Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 59-6 (2025) 403–408

Application of pqEDMD for modeling open raceway ponds

Camilo Garcia-Tenorio * **José Luis Guzmán** **
Laurent Dewasme * **Alain Vande Wouwer** *

* *Systems, Estimation, Control, and Optimization (SECO), University of Mons, 7000, Mons, Belgium (e-mail: camilo.garciatenorio, laurent.dewasme, alain.vandewouwer@umons.ac.be)*
** *Department of Informatics, ceiA3, CIESOL, University of Almeria, 04120, Almeria, Spain (e-mail: joseluis.guzman@ual.es)*

Abstract: Extended Dynamic Mode Decomposition (EDMD) has received increasing attention in the last decade, but neural networks remain the most popular approach to the data-driven representation of biochemical processes in the published literature. This study explores the potential of pqEDMD—a variant of EDMD using a reduced set of orthogonal polynomials—to approximate the dynamics of a complex system, i.e., a raceway pond for the biological treatment of wastewater and the production of algal biomass. We carefully discuss the main ingredients of the method, and illustrate the performance of the method with numerical results, showing promising prospects.

Keywords: Mathematical modeling, extended dynamic mode decomposition, machine learning, raceway ponds, wastewater treatment

1. INTRODUCTION

In recent years, the modeling of raceway ponds has gained increasing interest due to their pivotal role in producing microalgae, which have a wide array of industrial and environmental applications (Fernández et al., 2017). Raceway ponds are not only essential for wastewater treatment (Rayen et al., 2019) but also drive the production of diverse bioproducts such as animal feed and biofuels, making them a significant renewable energy source (Kiran et al., 2014; Moreno-Garcia et al., 2017). In addition, microalgae from these systems contribute to carbon capture, bioremediation, and even food additives, making them a cornerstone of sustainable industrial processes (Sayre, 2010). Given the growing demand for these resources, optimizing raceway pond performance through accurate predictive models is crucial for improving efficiency, maximizing biomass yield, and minimizing costs and environmental impact (Otálora et al., 2024).

To achieve such optimization, it is necessary to understand and predict the biological and physical processes that yield microalgae proliferation in raceway ponds. To reach this level of understanding, two broad categories of mathematical models are usually considered: mechanistic and data-driven. Mechanistic models rely on fundamental principles of physics and biology, providing a detailed representation of the system's components and their interactions. By explicitly modeling processes like biomass growth, dissolved oxygen dynamics, nutrient uptake, light penetration, and carbon dioxide exchange, mechanistic approaches offer deep insights into how variables such as light intensity, temperature, and mixing affect microalgae productivity (Banerjee and Ramaswamy, 2017). For example, mechanistic models can rely on a biological model that describes the growth of microalgae in the reactor, a dynamic mass balance for oxygen and total inorganic carbon, and a thermal model of the heat exchange mechanisms between the reactor and the environment (Fernández et al., 2017; Rodríguez-Miranda et al., 2021; Otálora et al., 2024). However, these models require precise parameter estimation, which can be challenging in dynamic environments.

In contrast, data-driven models like those produced by the Extended Dynamic Mode Decomposition (EDMD) and neural networks rely on experimental data to capture the system behavior without necessarily detailing the underlying physical laws. EDMD extends traditional Dynamic Mode Decomposition by incorporating nonlinear observables, allowing it to model more complex dynamics by mapping the system into a higher-dimensional function space (Williams et al., 2015). Neural networks, particularly deep learning models, excel at modeling nonlinear relationships between inputs and outputs (Park and Sandberg, 1991; Hanin, 2019). Both EDMD and neural network methods can operate on the same input-output relationships, using variables such as temperature, solar radiation, dilution rates, harvesting schedules, CO_2 injection and aeration rates to make an approximation of the same set of measurements, such as volume, biomass, dissolved oxygen, and pH. They learn from historical data to predict system performance under varying scenarios, offering flexibility and adaptability. Even though it is possible to achieve similar objectives with the two frameworks, the EDMD algorithm, along with all its variants, like the pqEDMD (p-q-quasi norm EDMD) (Garcia-Tenorio and Vande Wouwer, 2022), can provide better performance when there are limited computational resources and limited quantity of data.

