation of autonomous system dynamics from the effects of exogenous inputs, making the resulting models more suitable for control-oriented analysis and prediction. Furthermore, the use of time-delay embeddings via Hankel matrices has proven effective in enriching the dynamic representation by capturing temporal correlations and latent states, leading to approaches such as Hankel-DMD [14].

Despite these advances, the accurate modeling of forced nonlinear multiple-input multiple-output (MIMO) systems remains an open and challenging problem. Strong nonlinearities and coupling effects often limit the effectiveness of Koopman-based representations constructed solely from delay embeddings or linear observables [15]. To overcome these limitations, lifting functions have emerged as an effective mechanism for enriching the observable space. By mapping measured variables into higher-dimensional feature spaces, lifting functions enable the approximation of complex nonlinear dynamics via linear evolution in an extended coordinate system, thereby enhancing the expressive power of Koopman-based models [16,17].

This paper introduces a fully data-driven modeling framework for forced nonlinear MIMO dynamical systems that integrates Hankel Dynamic Mode Decomposition with control and lifting functions (HDMDc+Lift). The proposed approach leverages Hankel embeddings to encode temporal correlations and lifting functions to enrich the observable space, enabling the identification of an augmented-order linear state-space representation directly from measured input–output data. Importantly, the methodology does not rely on explicit knowledge of the underlying physical equations, making it applicable to complex systems for which first-principles modeling is impractical or incomplete.

The effectiveness of the proposed framework is demonstrated through a case study based on a real multivariable tank system, using operational data that are independent

from those employed during the identification stage. The obtained results indicate that the HDMDc+Lift model provides accurate multi-step predictions while preserving the dominant dynamical structure of the physical system. Furthermore, spectral analysis of the identified linear operator suggests that the main dynamical modes are correctly captured, along with additional modes associated with nonlinear interactions. These characteristics highlight the potential of the proposed approach as a reliable foundation for subsequent tasks, including system analysis, monitoring, and fault diagnosis, in complex industrial processes.

2. Methodology

2.1. Problem Setting and Modeling Objective

Consider a forced nonlinear MIMO dynamical system described in discrete time by

$$\mathbf{x}_{k+1} = f(\mathbf{x}_k, \mathbf{u}_k), \quad (1)$$

$$\mathbf{y}_k = h(\mathbf{x}_k), \quad (2)$$

where $\mathbf{x}_k \in \mathbb{R}^n$ denotes the (possibly unmeasured) system state, $\mathbf{u}_k \in \mathbb{R}^m$ represents the control inputs or external perturbations, and $\mathbf{y}_k \in \mathbb{R}^p$ corresponds to the measured outputs. The nonlinear mappings $f(\cdot)$ and $h(\cdot)$ govern the state evolution and output generation, respectively.

In many practical applications, the functions f and h are unknown, partially known, or too complex to be accurately derived from first principles. Strong nonlinearities, multi-variable coupling effects, and unmodeled dynamics further complicate the development of reliable physics-based models. Consequently, the objective of this work is to identify an approximate dynamic representation directly from measured input–output data that enables accurate prediction while preserving the dominant dynamical structure of the underlying nonlinear system.

2.2. Data-Driven Approximations and DMD with Control

Dynamic Mode Decomposition (DMD) was originally introduced as a data-driven method for extracting coherent structures and dominant dynamics from time-series data [11,12]. From a Koopman perspective, DMD can be interpreted as a finite-dimensional approximation of the Koopman operator based on snapshot data. The resulting linear model captures the evolution of measured observables between consecutive time steps.

However, the classical formulation of DMD assumes autonomous system dynamics and does not explicitly account for the influence of external inputs. As a consequence, when applied to forced systems, the effect of control actions is implicitly absorbed into the identified dynamics, limiting the interpretability and applicability of the resulting model.

To overcome this limitation, Dynamic Mode Decomposition with control (DMDc) was proposed, extending the original framework by incorporating control inputs as exogenous variables [13]. In this case, the system is approximated by a linear model of the form

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k, \quad (3)$$

where the matrices $\mathbf{A}$ and $\mathbf{B}$ are identified directly from data. This formulation explicitly separates the intrinsic system dynamics from the effects of control inputs, making it more suitable for control-oriented analysis, prediction, and monitoring applications.

From a robustness and disturbance-rejection perspective, the DMDc formulation can be interpreted as an implicit separation between intrinsic system dynamics and exogenous inputs, a concept that is also central to several advanced control and estimation frameworks.

For example, equivalent-input disturbance (EID) approaches explicitly augment the system dynamics to estimate and compensate unknown perturbations acting on the system [18]. Although DMDc does not explicitly model disturbances, the identification of separate operators **A** and **B** enables the data-driven isolation of forced and autonomous components of the dynamics, which is conceptually aligned with disturbance-decoupling objectives.

Similarly, robust zeroing neural network formulations employ augmented dynamic structures to enforce convergence properties in the presence of uncertainties [19]. In comparison, DMDc achieves a linear representation directly from data, where robustness emerges implicitly from the quality and richness of the measured input–output trajectories rather than from explicitly designed nonlinear feedback terms. This highlights the suitability of DMDc-based models as a lightweight yet effective alternative for data-driven analysis of forced nonlinear systems.

2.3. Hankel Embedding and Motivation for Lifting

While DMDc improves the modeling of forced systems, its ability to capture complex nonlinear dynamics remains limited when only instantaneous measurements are used. To address this limitation, time-delay embeddings based on Hankel matrices are employed, incorporating delayed versions of the measured signals into the data representation.

The use of Hankel embeddings allows the model to encode temporal memory and latent dynamics, effectively enriching the observable space without requiring direct access to the full system state. This approach has led to variants such as Hankel-DMD and Hankel-DMDc (HDMDc), which have demonstrated improved performance in the identification of nonlinear and high-dimensional systems.

In addition to time-delay embedding, lifting functions provide a complementary mechanism for enhancing model expressiveness. Lifting maps the measured variables into a higher-dimensional feature space, where nonlinear relationships can be more effectively approximated by linear dynamics. From a Koopman perspective, lifting corresponds to enriching the observable dictionary, increasing the likelihood that the chosen subspace captures invariant or approximately invariant dynamics.

The combination of Hankel embedding and lifting constitutes the conceptual foundation of the HDMDc+Lift approach proposed in this work. By jointly exploiting temporal delays and lifted observables, the resulting framework enables the identification of extended-order linear models capable of accurately approximating forced nonlinear MIMO dynamics, as detailed in the following section.

The objective of the proposed methodology is to identify a low-order linear approximation of a nonlinear MIMO system directly from input–output data, while preserving its dominant dynamic structure and predictive capability. To this end, a data-driven modeling approach based on the Koopman operator is employed, combining three key elements:

- Hankel time-delay embedding to capture temporal dependencies and latent dynamics;
- Dynamic Mode Decomposition with control (DMDc) to explicitly incorporate control inputs;
- Lifting of observables to enhance the representation of nonlinear effects.

The resulting method, referred to as HDMDc+Lift, produces a linear state-space model suitable for prediction, dynamic analysis, and fault detection tasks.

The use of Hankel time-delay embeddings can also be interpreted through the lens of state augmentation strategies commonly employed in robust estimation and fault-tolerant control. In dynamic high-gain and decentralized fault-tolerant control schemes, augmented states are often introduced to capture hidden coupling effects, internal dynamics, or fault signatures [20]. Analogously, Hankel embeddings construct an extended state representa-

tion from measured data that implicitly captures latent dynamics and nonlinear interactions without requiring explicit physical modeling.

From a mathematical standpoint, delay-embedded observables approximate higher-order Markov representations of the underlying nonlinear system. This property enables the lifted linear model to retain memory of past system behavior, which is particularly beneficial in the presence of unmodeled dynamics or slowly varying disturbances. As a result, the Hankel-based formulation enhances the robustness of the identified model while preserving its linear structure.

