Chemical and Process Engineering: New Frontiers, 2023, 44(3), e24
DOI: 10.24425/cpe.2023.146726

A journey from mechanistic to data-driven models in process engineering: dimensionality reduction, surrogate and hybrid approaches, and digital twins

Katarzyna Bizon*

Cracow University of Technology, Faculty of Chemical Engineering and Technology, Warszawska 24, 31-155 Kraków, Poland

*** Corresponding author:**
e-mail:
katarzyna.bizon@pk.edu.pl

Presented at 24th Polish Conference of Chemical and Process Engineering, 13–16 June 2023, Szczecin, Poland.

Article info:
Received: 12 May 2023
Revised: 28 June 2023
Accepted: 20 July 2023

Abstract
The study examines various approaches oriented towards conceptual and numerical reduction of first-principle models, data-driven methodologies for surrogate (black box) and hybrid (gray box) modeling, and addresses the prospect of using digital twins in chemical and process engineering. In the case of numerical reduction of mechanistic models, special attention is paid to methodologies in which simulation data are used to construct light but robust numerical models while preserving all the physics of the problem, yielding reduced-order data-driven but still white-box models. In addition to reviewing various methodologies and identifying their applications in chemical engineering, including industrial process engineering, as well as fundamental research, the study outlines associated problems and challenges, as well as the risks posed by the era of big data.

Keywords
reduced-order model, data-driven model, surrogate model, hybrid model, digital twin

1. INTRODUCTION

The design, optimization, and control of chemical and process engineering systems, as well as the prediction of their behavior in time or in the operating parameter space, requires the formulation and then numerical implementation of mathematical models (Aris, 1976; Aris, 1993; McLean and McAuley, 2012) featuring different levels of complexity. This refers to fundamental phenomena, unit operations, and complex industrial processes. The use of detailed mathematical models based on first principles that incorporate many constituent physical and chemical phenomena occurring at different scales, and additionally consider the actual dimension of the system, could be economically justified at the design stage of new industrial equipment or processes. High computational effort may be preferable in this case to time- and cost-consuming experiments that usually need to be performed at different scales. Nevertheless, the application of complex mathematical models for optimization or process control, especially when dealing with multi-scale and multi-physics problems or complex process installations, becomes impossible. Consequently, it is necessary to use models with lumped parameter descriptions involving a range of ideal simplifying assumptions (Biegler et al., 2014).

The computational burden associated with numerical simulations of detailed mechanistic models can be decreased by using a variety of model reduction techniques or by replacing the entire model or just some of its elements with a data-driven model/submodel (Chakrabarty et al., 2016; Hou and Behdinan, 2022; Wang and Shan, 2007) (Fig. 1a).

Model reduction methodologies can be generally divided into those based on conceptual approach and those relying on mathematical and numerical techniques (Chakrabarty et al., 2016), whereby numerical dimensionality reduction of the model can also involve techniques based on data exploration, while preserving the underlying physics. The alternative of mechanistic models and their somehow reduced, i.e., conceptually, or mathematically/numerically, versions are the so-called surrogate models known also as emulators or metamodels (Bradley et al., 2022; Wang and Shan, 2007). Surrogate models generally have the form of a data-driven black-box input-output model, in which the physics of the problem is included only indirectly, since their structure does not incorporate the governing equations (von Stosch et al., 2014). Their use therefore raises many controversies and concerns in the scientific community (von Stosch et al., 2014; Coveney et al., 2016; Bradley et al., 2022). A compromise in the contest between traditional mechanistic modeling and data-driven surrogate approaches is offered in the form of hybrid models (or so-called gray-box models), which incorporate both physics and submodels based on data derived from the real object, and in effect are considered more dependable (von Stosch et al., 2014).

The extremely rapid development of hardware and computing power in recent years has not only contributed to the explosion of various types of modeling methodologies (Fig. 1b), based to a great extent on machine learning techniques or, more generally, data mining, but has also brought about the birth of a novel versatile tool: the so-called digital twin (Bárkányi et al., 2021; Wright and Davidson, 2020). Unlike