10.1016/j.ifacol.2025.07.179

Additionally, the family of EDMD algorithms provides an evolution equation that operates on the set of observables that has a traditional linear model structure. Therefore, it is possible to translate linear analysis and control methods to the new framework (Budišić et al., 2012; Korda and Mezić, 2018). This article explores the feasibility of using EDMD as an alternative method to neural networks for modeling the dynamics of a raceway pond system. To this end, this study uses a simulation tool (Fernández et al., 2017; Nordio et al., 2024) that provides the necessary outputs according to real temperature and solar radiation data. By demonstrating how EDMD can effectively model key components of the process, such as biomass growth and dissolved oxygen dynamics, with fewer data and fewer parameters, we hope to foster the use of EDMD in this and other application areas as well as to propose useful raceway pond models that can be exploited for monitoring and control.

2. RACEWAY POND

The in silico representation of the raceway pond comes from a first principle analysis and parameter fitting of the real experimental setup from IFAPA research center, under the collaboration agreement with the University of Almería, Spain (Fernández et al., 2017; Nordio et al., 2024). The reactor has two 40-meter channels connected by two 180° bends at the ends. An electric motor drives a paddle wheel that circulates the liquid around the pond and into a sump. The reactor instrumentation includes three pH-T probes and three dissolved oxygen sensors positioned at the end of each channel, at the paddle-wheel, and at the sump. Additionally, the system automatically injects air or CO_2 gas through a diffuser at the sump bottom to control the dissolved oxygen and pH levels in the culture, respectively.

The Simulink® model (Fernández et al., 2017; Rodríguez-Miranda et al., 2021; Nordio et al., 2024) relies on mass balances, transport phenomena, thermodynamic relationships, and biological kinetics occurring within the reactor. Therefore, it provides a complete dynamic simulation model. Using the model, we can predict the evolution of the system's main variables—including biomass concentration, pH, dissolved oxygen, and total inorganic carbon in the liquid phase in addition to oxygen and carbon dioxide exchange in the gas phase. The identification and validation of the model parameters come from an experimental dataset of the 80 m^2 pilot-scale plant, providing a tool to determine the influence of design parameters on system performance or design controllers of critical variables like the pH (Banerjee and Ramaswamy, 2017; Nordio et al., 2024). For this study, the simulation tool will provide synthetic experimental data to feed into the pqEDMD algorithm and determine the feasibility of this algorithm for data-driven modeling of the raceway pond.

Figure 1 shows the block diagram of the simulation tool in Simulink®. The rightmost block contains the raceway dynamics that is unknown to the pqEDMD algorithm. The inputs of the system are solar radiation, water temperature, dilution rate (inflow), harvesting flow rate (outflow), and air and CO_2 flow rates. The outputs are the volume of water, biomass concentration, dissolved oxygen concentration, and pH. From the set of outputs, only the last three are relevant for modeling because the volume depends on the dilution and harvesting rates in a trivial way. Especially, the regular operation of the reactor uses the same dilution and harvest during the day, making the volume *constant* in between cycles. The blocks to the left are the on-off feedback controllers for the air and CO_2 flow rates. These variables are in a closed loop, which is required to maintain favorable conditions in the pond.

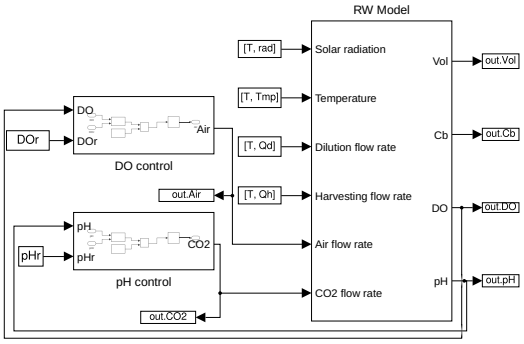

Fig. 1. Simulink® model of the raceway pond.

Air injection serves two purposes: mixing and oxygen removal. Air bubbles help to circulate the algae, ensuring that sufficient amounts of light and nutrients reach the whole population. During photosynthesis, the algae population produces oxygen, and without mixing, oxygen accumulates in the pond. High oxygen levels are detrimental, leading to oxidative stress for the algae and creating harmful conditions for biomass growth.