2.4. Hankel-Based Data Organization

Consider a set of measured output data of length T

$$\{\mathbf{y}_k\}_{k=1}^T, \quad \mathbf{y}_k \in \mathbb{R}^p, \quad (4)$$

and the corresponding control inputs

$$\{\mathbf{u}_k\}_{k=1}^T, \quad \mathbf{u}_k \in \mathbb{R}^m. \quad (5)$$

To incorporate temporal memory into the model, Hankel matrices of order q are constructed for both outputs and inputs.

The output Hankel matrix is defined as

$$\mathbf{X}_H = \begin{bmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_{T-q} \\ \mathbf{y}_2 & \mathbf{y}_3 & \cdots & \mathbf{y}_{T-q+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{y}_q & \mathbf{y}_{q+1} & \cdots & \mathbf{y}_T \end{bmatrix} \in \mathbb{R}^{qp \times (T-q+1)}. \quad (6)$$

Similarly, the input Hankel matrix is constructed as

$$\mathbf{U}_H = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_{T-q} \\ \mathbf{u}_2 & \mathbf{u}_3 & \cdots & \mathbf{u}_{T-q+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{u}_q & \mathbf{u}_{q+1} & \cdots & \mathbf{u}_T \end{bmatrix} \in \mathbb{R}^{qm \times (T-q+1)}. \quad (7)$$

This representation allows the model to encode temporal correlations and delayed nonlinear effects using only measured data.

2.5. Incorporation of Lifting Functions

To further enhance the expressive power of the model, the measured outputs are mapped to a lifted observable space through a function

$$\psi : \mathbb{R}^p \rightarrow \mathbb{R}^d. \quad (8)$$

In this work, the lifting is defined as a linear augmentation based on delayed measurements:

$$\psi(\mathbf{y}_k) = \begin{bmatrix} \mathbf{y}_k \\ \mathbf{y}_{k-1} \\ \vdots \\ \mathbf{y}_{k-q+1} \end{bmatrix}. \quad (9)$$

Although nonlinear lifting functions may be employed, empirical evaluation showed that delay-based lifting provides a favorable balance between model accuracy, numerical

stability, and computational cost for the multitank system. This choice is consistent with the Koopman framework, where delay coordinates are known to approximate invariant subspaces under mild observability conditions.

The lifted Hankel matrix is therefore constructed as

$$\mathbf{X}_{H,\psi} = \begin{bmatrix} \psi(\mathbf{y}_1) & \psi(\mathbf{y}_2) & \cdots & \psi(\mathbf{y}_{T-q}) \end{bmatrix}. \quad (10)$$

2.6. HDMDc Formulation in the Lifted Space

The lifted output and input Hankel matrices are concatenated to form the extended data matrix

$$\mathbf{\Omega} = \begin{bmatrix} \mathbf{X}_{H,\psi} \\ \mathbf{U}_H \end{bmatrix}. \quad (11)$$

The corresponding future-state matrix is defined as

$$\mathbf{X}'_{H,\psi} = \begin{bmatrix} \psi(\mathbf{y}_2) & \psi(\mathbf{y}_3) & \cdots & \psi(\mathbf{y}_{T-q+1}) \end{bmatrix}. \quad (12)$$

The objective is to identify a linear operator $\mathbf{G}$ such that

$$\mathbf{X}'_{H,\psi} \approx \mathbf{G}\mathbf{\Omega}. \quad (13)$$

This operator is estimated via truncated singular value decomposition (SVD):

$$\mathbf{\Omega} \approx \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^\top, \quad (14)$$

where r denotes the truncation rank.

The system operator is obtained as

$$\mathbf{G} = \mathbf{X}'_{H,\psi} \mathbf{V}_r \mathbf{\Sigma}_r^{-1} \mathbf{U}_r^\top. \quad (15)$$

The matrix $\mathbf{G}$ is partitioned as

$$\mathbf{G} = \begin{bmatrix} \mathbf{A}_H & \mathbf{B}_H \end{bmatrix}, \quad (16)$$

leading to the lifted linear state-space model

$$\mathbf{z}_{k+1} = \mathbf{A}_H \mathbf{z}_k + \mathbf{B}_H \mathbf{u}_k, \quad (17)$$

where $\mathbf{z}_k = \psi(\mathbf{y}_k)$ and, as demonstrated in (9), the first p rows correspond to the measured output at time instant k . Since the dimension of the identity matrix is equal to the number of system outputs, therefore, the output matrix $\mathbf{C}$ can be conveniently constructed as a combination of identity and zero matrices, as follows:

$$\mathbf{C} = [\mathbf{I}_{p \times p} \quad \mathbf{0}_{d-p \times d-p}] = \begin{bmatrix} 1_1 & 0 & \cdots & 0_p & 0_{p+1} & \cdots & 0_{d-p} \\ 0 & 1_2 & \cdots & 0_p & 0_{p+1} & \cdots & 0_{d-p} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1_p & 0_{p+1} & \cdots & 0_{d-p} \end{bmatrix} \quad (18)$$

The identification of the lifted operator $\mathbf{G}$ can be interpreted as constructing a linear surrogate of the nonlinear input–output dynamics in an augmented coordinate system. This perspective is closely related to hybrid modeling approaches that combine data-driven components with structured estimators. For instance, hybrid neural network- and physics-based estimators aim to balance expressive power with interpretability by embedding

nonlinear effects into extended state representations [21]. In contrast, the HDMDc+Lift framework embeds nonlinearities through delay coordinates and lifted observables, yielding a fully linear evolution model in the lifted space.

An important distinction is that the spectral properties of the operator $\mathbf{A}_H$ can be directly analyzed, enabling stability assessment, modal interpretation, and integration with observer-based monitoring schemes. This characteristic contrasts with purely neural-network-based approaches, where stability and robustness analysis often remain challenging. Consequently, the lifted Koopman-based representation provides a mathematically transparent and computationally efficient alternative for modeling complex nonlinear MIMO systems under external forcing.

2.7. Algorithmic Summary

The complete HDMDc+Lift procedure is summarized in Algorithm 1.

Algorithm 1 HDMDc with Lifting for Forced Nonlinear MIMO Systems.

Require: Output data $\{\mathbf{y}_k\}_{k=1}^T$, input data $\{\mathbf{u}_k\}_{k=1}^T$, Hankel order q , truncation rank r

Ensure: Lifted linear state-space model $(\mathbf{A}_H, \mathbf{B}_H)$

- 1: Construct output Hankel matrix $\mathbf{X}_H$
 - 2: Construct input Hankel matrix $\mathbf{U}_H$
 - 3: Define lifting function $\psi(\cdot)$ using delayed measurements
 - 4: Form lifted Hankel matrix $\mathbf{X}_{H,\psi}$
 - 5: Construct extended data matrix $\mathbf{\Omega}$
 - 6: Compute truncated SVD of $\mathbf{\Omega}$
 - 7: Estimate operator $\mathbf{G} = [\mathbf{A}_H \ \mathbf{B}_H]$
 - 8: **return** Linear lifted model $\mathbf{z}_{k+1} = \mathbf{A}_H \mathbf{z}_k + \mathbf{B}_H \mathbf{u}_k$
-

The proposed algorithm identifies a finite-dimensional linear approximation of the Koopman operator restricted to a lifted and delay-embedded observable space, enabling accurate prediction of forced nonlinear dynamics using only input–output data.

3. Case Study and Experimental Setup

3.1. Multitank System Description

The proposed methodology is validated using a laboratory-scale multitank system with three control inputs and two measured outputs. This system constitutes a well-established benchmark in control engineering research due to its nonlinear dynamics, multivariable interactions, and strong coupling effects among state variables.

The system dynamics are governed by nonlinear flow relationships between interconnected tanks, where variations in pump actuation affect multiple liquid levels simultaneously. These characteristics make the multitank system particularly suitable for evaluating data-driven modeling techniques and Koopman-based linear representations, as they challenge both model expressiveness and generalization capability.

The variables considered in this study are summarized as follows:

- Inputs: control signals applied to the feed pumps.
- Outputs: liquid levels measured in the tanks.