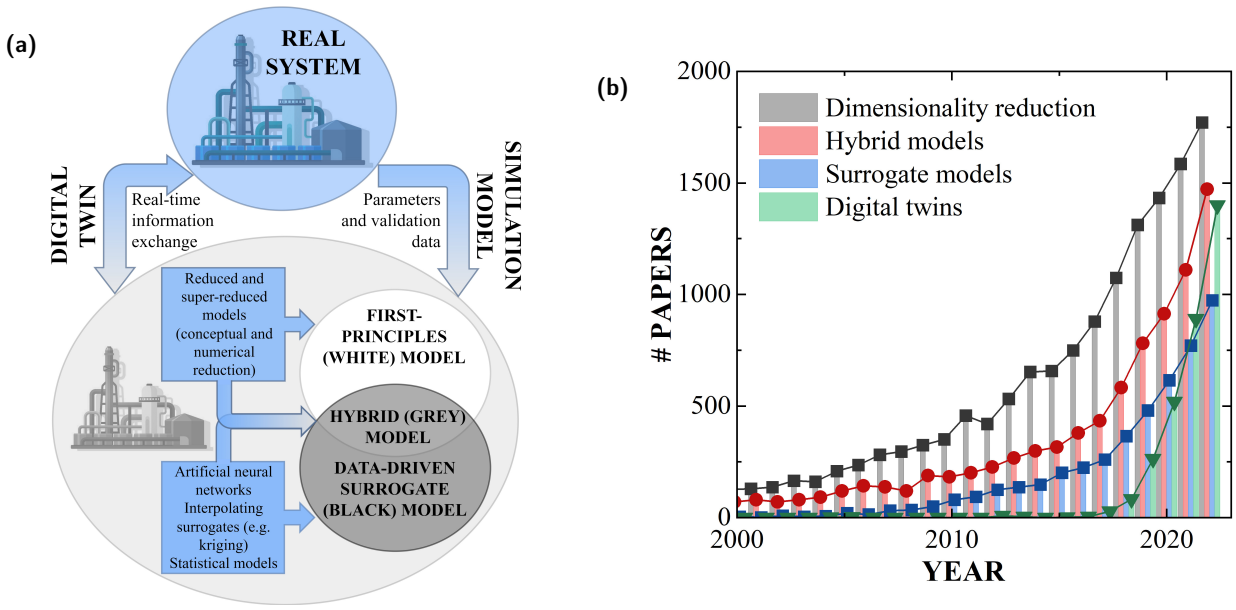

Figure 1. Illustration of the general classification of reduced and data-driven models (a) and the number of publications (in Scopus indexed journals on 5th May 2023) over the years, on dimensionality reduction, hybrid and surrogate models, and digital twins, addressing issues in chemical engineering, chemistry, material and environmental science, and energy (b).

the classical surrogate or hybrid model, a digital twin combines numerical models of different nature with their constant updating and real-time interaction with the simulated real-world asset (Fig. 1a).

The study examines various approaches to mechanistic model reduction, surrogacy, and hybridization, and additionally discusses the perspective of using digital twins in chemical and process engineering. Due to the extremely broad scope of the topic in the case of mathematics-based reduction of mechanistic models, special attention is paid to methodologies in which data of both simulation and experimental origin are used to construct light yet robust numerical models while preserving all the physics of the problem (white-box data-driven reduced-order or super-reduced models). In addition to reviewing various methodologies and applications in chemical engineering, including industrial process engineering (with particular emphasis on emerging technologies) as well as fundamental research, the related problems and challenges are outlined that are likely to boost future research.

2. DIMENSIONALITY REDUCTION – FROM CONCEPTUAL TO NUMERICAL APPROACHES

Back in the mid-70s, Rutherford Aris, a world-renowned mathematician by background, with a great academic record in the field of chemical engineering and a major influence on the development of the field in the second half of the 20th century, in an article published in the journal Chemical Engineering Education (Aris, 1976), came up with a list of principles for mathematical modeling. He termed these principles as the "maxims." Among the 13 maxims formulated by Aris, several of them deserve particular attention:

- "*Think geometrically. See when you can reduce the number of variables (. . .), but keep in mind the needs of the general case.*"
- "*Use crude approximations, e.g., 1-point collocation.*"
- "*Neglect small terms but distinguish between regular and singular perturbations.*"
- "*These maxims will self-destruct. Make your own!*" (Aris, 1976)

These more than half century old maxims are the quintessence of the conceptual and mathematical (numerical) reduction of mathematical models, with the final one that can be treated as a potential for all other alternatives. Such alternatives of today, for example, are data-driven (meta-)models. In this context, it is also worth quoting the words of another world-renowned chemical engineering specialist, Gilbert F. Froment, dating back to the same period (Froment, 1974): "*Owing to the increasing possibilities of computers the modelling of fixed bed catalytic reactors has been a rapidly developing field in the last decade. Models with more and more complexity have been set up and compared with more elementary ones (. . .). This has now paved the way to judicious model reduction (. . .). Finally, it is hoped that model studies and analysis of phenomena in reactors will be based on real data and real reactor configurations* (. . .)."