Injecting CO_2 has two competing purposes. The algae require carbon dioxide to perform photosynthesis, required for biomass growth, and CO_2 injection satisfies this carbon requirement. The problem of having an excess of CO_2 injection is the drop in pH levels that triggers growth inhibition, and under very low conditions, it can harm cells. Therefore, it is essential to maintain an optimal CO_2 injection to balance pH. This optimal flow corresponds to the matching of injection and consumption. For the experimental 80 m^2 raceway pond, the matching occurs at a pH level around 8 (Banerjee et al., 2024). Even though that is the optimal operation of the raceway, the objective of this work is to have reliable data-driven models for the subsequent analysis and control of the system. Therefore, we will use feedback methods that guarantee an acceptable level of dissolved oxygen and pH while driving the variables around a sufficiently large portion of the state space for an accurate identification of the nonlinear dynamics.

2.1 Simulation Setup

It is important, when identifying a system, to ensure that there is a dataset of sufficient size and information. For the data-driven identification of the raceway pond, the available data is a set of six nonconsecutive days of solar radiation and water temperature. We provide the pqEDMD with four days of data for training and two days for testing. Figure 2 shows the first day of data, where the water temperature and solar radiation come from measured data from the 24^{th} until the 31^{st} of March 2023

at the raceway plant in Almería, Spain, and the remaining inputs come from the simulation results. The definition of the signals in the block diagram for the simulation in Figure 1 is: `rad` [W/m^2] for the solar radiation, `Tmp` [°C] for the water temperature, `Qd` [m^3/60s] for the dilution rate, and `Qh` [m^3/60s] for the dilution and harvesting rates. These last are set to 0.001 [m^3/60s] or 60 [L/h] and are active for one hour during the day: from 09:00 until 10:00 for the dilution, and from 11:00 until 12:00 for the harvesting. Consequently, these profiles effectively keep the volume constant during a 24-hour cycle. Even though this behavior is simulated, the experimental setup works the same way. Finally, the two remaining inputs are the air and CO_2 injection in [m^3/60s] that come from two on-off feedback control loops. In this set of inputs, solar radiation and temperature are environmental variables that are out of the system's control, and dilution and harvest rates are set as part of the standard system operation. Even if there is no manipulation of the variables other than the real data or the dilution/harvesting schedule, the solar radiation and the water temperature have a sufficiently large variation range to produce informative data for identification. For the remaining two input variables (air and CO_2), and arguably, the main drivers of the system, there is some room to establish experimental conditions that can give a rich enough spectrum of the system dynamics for identification while keeping the dissolved oxygen and pH levels at acceptable values. The objective of the model is to replicate the behavior of the biomass, DO, and pH during a 24-hour cycle for subsequent control and optimization of the process.

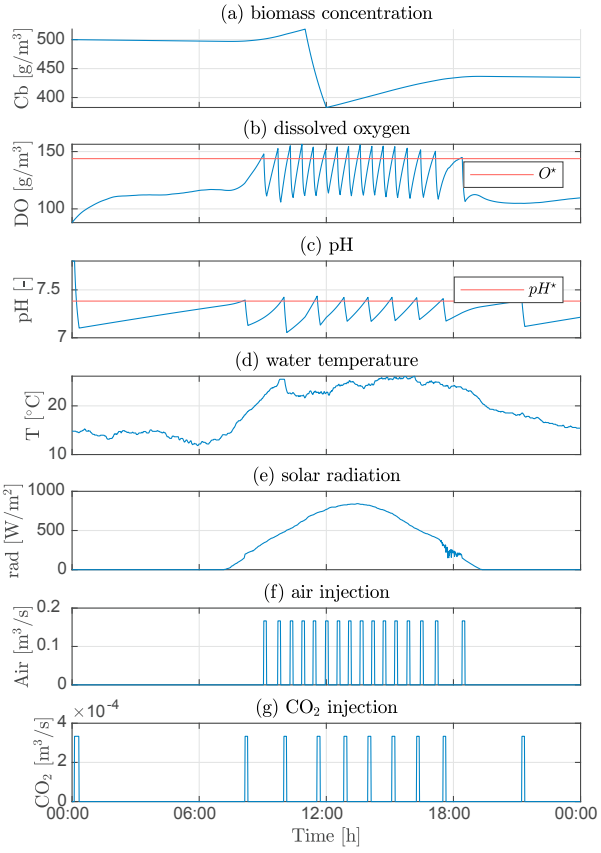

Fig. 2. First set of experimental data for system identification, March 26, 2023.