Due to inherent cross-couplings, saturation effects, and unmodeled disturbances, deriving accurate physics-based models for this system typically requires complex parameter identification procedures. As such, it provides a representative test case for assessing the effectiveness of fully data-driven modeling approaches. Further details of this experimental system in [12,22,23].

3.2. Data Acquisition and Preprocessing

The data used for model identification and validation were acquired directly from the real multitank system via a supervisory control and data acquisition (SCADA) platform. Several hours of operation under nominal conditions were recorded, ensuring sufficient excitation of the system dynamics across relevant operating regimes.

Prior to model construction, the collected data were subjected to a preprocessing pipeline consisting of the following steps:

- Removal of outliers and segments affected by sensor anomalies or operational transients.
- Normalization of input and output signals to mitigate scaling effects and improve numerical conditioning.
- Partitioning of the dataset into two disjoint subsets:
 - Training data, used exclusively for model identification.
 - Validation data, not employed during training and reserved for performance assessment.

This strict separation between training and validation datasets enables an objective evaluation of the generalization capability of the identified model and reduces the risk of overfitting.

3.3. HDMDc+Lift Model Configuration

The HDMDc+Lift model was constructed following the methodology described in Algorithm 1. The performance of the proposed HDMDc+Lift method depends on two key parameters: the Hankel matrix order q , which determines the temporal memory depth, and the SVD truncation rank r , which controls the number of retained dynamical modes.

A systematic parameter sweep was conducted to evaluate the trade-off between predictive accuracy and model complexity within the HDMDc+Lift framework. This analysis considered multiple combinations of the Hankel order q and truncation rank r , assessing both prediction performance and computational cost. The obtained results, summarized in Table 1, include the coefficient of determination for both system outputs, together with the corresponding training and prediction times.

The results indicate that increasing either the Hankel order q or the truncation rank r generally improves prediction accuracy up to a certain point; however, this improvement is accompanied by a significant increase in computational burden due to the higher-dimensional lifted state representation. This behavior is directly linked to the growth in the dimensions of the identified state-space matrices $\mathbf{A}$ and $\mathbf{B}$, which scale with both the number of delays and the dimension of the lifted observable space. Although several configurations with larger values of q and r achieve predictive performance comparable to the selected configuration ($q = 5, r = 15$), the associated training and prediction times grow markedly, reducing their practical suitability.

These values were found to provide an effective balance between predictive accuracy, numerical stability, and computational efficiency.

The resulting model is expressed in discrete-time linear state-space form as

$$\mathbf{z}_{k+1} = \mathbf{A}\mathbf{z}_k + \mathbf{B}\mathbf{u}_k, \quad (19)$$

$$\hat{\mathbf{y}}_k = \mathbf{C}\mathbf{z}_k, \quad (20)$$

where $\mathbf{z}_k$ denotes the lifted state vector, $\hat{\mathbf{y}}_k$ represents the estimated outputs, and the matrices $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ are identified directly from data.

Table 1. Comparative results obtained for the best selection of parameters related to training time and prediction.

Hankel Order (q)	Truncation Rank (r)	H1 (Fit %)	H2 (Fit %)	Training Time (s)	Prediction Time (s)
5	5	0.65004	0.544178	0.675	0.329
5	10	0.85847	0.83391	0.655	0.316
5	15	0.87794	0.86953	0.656	0.32
5	20	0.15297	NC	0.764	0.354
10	5	0.6578	0.55373	1.693	0.854
10	10	0.87504	0.8499	1.535	0.783
10	15	0.87504	0.8499	1.573	0.813
10	20	0.28832	NC	1.56	0.76
15	5	0.66303	0.55997	4.694	3.196
15	10	0.86744	0.84664	4.97	3.615
15	15	0.88051	0.87412	3.129	1.941
15	20	0.065866	NC	4.4195	2.71
20	5	0.64222	0.56521	5.1838	2.988
20	10	0.85543	0.84101	5.086	2.838
20	15	0.87962	0.87413	6.906	4.414
20	18	0.87189	0.88594	7.3876	4.73
20	20	0.84744	0.7987	5.7104	3.62

3.4. Validation Strategy and Performance Metrics

Model validation was conducted by comparing the estimated outputs $\hat{y}_k$ with the corresponding measured outputs of the multitank system using validation data that were not employed during the identification stage.

The performance of the identified model was assessed using the following criteria:

- Mean squared error (MSE), to quantify prediction accuracy.
- Coefficient of determination (R^2), to evaluate the proportion of explained output variance.
- Visual inspection of time-domain trajectories, to assess dynamic consistency and transient behavior.

In addition to predictive accuracy, the stability properties of the identified model were examined through spectral analysis of the state transition matrix **A**. The dominant eigenvalues were verified to lie within the unit circle, which constitutes a necessary condition for stability in discrete-time linear systems. This analysis provides further insight into the dynamical consistency of the learned model and its suitability for subsequent analysis and monitoring tasks.

4. Results and Discussion

4.1. Predictive Performance of the HDMDc+Lift Model

The model identified using the HDMDc+Lift methodology was evaluated using a dataset completely independent from that employed during the identification process. This strategy enabled an objective assessment of the model's generalization capability and prevented overfitting.

Figures 1 and 2 show the comparison between the real outputs of the multitank system and the outputs estimated by the model for both measured variables. A high correspondence between the real and estimated trajectories is observed, both during transient and steady-state regimes.

Quantitatively, the model achieved an average coefficient of determination of 87%, demonstrating strong predictive capability given the nonlinear and coupled nature of the system. This level of accuracy was achieved without relying on physical models or explicit nonlinear functions, reinforcing the effectiveness of the data-driven approach.

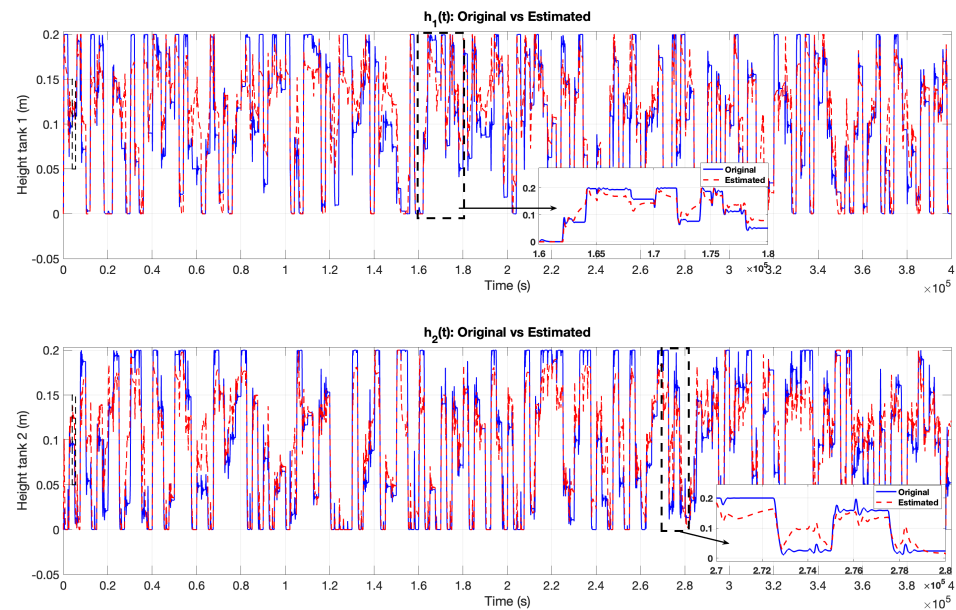


Figure 1. Performance of the HDMDc+Lift model in comparison with the training data ($q = 5, r = 15$).

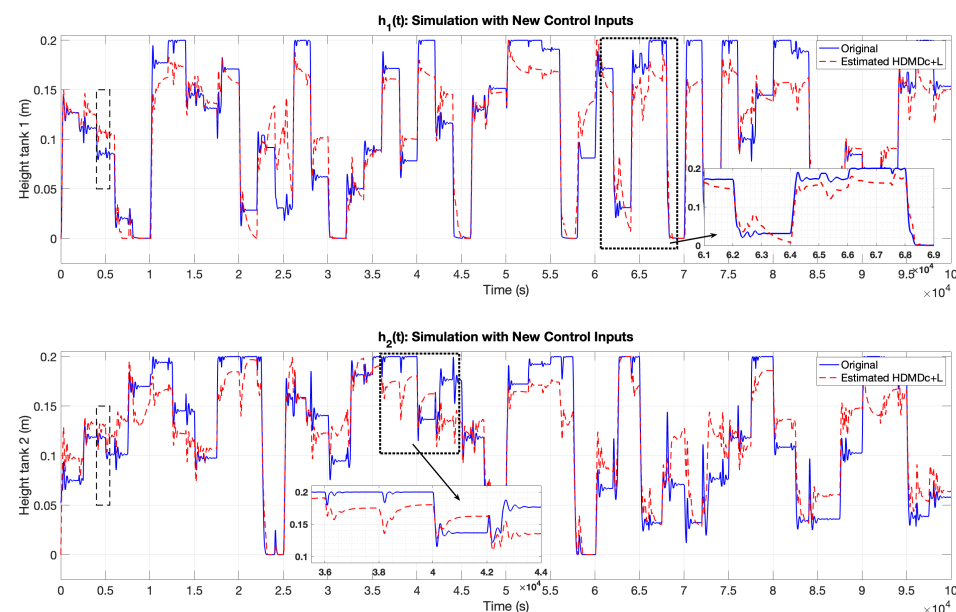


Figure 2. Performance of the HDMDc+Lift model in comparison with the validation set ($q = 5, r = 15$).

4.2. Selection of the Lifting Dictionary

The choice of the observable dictionary used for lifting is a critical design aspect of the HDMDc+Lift framework, as it directly determines the expressiveness and numerical properties of the resulting lifted linear model. An inadequate selection may lead to poor generalization, overfitting, or excessive computational cost, despite achieving acceptable short-term prediction accuracy.

To address this issue, a systematic evaluation was conducted using several candidate dictionaries belonging to different functional classes, including polynomial, radial basis, trigonometric, logarithmic-rational, truncated Fourier, and delay-based functions. For each dictionary, the identified model was assessed using a joint criterion that accounts for multi-step prediction accuracy, training time, and prediction time.

The comparative results, summarized in Table 2, indicate that the delay-based lifting functions provide the most favorable trade-off between predictive performance and computational efficiency; while some alternative dictionaries achieve comparable accuracy levels, they exhibit significantly higher training or prediction times, which limits their suitability for monitoring-oriented applications. Based on these results, the delay-based dictionary was selected for the remainder of this study, as it offers robust predictive performance while maintaining low computational complexity.

Table 2. Comparative results obtained for the best dictionary between predictive performance and computational efficiency.

Dictionary	H1train (Fit %)	H2train (Fit %)	H1predd (Fit %)	H2predd (Fit %)	Predd_Time (s)
Del-Bas Functions	0.877	0.869	0.878	0.862	0.1654
RBF	0.8068	0.773	0.814	0.727	0.347
Trigonometric	0.858	0.875	0.851	0.841	0.403
Logarithmic-Rational	0.683	0.709	0.662	0.625	0.168
Fourier (Trunc Functions)	0.85406	0.8493	0.8521	0.804	0.522

4.3. Hyperparameter Sensitivity and Selection

To address the sensitivity of the HDMDc+Lift model with respect to its hyperparameters and to justify their selection, a systematic optimization-based search was conducted over the Hankel embedding order q and the SVD truncation rank r . In particular, a Bayesian optimization strategy was employed to explore the two-dimensional hyperparameter space within predefined bounds, using a composite cost function based on multi-step prediction errors evaluated on independent validation data, as shown in Figure 3.

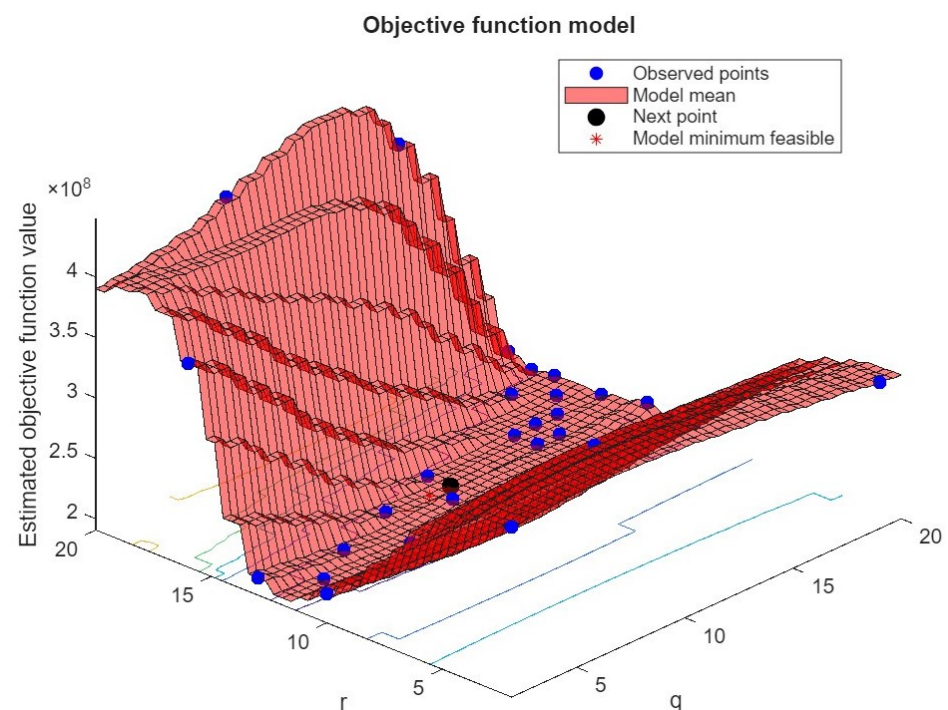


Figure 3. Optimal parameter selection based on the exploration of the (q, r) hyperparameter space.

The optimization results indicate that larger values of q and r , such as $q = 20$ and $r = 18$, can achieve prediction accuracies comparable to those obtained with the originally selected parameters ($q = 5, r = 15$), yielding coefficients of determination close to 87% for

both outputs. However, this increase in hyperparameter complexity does not result in a significant improvement in predictive performance.

In contrast, the computational cost associated with both model training and multi-step prediction increases substantially as the hyperparameters grow. Specifically, the average prediction time increases from approximately 0.87 s for $q = 5$, $r = 15$ to 4.90 s for the optimized configuration of $q = 20$, $r = 18$. At the same time, the dimensions of the identified state-space matrices **A** and **B** increase considerably, leading to higher memory requirements and reduced scalability in real-time or monitoring-oriented applications.

Based on these results, the parameters $q = 5$ and $r = 15$ were selected, as they provide a favorable trade-off between predictive accuracy, numerical robustness, and computational efficiency. These findings indicate that, although the HDMDc+Lift framework exhibits limited sensitivity to moderate variations in hyperparameters in terms of prediction accuracy, practical considerations related to computational cost and model size are decisive for parameter selection in real-world deployments.

4.4. Evaluation Under Non-Nominal Operating Conditions

To assess the robustness and generalization capability of the proposed HDMDc+Lift model beyond nominal operating conditions, its predictive performance was evaluated under multiple non-nominal scenarios. Three representative operating regimes were considered: stochastic disturbances acting on the actuators, a persistent actuator bias introducing a steady offset, and a change in the operating regime that places the system dynamics outside the training distribution.

For each scenario, independent validation datasets were generated, and the multi-step prediction performance was quantified using the coefficient of determination for both output variables. The resulting performance metrics are reported in Table 3. In addition, the corresponding time-domain predictions are shown in Figures 4–6, allowing a qualitative comparison of transient and steady-state behavior across the different operating regimes.