The implementation of the on-off controllers is the same for the two variables. For the air injection, when the dissolved oxygen is above the setpoint, the subsequent flow into the pond is 1/6 [m^3/60s]. For the CO_2 injection, when the pH is above the setpoint, the subsequent gas flow into the reactor is 1/3000 [m^3/60s]. These values come from empirical testing. Figure 2 shows the behavior of these controllers: sub-figures (b) and (c) show how the dissolved oxygen and the pH vary around their respective *setpoints*, $O^\star = 143.8$ [g/m^3], and $pH^\star = 7.38$ respectively, according to their respective inputs in sub-figures (f) and (g). In order to improve the information content of the dataset, the strategy is to set a different setpoint for all the various days of the experiment. The setpoints come from drawing a random number within a range from a uniform probability distribution, i.e., $O^\star \sim \mathcal{U}(130, 180)$, and $pH^\star \sim \mathcal{U}(7, 9)$.

The final detail regarding the simulation setup is the sampling time of the experiments. The original formulation of the simulator samples the outputs and updates the inputs every second. For the pqEDMD identification, having such fast sampling is detrimental to the solution, not only in terms of computational burden but also in terms of the numerical conditioning of the least squares solution. After empirically testing different sampling frequencies, a 60-second sampling performs well for the algorithm.

3. MODELING

The extended dynamic mode decomposition from Williams et al. (2015) and its evolution to include forcing signals from Korda and Mezić (2018) offer a powerful alternative that can address some of the limitations of machine learning methods like neural networks. The algorithm provides a linear evolution rule for the set of observables with the same structure as a traditional discrete-time linear state space representation and is, therefore, interpretable. It is possible to calculate and analyze the eigenvalues of the transition matrix. Also, under certain circumstances, it is possible to get an approximation of the Koopman operator and analyze the eigenfunctions of the operator (Budišić et al., 2012). Another advantage of the family of EDMD algorithms is the amount of data to get accurate approximations, which is, in general, orders of magnitude less than machine learning methods such as neural networks Otálora et al. (2023). The last point of contention is the ability of the algorithm to extrapolate outside the training data distribution.

3.1 Problem Setup

Consider a discrete-time system with state variables $x(k) \in \mathbb{R}^n$, and forcing signal, or input vector $u(k) \in \mathbb{R}^m$, at a specific instance of time k,

$$x(k+1) = T(x(k), u(k)) \tag{1}$$

where $T\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is the differentiable vector-valued evolution map. The goal of pqEDMD is to approximate the nonlinear mapping in a higher-dimensional function space or feature space whose evolution is linear. Thus, the development gets linearity while sacrificing dimensionality. Defining the function space as a set of observable functions, i.e., a vector of functions of the state,

$$\Psi(x) = [\psi_1(x)\ \psi_2(x)\ \cdots\ \psi_d(x)]^T, \tag{2}$$

where each element comes from a family of orthogonal polynomials, and a selection rule based on p-q-quasi norms determines the inclusion or exclusion of a particular polynomial in the basis (more on that later), the linear evolution mapping on the set of observables takes the form

$$\Psi(x(k+1)) = A\Psi(x(k)) + Bu(k) \tag{3a}$$
$$x(k) = C\Psi(x(k)), \tag{3b}$$

where $A \in \mathbb{R}^{d\times d}$ is the *observables* transition matrix, a linear operator acting on the function space $\Psi(x)$, $B \in \mathbb{R}^{d\times m}$ is the input matrix, and $C \in \mathbb{R}^{n\times d}$ is the output matrix that brings an arbitrary value of the function space back to the state space. Equation (3) is a nonlinear evolution mapping on the state, but linear on the set of functions $\Psi(x)$. Hence, to evolve the state, it is sufficient to evaluate the state at time k with the observables, perform the evolution with A, B and $u(k)$, and finally, bring the value back to the state space with C.

3.2 Data Collection

To approximate the A, B and C matrices in (3), it is necessary to collect sequences of snapshots $\{(x_i^+, x_i^-, u_i)\}_{i=0}^{N_j}$ from experiments on a real system or, like in this case, data coming from a simulator, from the successive application of (1) from an initial condition $x(0) = x_0$ according to a known sequence of inputs $\{u(k)\}_{k=0}^{N_j}$, where N_j is the number of points per experiment and $N = \sum_j N_j$ is the total number of points, and with $x_i^+ = T(x_i^-, u_i)$, so that the following data matrices can be formed

$$X^+ = \begin{bmatrix} x_1^+ & \dots & x_N^+ \end{bmatrix},\ X^- = \begin{bmatrix} x_1^- & \dots & x_N^- \end{bmatrix},\ U = \begin{bmatrix} u_1 & \dots & u_N \end{bmatrix}. \tag{4}$$

Notice that X^+ is just one step ahead of X^-. If there are several sequences from different initial conditions, it is possible to put them alongside in the data matrices (4) and perform the same calculation. For nonlinear systems with multiple asymptotically stable equilibrium points, it is necessary to have multiple samples that converge to different attractors.

3.3 Observables

There are many alternatives to select the observables, e.g., radial basis functions, the set of monomials, trigonometric functions, or an arbitrary set of functions (Williams et al., 2015; Garcia-Tenorio et al., 2021; Brunton et al., 2016). Arguably, the best performance comes from using a family of orthogonal polynomials under a p-q-quasi norm reduction method because it improves the numerical stability of the solution (Garcia-Tenorio and Vande Wouwer, 2022). A vast panel of orthogonal polynomials exists under different inner products. An important property is that the product of two polynomials from the same family is still orthogonal. Hence, each element of the observables set is the product of n univariate polynomials on each of the original states of the system,

$$\psi(x) = \prod_{j=1}^{n} \pi^{\alpha_j}(x_j), \tag{5}$$

where $\pi^{\alpha_j}(x_j)$ is a polynomial of degree α_j. Thus, the only necessary information to define an element of the set is the tuple of orders or degrees of the polynomial for each variable, $\alpha = (\alpha_j)_{j=1}^n$. If there is no restriction on the maximum order of the observables, a set that contains all possible combinations of α_j up to order p produces a large basis, with some high-order terms, possibly affecting the numerical stability and the overall error of the approximation. To prevent this situation, pqEDMD applies a p-q-quasi norm reduction to the set and only retains the elements whose q-quasi norm is less than p. In other words, the α vector that defines a multivariate polynomial $\Psi(x)$ satisfies

$$\alpha = \left\{ \alpha \in \mathbb{N}_+^n \colon \|\alpha\|_q \le p \right\}, \tag{6}$$

where $q \in \mathbb{R}_+$, and $\|\alpha\|_q = (\sum_{j=1}^n \alpha_j^q)^{1/q}$ is the q-quasi norm of the set of orders, and $p \in \mathbb{N}$ is a positive integer that determines the maximum order of the multivariate polynomial function $\psi(x)$.

With the set of observables, the next step is to evaluate the snapshot data and calculate a solution to the two linear problems

$$\Psi(X^+) = [A\ B] \underbrace{\begin{bmatrix} \Psi(X^-) \\ U \end{bmatrix}}_{\mathcal{X}} + r_{A,B} \tag{7a}$$
$$X^- = C\Psi(X^-) + r_C, \tag{7b}$$

where $r_\bullet$ is the residual term to minimize in either equation and $\mathcal{X} \in \mathbb{R}^{(d+m)\times N}$ is the regression matrix of the first linear problem.

3.4 Solution

For the two linear systems of equations in (7), it is feasible to use a traditional least squares solution: based on the pseudo inverse of the right-hand-side of (7a) like in the original formulation by Williams et al. (2015), a regularized least squares solution like the SINDY algorithm by Brunton et al. (2016), or even a maximum likelihood approach for uncertain systems (Garcia-Tenorio and Wouwer, 2022), or any other optimization method that minimizes the residual term $r_{A,B}$.