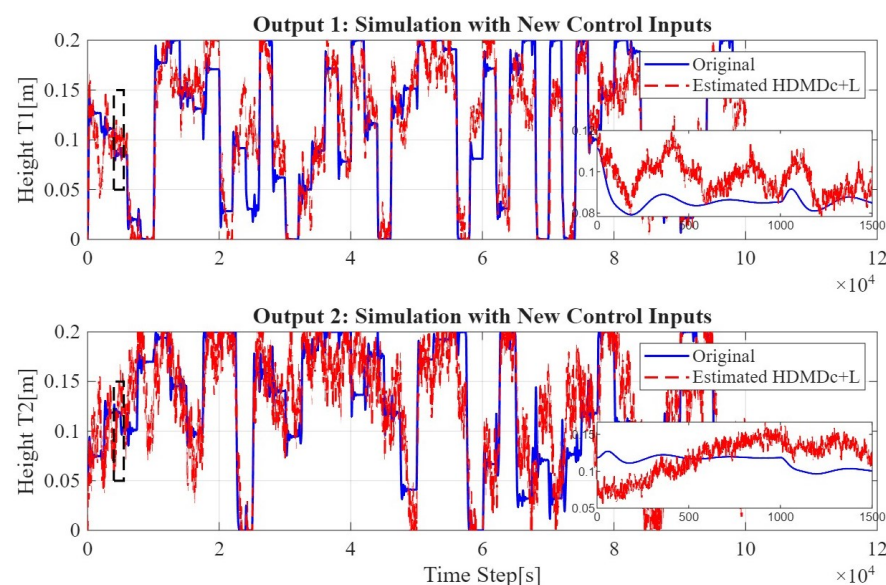


Figure 4. Performance of the HDMDc+Lift model with unknown entries with stochastic disturbance in actuators.

The results indicate that the HDMDc+Lift model preserves satisfactory predictive accuracy under stochastic disturbances and moderate regime changes, while its performance degrades in the presence of persistent actuator bias. Stochastic disturbances primarily introduce zero-mean fluctuations that are partially attenuated by the identified Koopman

subspace, resulting in only a moderate reduction in prediction accuracy. In contrast, a persistent actuator bias induces a systematic shift in the system trajectories that is not represented in the training data, leading to a structural mismatch in the learned linear operator and a pronounced collapse of the R^2 metric. Nevertheless, the model maintains stable and bounded predictions across all considered scenarios, highlighting its robustness and generalization capability despite structural deviations.

Table 3. Comparative results for HDMDc+Lift model under Non-Nominal Operating Conditions.

Fault	H1_pred (Fit %)	H2_pred (Fit %)
Stochastic disturbances in actuators	0.762	0.696
Actuator bias (persistent offset)	0.246	0.314
Change of operating regime	0.637	0.522

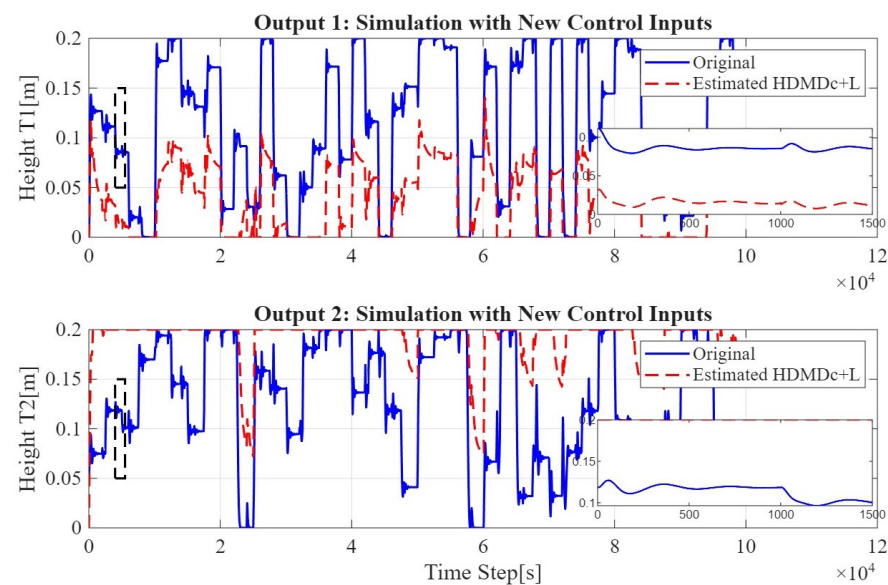


Figure 5. Performance of the HDMDc+Lift model with unknown entries with actuator bias (persistent offset).

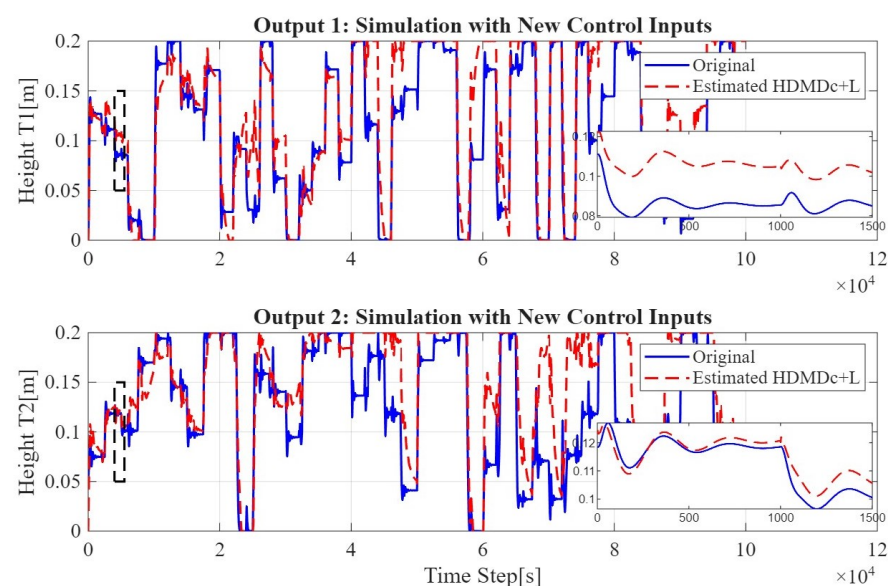


Figure 6. Performance of the HDMDc+Lift model with unknown entries with change of operating regime.

4.5. Predictive Performance and Uncertainty Quantification

In addition to pointwise prediction accuracy, the uncertainty associated with the proposed HDMDc+Lift model was quantified through confidence bounds derived from the training residuals. Specifically, the modeling error was computed as the difference between the measured outputs and the reconstructed outputs obtained during the identification phase. From these residuals, an empirical standard deviation was estimated for each output channel.

Assuming a Gaussian distribution of the modeling error, 95% confidence bounds were defined as $\pm 1.96\sigma$ around the predicted output trajectories. These bounds provide a statistical characterization of the uncertainty induced by truncation effects, finite data length, and modeling approximations inherent to the Koopman-based formulation. It is important to emphasize that the confidence bounds were computed exclusively from the training data and subsequently propagated to the validation phase, without recalibration during prediction.

Figures 7 and 8 illustrate the results for both the training and validation datasets, including the corresponding confidence bounds. As observed, the measured outputs remain largely contained within the estimated bounds. This behavior confirms that the identified HDMDc+Lift model captures the dominant system dynamics while providing meaningful error envelopes for prediction-based analysis and monitoring applications.