Arguably, the best method to solve the first linear system of equations, in terms of the possible rank deficiency of the regression matrix, is with a singular value decomposition of it $\mathcal{X}^T = USV^T$, the calculation of the effective rank and the orthonormal transformation of the matrices. The effective rank or ϵ-rank is,

$$r_\epsilon = \min\{r \colon \sigma_r \le \epsilon N \sigma_1\} \tag{8}$$

where $\epsilon << 1$, an arbitrarily small number, e.g., `eps` in Matlab®, and $\{\sigma_i\}_{i=1}^{d+m}$ are the singular values of $\mathcal{X}^T$. The effective rank eliminates the zero singular values and the arbitrarily small singular values from the range of $\mathcal{X}$. With the effective rank, the solution is an orthonormal transformation of the matrices in the linear system of equations (7a). Notice that the first r_ϵ right singular vectors, i.e., $V_{r_\epsilon} = V_{:,1:r_\epsilon}$ are already an orthonormal transformation of $\mathcal{X}^T$, because $S_{r_\epsilon}^{-1} U_{r_\epsilon} \mathcal{X}^T = V_{r_\epsilon}^T$. Then by applying the same orthonormal transformation to $\Psi(Y)^T$, i.e., $S_{r_\epsilon}^{-1} U_{r_\epsilon}^T \Psi(Y)^T = D_{r_\epsilon}$, the solution of the linear system of equations is,

$$[A\ B]^T = V_{r_\epsilon} D_{r_\epsilon}, \tag{9}$$

where all the r_ϵ subscripts denote the appropriate slicing of the matrices.

The second linear system of equations to solve is (7b), but rather than using an additional least-squares solution, which will inherently introduce numerical error, we consider the first-order elements of the polynomial basis of observables. These elements have a unique functional inverse, which translates into matrix form to produce the matrix C that returns the state from the function space (Garcia-Tenorio and Vande Wouwer, 2022).

4. NUMERICAL RESULTS

4.1 Simulation Parameters

For quick reference, Table 1 summarizes the simulation parameters to produce the training and testing sets.

Table 1. Simulation parameters to generate the training and testing sets.

Symbol	Value	Units
Δt	60	[s]
T	Sensor data	[°C]
rad	Sensor data	[W/m$_2$]
Qh	0.001 @09h–10h	[m^3/$\Delta t \cdot s$]
Qd	0.001 @11h–12h	[m^3/$\Delta t \cdot s$]
pHr	$\sim \mathcal{U}(7, 9)$	[g/m^3]
DOr	$\sim \mathcal{U}(130, 180)$	[g/m^3]
Air fb-gain	1/6	[-]
CO_2 fb-gain	$1/(50 \cdot 60)$	[-]

In the six days of experiments, the second day provides only half a day of data (about 700 data points instead of 1400). The training set considers days one until four, and the testing set considers days five and six. In total, there are 5011 points for training, and 2880 for testing. In comparison with machine learning algorithms, this is a minimal set of points for the approximation.

An important remark regarding the data gathering is the on-off feedback controller loops. Independently of the data-driven strategy, a direct approach to system identification under feedback control is a strategy that disregards the presence of the controller and makes an approximation using input/output data (Ljung, 1998). Even though there is no consideration of the closed-loop data dependency, the on-off architecture of the controller, especially a "bad" implementation of it, works in favor of the algorithm. The oscillation of the dissolved oxygen and pH around their respective working points does not limit the range of system responses, there is limited noise amplification, and there are no controller induced dynamics.

4.2 Training and Testing

From the description of the pqEDMD algorithm in Section 3, there are three parameters to choose for the approximation: the family of orthogonal polynomials and the p-q pair of parameters. The relationship between the polynomial type and a particular system is still an open question, especially since there are some systems where all the polynomials produce the same result. There is some work in the related field of sparse identification and algorithms like SINDY (Brunton et al., 2016), which can provide some insight to make the best choice. For this work, the approximation of the system is not independent of the polynomial family, and the choice comes from testing the different families and comparing the resulting empirical error for the same p-q parameter sweep. The definition of the error is

$$\epsilon = \frac{1}{N} \sum_{i=1}^{N} \frac{|\hat{y}_i - y_i|}{|y_i|}, \tag{10}$$

where $\hat{y}$ is the approximation of the output and y is the output from the dataset.

The results for a parameter sweep where $p = [2, 3, 4]$ and $q = [0.5, 1.0, 1.5]$ gives two alternatives to select the model. Table 2 highlights the best-performing p-q pair for the best-performing polynomial bases, and for comparison, it also shows the counterpart of the best p-q pair in the other polynomial, where the best-performing families are Laguerre and Legendre.

Table 2. pqEDMD parameters, number of observables and error for the best polynomials.

	Polynomial			
	Laguerre		Legendre	
p	2	3	2	3
q	0.5	0.5	5.0	0.5
d	9	13	9	13
ϵ	0.0194	0.0189	0.0194	NaN

Table 2 provides some interesting results. First, for two cases, the best-performing p-q pair is the most restrictive in terms of the q-quasi norm. Lower values of q give stronger truncations and eliminate more elements of the set of observables. $p = 3$, $q = 0.5$ for the Legendre polynomials produce an A matrix that is not Hurwitz, and the trajectories of the testing set diverge, showing the importance of testing several p-q pairs. For this example, the choice of having the first four days of data for the training set and the last two for testing is arbitrary, and a different combination could result in different approximations for the same set of parameters.

Notice that the empirical error for the same p-q parameterization, i.e., $p = 2$ and $q = 0.5$ has precisely the same error, $\epsilon = 0.0194$. This result is related to the open question regarding the dependence of the solution on the family of polynomials and the hypothesis that some families are better than others on a particular application. Notice also that the best-performing Laguerre polynomial is only slightly better than its lower-dimensional counterparts. Comparing these two solutions, there is a difference of 4 elements in the set of observables. Even though it is not much, the difference in the cardinality of two competing solutions may be the most important factor when selecting one or the other. For example, a smaller A matrix may have better chances of producing feasible solutions if the model is the basis of a model predictive control algorithm.

Figure 3 shows the approximation of the best solution on the two experiments of the testing set. The simulation starts at the same initial conditions as the experimental data, and the evolution comes from the iterative application of 3 according to the sequence of inputs of the experiments. This is, therefore, an open-loop approximation of the closed-loop dynamics of the raceway pond. These results show that the model is accurate enough to predict the system behavior, and it is a good candidate for testing on actual data from the reactor.

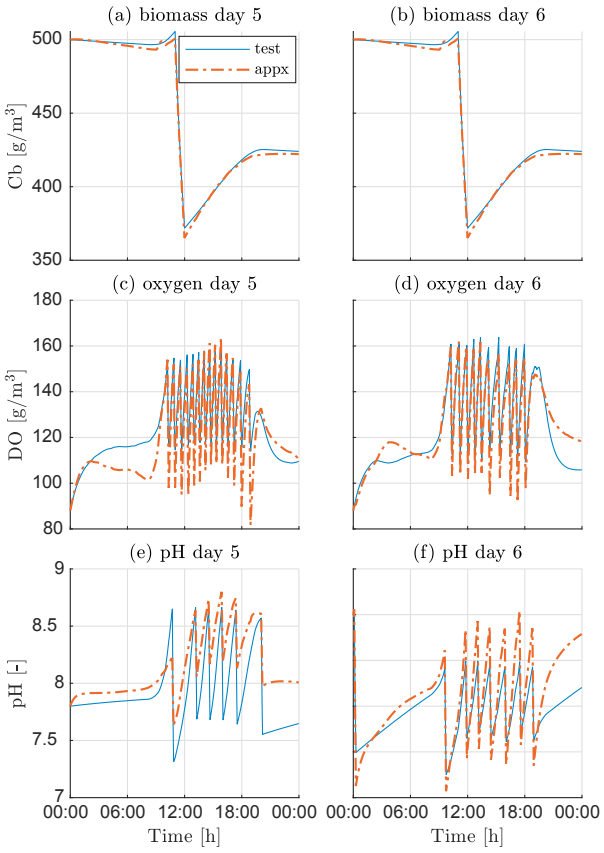

Fig. 3. pqEDMD approximation of the testing set of experimental data, march 30-31 2023.

5. CONCLUSIONS

This paper deals with a data-driven approximation of an in silico model of a raceway pond bioreactor using the pqEDMD algorithm. We discuss the procedure to derive accurate approximations and some aspects of the algorithm, like the p-q parameterization and the choice of polynomials for the set of observables. This study's source code and example data can be found in the `examples/Raceway` directory of the `garten-cam` GitHub repository.