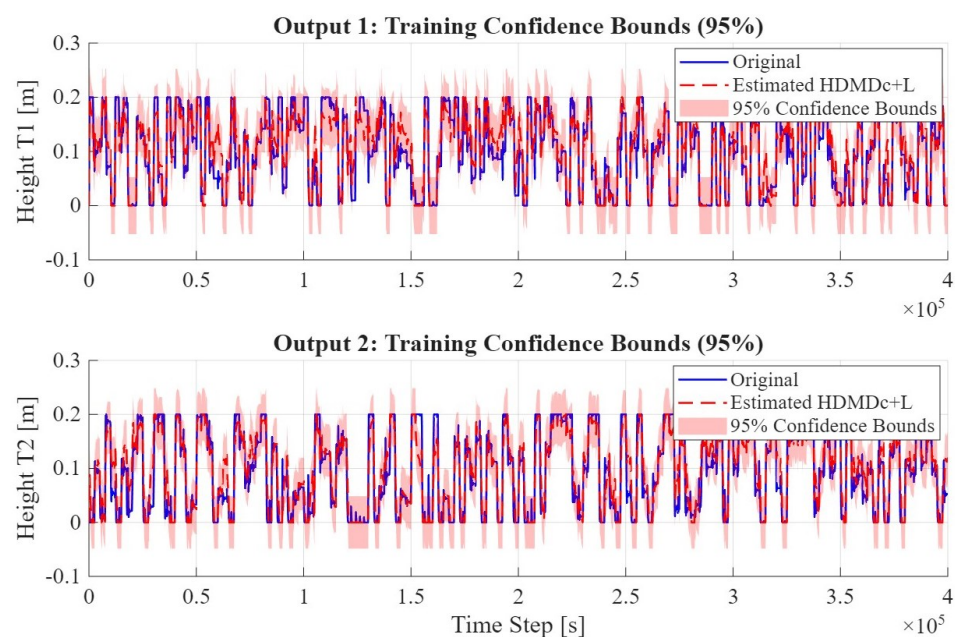


Figure 7. Performance of the HDMDc+Lift including 95% confidence bounds with the training data ($q = 5, r = 15$).

4.6. Comparison with the Classical Linearized Model, HAVOK and HDMDc

To contextualize the performance of the proposed model, a direct comparison was conducted with a classical linearized model derived from the physical equations of the multitank system around a nominal operating point, which we consider the worst possible performance of an identified model.

The results show that the HDMDc+Lift model consistently outperforms the linearized model, particularly during:

- Fast transients;
- Changes in operating conditions;
- Regions far from the linearization point.

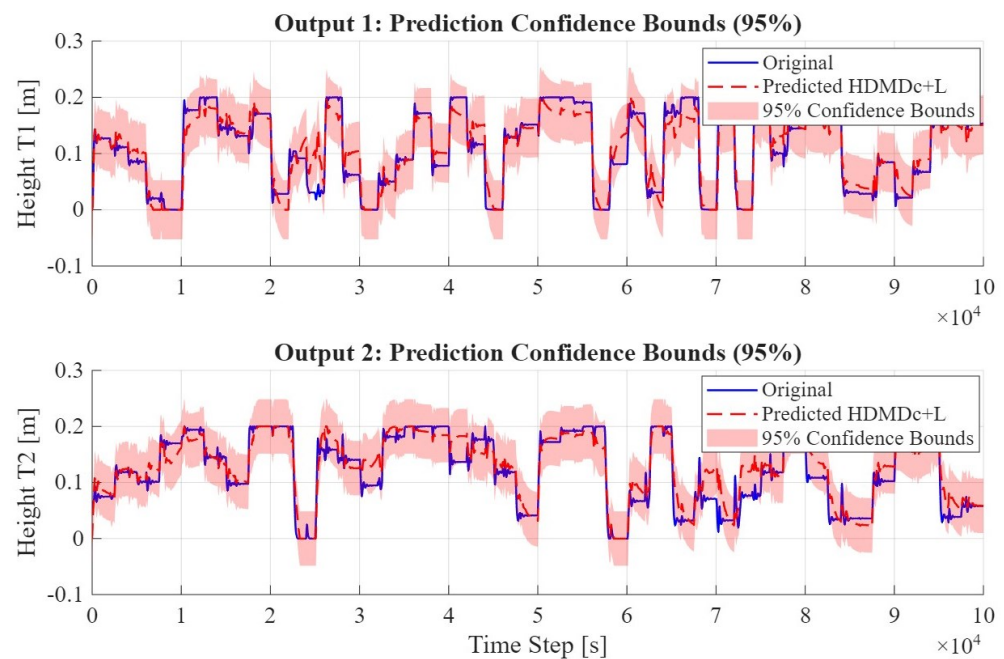


Figure 8. Performance of the HDMDc+Lift including 95% confidence bounds with the validation set ($q = 5, r = 15$).

While the linearized model exhibits significant errors under non-small variations, the data-driven model maintains a coherent response, effectively capturing dynamic coupling effects between variables.

This comparison highlights a key advantage of the proposed approach: the ability to represent complex nonlinear dynamics using an extended-order linear model. Tables 4 and 5 present a statistical analysis of prediction errors for both models.

Table 4. Statistical analysis of prediction error for level h_1 for the HDMDc+Lift model versus the linearized model.

Statistic	HDMDc+Lift	Linearized Model
Mean	0.0021	0.0879
Standard deviation	0.0241	0.0581
Skewness	0.4134	−0.1460
Kurtosis	32.5780	2.0672
MSE	0.0006	0.0111

To assess the relative predictive performance of the proposed HDMDc+Lift framework, a comparative analysis was conducted against HAVOK and standard HDMDc using the same validation dataset and evaluation metrics. The results are summarized in Table 6.

Table 5. Statistical analysis of prediction error for level h_2 for the HDMDc+Lift model versus the linearized model.

Statistic	HDMDc+Lift	Linearized Model
Mean	−0.0053	0.0897
Standard deviation	0.0225	0.0714
Skewness	0.5224	0.2000
Kurtosis	42.8972	1.7047
MSE	0.0005	0.0131

As shown in the table, the HAVOK model yields the lowest prediction accuracy for both outputs. This behavior is consistent with its linear representation of delay-embedded dynamics, which limits its ability to capture nonlinear input–output relationships, particularly in forced MIMO systems.

The HDMDc approach improves prediction performance by explicitly incorporating control inputs within a Hankel-embedded formulation. However, since the model operates in the original measurement space without nonlinear lifting, its representational capacity remains constrained.

In contrast, the proposed HDMDc+Lift model achieves the highest prediction accuracy for both outputs, with coefficients of determination exceeding 0.87.

The corresponding time-domain predictions, shown in Figure 9, visually confirm these quantitative results, highlighting the superior tracking capability and reduced prediction error of the lifted formulation over the entire prediction horizon.

Table 6. Prediction analysis for Classical Linearized Model, HAVOK, HDMDc and HDMDc+L.

Model	H1_Pred (Fit %)	H2_Pred (Fit %)
Classical Linearized Model	NC	NC
HAVOK	0.358	0.43
HDMDc	0.571	0.703
HDMDc+Lift	0.878	0.869

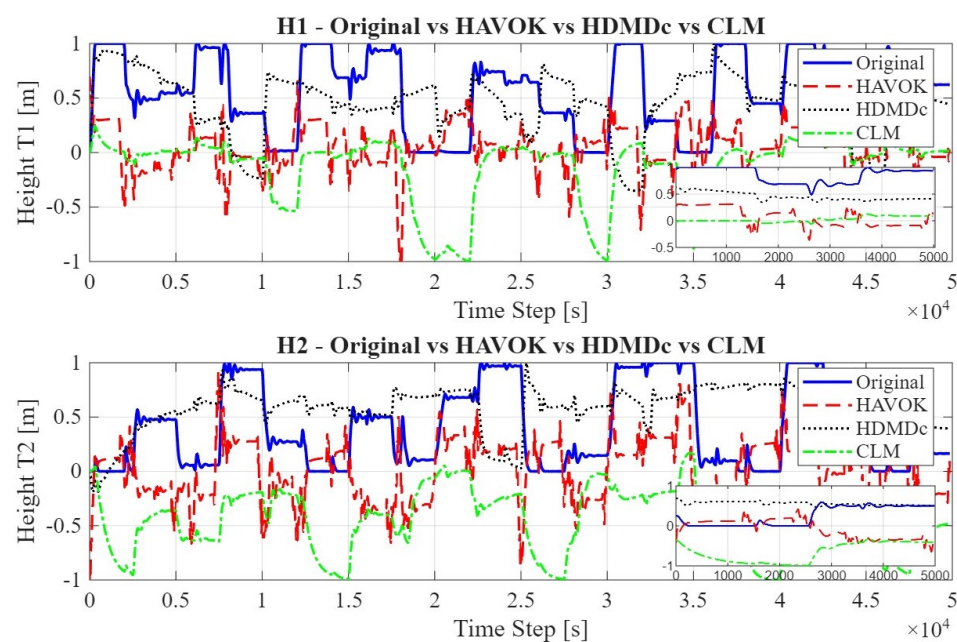


Figure 9. Comparison of performance between Classical Linearized Model, HAVOK, HDMDc and HDMDc+L with the training data ($q = 5, r = 15$).