REFERENCES

Banerjee, S., Otálora, P., El Mistiri, M., Khan, O., Guzmán, J.L., and Rivera, D.E. (2024). Control-relevant input signal design for integrating processes: Application to a microalgae raceway reactor. *IFAC-PapersOnLine*, 58(15), 360–365.

Banerjee, S. and Ramaswamy, S. (2017). Dynamic process model and economic analysis of microalgae cultivation in open raceway ponds. *Algal Research*, 26, 330–340.

Brunton, S.L., Proctor, J.L., and Kutz, J.N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15), 3932–3937.

Budišić, M., Mohr, R., and Mezić, I. (2012). Applied Koopmanisma). *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 22(4), 047510.

Fernández, I., Guzmán, J.L., Berenguel, M., and Acién, F.G. (2017). *Dynamic Modeling of Microalgal Production in Photobioreactors*, 49–87. Springer.

Garcia-Tenorio, C., Delansnay, G., Mojica-Nava, E., and Vande Wouwer, A. (2021). Trigonometric embeddings in polynomial extended mode decomposition—experimental application to an inverted pendulum. *Mathematics*, 9(10).

Garcia-Tenorio, C. and Vande Wouwer, A. (2022). A matlab toolbox for extended dynamic mode decomposition based on orthogonal polynomials and p-q quasi-norm order reduction. *Mathematics*, 10(20).

Garcia-Tenorio, C. and Wouwer, A.V. (2022). Maximum likelihood pqedmd identification. In *2022 26th International Conference on System Theory, Control and Computing (ICSTCC)*, 540–545.

Hanin, B. (2019). Universal function approximation by deep neural nets with bounded width and relu activations. *Mathematics*, 7(10).

Kiran, B., Kumar, R., and Deshmukh, D. (2014). Perspectives of microalgal biofuels as a renewable source of energy. *Energy Conversion and Management*, 88, 1228–1244.

Korda, M. and Mezić, I. (2018). linear predictors for nonlinear dynamical systems: koopman operator meets model predictive control. *Automatica*, 93, 149–160.

Ljung, L. (1998). *System Identification: Theory for the User*. Pearson Education.

Moreno-Garcia, L., Adjallé, K., Barnabé, S., and Raghavan, G. (2017). Microalgae biomass production for a biorefinery system: Recent advances and the way towards sustainability. *Renewable and Sustainable Energy Reviews*, 76, 493–506.

Nordio, R., Rodríguez-Miranda, E., Casagli, F., Sánchez-Zurano, A., Guzmán, J.L., and Acién, G. (2024). Abaco-2: a comprehensive model for microalgae-bacteria consortia validated outdoor at pilot-scale. *Water Research*, 248, 120837.

Otálora, P., Guzmán, J.L., Berenguel, M., and Acién, F.G. (2023). Data-driven ph model in raceway reactors for freshwater and wastewater cultures. *Mathematics*, 11(7).

Otálora, P., Skogestad, S., Guzmán, J.L., and Berenguel, M. (2024). Modeling, control and online optimization of microalgae-based biomass production in raceway reactors. *IFAC-PapersOnLine*, 58(14), 235–240.

Park, J. and Sandberg, I.W. (1991). Universal approximation using radial-basis-function networks. *Neural Computation*, 3(2), 246–257.

Rayen, F., Behnam, T., and Dominique, P. (2019). Optimization of a raceway pond system for wastewater treatment: a review. *Critical reviews in biotechnology*, 39(3), 422–435.

Rodríguez-Miranda, E., Acién, F.G., Guzmán, J.L., Berenguel, M., and Visioli, A. (2021). A new model to analyze the temperature effect on the microalgae performance at large scale raceway reactors. *Biotechnology and Bioengineering*, 118(2), 877–889.

Sayre, R. (2010). Microalgae: The Potential for Carbon Capture. *BioScience*, 60(9), 722–727.

Williams, M.O., Kevrekidis, I.G., and Rowley, C.W. (2015). A Data-Driven Approximation of the Koopman Operator: Extending Dynamic Mode Decomposition. *Journal of Nonlinear Science*, 25(6), 1307–1346.