4.7. Multi-Step Ahead Prediction Performance

To further evaluate the predictive robustness of the proposed HDMDc+Lift model, a multi-step ahead prediction analysis was conducted under a constant input assumption. The prediction performance was quantified using the mean squared error (MSE) for increasing prediction horizons, as reported in Table 7.

As expected, the prediction error gradually increases with the length of the prediction horizon due to the accumulation of modeling inaccuracies and truncation effects inherent to data-driven approximations. Nevertheless, the error growth remains bounded and does

not exhibit divergence, indicating that the identified lifted Koopman model preserves the dominant dynamics and stability properties of the multitanque system.

The results demonstrate that the HDMDc+Lift formulation maintains reliable predictive performance even for long prediction horizons, which is a desirable property for monitoring and forecasting applications. This behavior highlights the ability of the proposed approach to capture the essential system dynamics while limiting long-term error propagation.

Table 7. Multi-step ahead prediction performance of the HDMDc+Lift model using Fit % and MSE metrics.

Prediction Horizon	H1_Pred (Fit %)	H2_Pred (Fit %)	H1 MSE	H2 MSE
1	0.868	0.843	7.43×10^{-3}	9.40×10^{-3}
5	0.861	0.839	7.53×10^{-3}	9.59×10^{-3}
10	0.853	0.826	7.61×10^{-3}	9.72×10^{-3}
50	0.823	0.815	8.30×10^{-3}	10.59×10^{-3}
100	0.812	0.809	9.34×10^{-3}	11.86×10^{-3}

4.8. Modal Analysis and Stability of the Identified Model

Beyond numerical accuracy, it is essential to verify that the HDMDc+Lift model preserves the dynamic structure of the original system. To this end, a pole analysis was conducted for both the linearized model and the data-driven model.

In control theory, system poles correspond to the eigenvalues of the system matrix **A** and are directly associated with stability, response speed, and transient behavior. Therefore, if the HDMDc+Lift model provides a structurally accurate approximation, it should reproduce at least the dominant poles of the linearized model.

Unlike the linearized model, which has a fixed number of poles determined by the physical system order, the HDMDc+Lift model can capture additional poles corresponding to nonlinear, coupled, or resonant dynamics not present in the classical formulation.

It is worth noting that, even for larger Hankel embedding orders (up to $q = 20$), no eigenvalues were observed outside the unit circle. The poles located closest to the unit circle correspond to the dominant dynamical modes of the system, which are consistently preserved across different values of q . As the embedding order increases, additional poles appear; however, these modes are additional dynamic modes with weak or hidden dynamics and remain clustered near the origin of the complex plane. Consequently, their contribution to the overall system response is negligible, and the truncated SVD effectively preserves stability while retaining the dominant spectral content.

Figure 10 compares the distances between the poles of the linearized model and those of the HDMDc+Lift model. Figure 11 shows the complete set of poles estimated by the HDMDc+Lift model, highlighting the presence of additional dynamic modes.

Based on the spectral analysis presented in Figure 10, it is observed that the HDMDc+Lift model not only achieves superior temporal fitting compared to the linearized model, but also accurately captures the essential poles of the real system. The Euclidean distance between the dominant poles of both models is minimal, as shown in Table 8, indicating that the data-driven model reproduces the fundamental system dynamics with high fidelity.

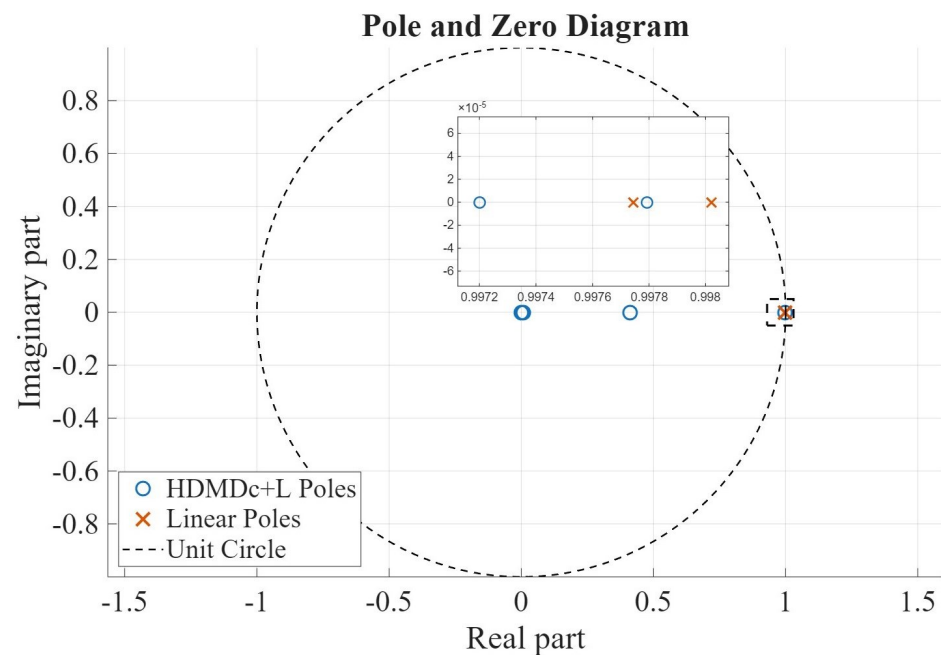


Figure 10. Pole comparison: linearized model versus HDMDc+Lift.

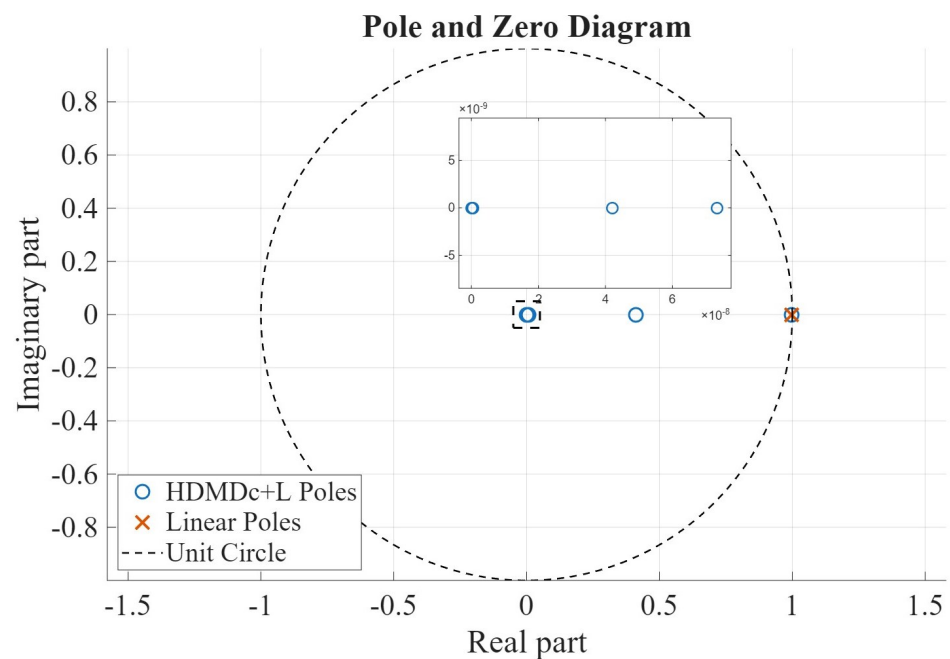


Figure 11. Complete pole distribution estimated by the HDMDc+Lift model.

Table 8. Distance analysis of dominant poles for the HDMDc+Lift model versus the linearized model.

Model	First Pole	Second Pole
HDMDc+Lift	0.9985	0.9985
Linearized model	0.9980	0.9977
Pole distance	0.00069944	0.0009064

4.9. Implications for Monitoring and Diagnosis

Although the primary focus of this work is on data-driven modeling, the proposed framework has important and immediate implications for system monitoring and fault diagnosis. The high fidelity of the identified model, together with its demonstrated structural stability, makes it well suited for integration into state estimation, observer-based monitoring, and residual generation schemes.

In particular, the linear representation in the lifted space enables the direct application of well-established diagnostic tools, such as Luenberger observers, Kalman filters, and model-based residual evaluators, while preserving the ability to capture nonlinear system dynamics. This constitutes a key advantage over purely black-box models, which often lack interpretability and robustness when deployed in diagnostic settings.

Moreover, the model's ability to generalize beyond the training dataset is a critical requirement for reliable fault detection and isolation (FDI) in real-world systems. In practical operation, faults and abnormal conditions typically manifest as deviations from nominal behavior under operating regimes that may not have been explicitly observed during identification. The observed generalization capability suggests that the proposed approach can distinguish between normal variability and abnormal behavior, thereby reducing false alarms and improving diagnostic sensitivity.

5. Conclusions

This work presented a fully data-driven modeling framework for forced nonlinear multiple-input multiple-output (MIMO) dynamical systems, grounded in finite-dimensional approximations of the Koopman operator through Hankel time-delay embeddings, explicit incorporation of control inputs, and observable lifting. The proposed HDMDc+Lift methodology enables the identification of an extended-order linear state-space representation directly from input–output data, without requiring explicit knowledge of the underlying physical equations.

The effectiveness of the approach was demonstrated on a real multitank system, where the identified model exhibited accurate multi-step predictions when evaluated on independent validation data. These results indicate that the proposed framework captures both transient dynamics and nonlinear coupling effects while maintaining numerical stability and structural coherence. The preservation of dominant dynamical modes, as confirmed by spectral analysis, further supports the validity of the Koopman-based approximation using delay-coordinate observables.

In addition to its modeling accuracy, the proposed method exhibits favorable computational performance, making it suitable for practical deployment in industrial monitoring and analysis settings. Overall, the results establish HDMDc+Lift as a flexible and effective data-driven modeling tool for forced nonlinear systems. Future work will focus on extending the framework toward state estimation, fault detection, and health monitoring applications, leveraging the identified linear representations within advanced control and diagnostic architectures.

Author Contributions: Conceptualization, L.I.M.; Methodology, M.V.-E., J.F.D.-S. and L.I.M.; Software, M.V.-E. and J.F.D.-S.; Validation, M.V.-E. and L.I.M.; Formal analysis, J.F.D.-S. and L.I.M.; Investigation, M.V.-E.; Resources, L.I.M.; Data curation, J.F.D.-S.; Writing—original draft, J.F.D.-S.; Writing—review & editing, M.V.-E. and L.I.M.; Supervision, L.I.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the “Vicerrectorado de Investigación e Innovación” at the University of Cuenca (VIUC) which provided financial backing for the project titled “Control based on the Koopman Operator applied to nonlinear MIMO Systems”. The APC was also funded by the VIUC.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

References

1. Ljung, L. *System Identification: Theory for the User*, 2nd ed.; Prentice Hall: Toronto, ON, Canada, 1999.
2. Sontag, E.D. *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, 2nd ed.; Springer: New York, NY, USA, 1998.
3. Khalil, H.K. *Nonlinear Systems*, 3rd ed.; Prentice Hall: Toronto, ON, Canada, 2002.
4. Brunton, S.L.; Proctor, J.L.; Kutz, J.N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 3932–3937. [[CrossRef](#)] [[PubMed](#)]
5. Verhaegen, M.; Verdult, V. *Filtering and System Identification: A Least Squares Approach*; Cambridge University Press: Cambridge, UK, 2007.
6. Billings, S.A. *Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains*; Wiley: New York, NY, USA, 2013.
7. Katayama, T. *Subspace Methods for System Identification*; Springer: London, UK, 2005.
8. Brunton, S.L.; Noack, B.R.; Koumoutsakos, P. Machine learning for fluid mechanics. *Annu. Rev. Fluid Mech.* **2019**, *52*, 477–508. [[CrossRef](#)]
9. Koopman, B.O. Hamiltonian systems and transformations in Hilbert space. *Proc. Natl. Acad. Sci. USA* **1931**, *17*, 315–318. [[CrossRef](#)] [[PubMed](#)]
10. Mezić, I. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dyn.* **2005**, *41*, 309–325. [[CrossRef](#)]
11. Schmid, P.J. Dynamic Mode Decomposition of Numerical and Experimental Data. *J. Fluid Mech.* **2009**, *656*, 5–28. [[CrossRef](#)]
12. Durán-Sigüenza, J.F.; Minchala, L.I.; Garza-Castañón, L.E.; Zhang, H. Control based on the Koopman operator: A comprehensive review. *J. Frankl. Inst.* **2025**, *362*, 108256. [[CrossRef](#)]
13. Proctor, J.L.; Brunton, S.L.; Kutz, J.N. Dynamic mode decomposition with control. *SIAM J. Appl. Dyn. Syst.* **2016**, *15*, 142–161. [[CrossRef](#)]
14. Brunton, S.L.; Budišić, M.; Kaiser, E.; Kutz, J.N. Modern Koopman Theory for Dynamical Systems. *arXiv* **2021**, arXiv:2102.12086. [[CrossRef](#)]
15. Williams, M.O.; Kevrekidis, I.G.; Rowley, C.W. A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition. *J. Nonlinear Sci.* **2015**, *25*, 1307–1346. [[CrossRef](#)]
16. Klüs, S.; Schuster, I.; Muandet, K.; Fackler, F.; Schmidt, F. Data-driven model reduction and transfer operator approximation. *J. Nonlinear Sci.* **2018**, *28*, 985–1010. [[CrossRef](#)]
17. Korda, M.; Mezić, I. Linear predictors for nonlinear dynamical systems: Koopman operator meets model predictive control. *Automatica* **2018**, *93*, 149–160. [[CrossRef](#)]
18. Lu, Q.; Wu, X.; She, J.; Guo, F.; Yu, L. Disturbance Rejection for Systems with Uncertainties Based on Fixed-Time Equivalent-Input-Disturbance Approach. *EEE/CAA J. Autom. Sin.* **2024**, *11*, 2384–2395. [[CrossRef](#)]
19. Jin, J.; Zhao, L.; Chen, L.; Chen, W. A robust zeroing neural network and its applications to dynamic complex matrix equation solving and robotic manipulator trajectory tracking. *Front. Neurobot.* **2022**, *16*, 1065256. [[CrossRef](#)] [[PubMed](#)]
20. Yang, X.; Puig, V.; Wang, X.; Wang, S.; Sun, C.; Zhang, Y. Dynamic-High-Gain-Based Decentralized Optimal Fault-Tolerant Control for a Class of Interconnected Nonlinear Systems. *IEEE Trans. Autom. Control* **2025**, *70*, 5823–5835. [[CrossRef](#)]
21. Zhou, Z.; Wang, Y.; Liu, X.; Li, Z.; Wu, M.; Zhou, G. Hybrid of Neural Network and Physics-Based Estimator for Vehicle Longitudinal Dynamics Modeling Using Limited Driving Data. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 16735–16746. [[CrossRef](#)]
22. Minchala, L.I.; Peralta, J.; Mata-Quevedo, P.; Rojas, J. An Approach to Industrial Automation Based on Low-Cost Embedded Platforms and Open Software. *Appl. Sci.* **2020**, *10*, 4696. [[CrossRef](#)]
23. Guillén, G.F.G.; Sigüenza, J.F.D.; Esquivel, M.L.V.; Minchala, L.I. An experimental comparison of model-based fault detection and isolation techniques in a multi-tank system. In Proceedings of the 2024 IEEE Eighth Ecuador Technical Chapters Meeting (ETCM), Cuenca, Ecuador, 15–18 October 2024; pp. 1–6. [[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.