The conceptual methods of model reduction, that is, straightforwardly speaking, simplifying them both in terms of dimensionality (usually treating a real three-dimensional system as a one- or zero-dimensional one) and the phenom-

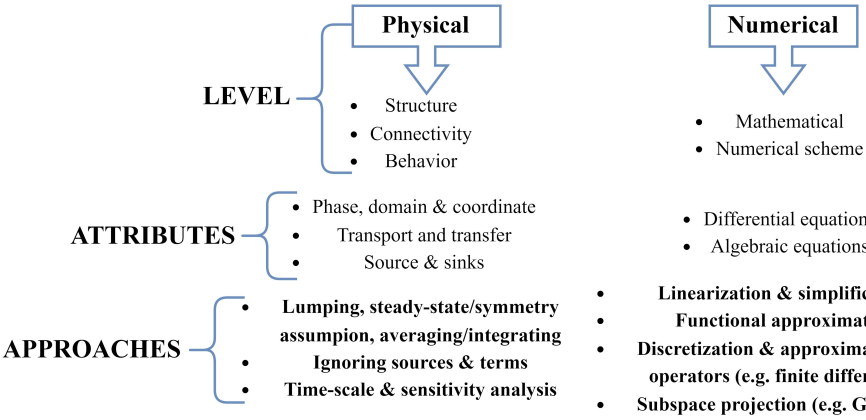

Figure 2. General classification of conventional model reduction approaches used in chemical engineering; based on Chakrabarty et al. (2016).

ena considered (Fig. 2), until about 50 years ago were the only of a very few practicable options. Other approaches included several types of mathematical methods, such as linearization, or numerical methods, like classical discretization methods based on finite differences or orthogonal collocation. The application of the latter, however, was largely limited to at most one-dimensional systems.

It should be emphasized that the classical and still widely used numerical approaches nowadays, such as the method of lines, finite difference, finite volume methods (Kurtz et al., 1978; Vinokur, 1989), collocation and spectral methods (Hesthaven et al., 2007; Villadsen and Michelsen, 1978), formally belong to the group of reduction methods (Bizon and Continillo, 2012). This is evident from the fact that mathematical models of distributed-parameter systems encountered in chemical engineering take typically the form of partial differential equations (PDEs). For practical computational purposes, such original infinite-dimensional PDEs must be reduced to finite-dimensional systems of algebraic or ordinary differential equations (ODEs). The aforementioned classical methods of discretization of governing equations lead to their transformation into a large system of algebraic equations or ODEs. Building efficient, in terms of computational time, and yet still faithful numerical models requires further reduction of such high-dimensional discretized models. Therefore, in the modern sense, the term model reduction, or more precisely, model-order reduction, is commonly used rather to describe the latter stage, and refers to a broad class of techniques for obtaining low-dimensional approximations (Swischuk et al., 2018).

Systematic model reduction based on conceptual, i.e., physical approach (Chakrabarty et al., 2016) involves simplifying the problem via lumping or averaging, ignoring/neglecting terms in the equations, and introducing (quasi-)steady-state or symmetry assumptions (Fig. 2). Lumping is particularly attractive for formulating and then solving models of catalytic fixed-bed reactors (Froment, 1974) as well as catalytic fluidized-bed reactors (Bizon, 2016; Bizon, 2021), in which fluid-solid processes are conducted. Two-scale reactor (fluid)-particle (solid) models are usually simplified by ignoring the intra-particle distribution of state variables. It should be, however, recognized that while this approach significantly reduces the complexity of the model, it may also lead to its oversimplification, which in turn can produce erroneous predictions of apparatus performance.

To visualize the problem, Fig. 3 shows the results of numerical simulations obtained from four different models of a catalytic bubbling fluidized-bed reactor (BFBR) described in detail in (Bizon, 2021). In all cases, ideal mixing and plug flow of the gas were assumed in the dense and bubbling phases of the bed, respectively. However, the models differ in the description of the mass and heat transport phenomena at the gas-solid interface and within the catalyst particle:

- Model P, pseudo-homogeneous model: mass and heat transport resistances between interstitial gas in the dense phase and the catalyst particles are neglected, the same applies to intra-particle transport resistances.
- Model H1, with distributed-parameter catalyst submodel: both external and intra-particle mass and heat transport resistances are accounted for.
- Model H2, with lumped-thermal catalyst submodel: similar to H1, but with uniform intra-particle temperature distribution.
- Model H3, with lumped-particle catalyst submodel: similar to H2, but additionally with uniform intra-particle concentration distribution.

The results shown in Fig. 3 represent the start-up stage of catalytic BFBR, presuming that a single irreversible exothermic reaction A $\rightarrow$ B takes place in the reactor. As can be clearly observed, the two simplest models employed (i.e., P and H3) fail completely to predict ignition. The erroneous prediction obtained from the P and H3 models is due to the existence of multiple steady states, which are typical of exothermic catalytic processes (Bizon, 2016; Sun et al.,

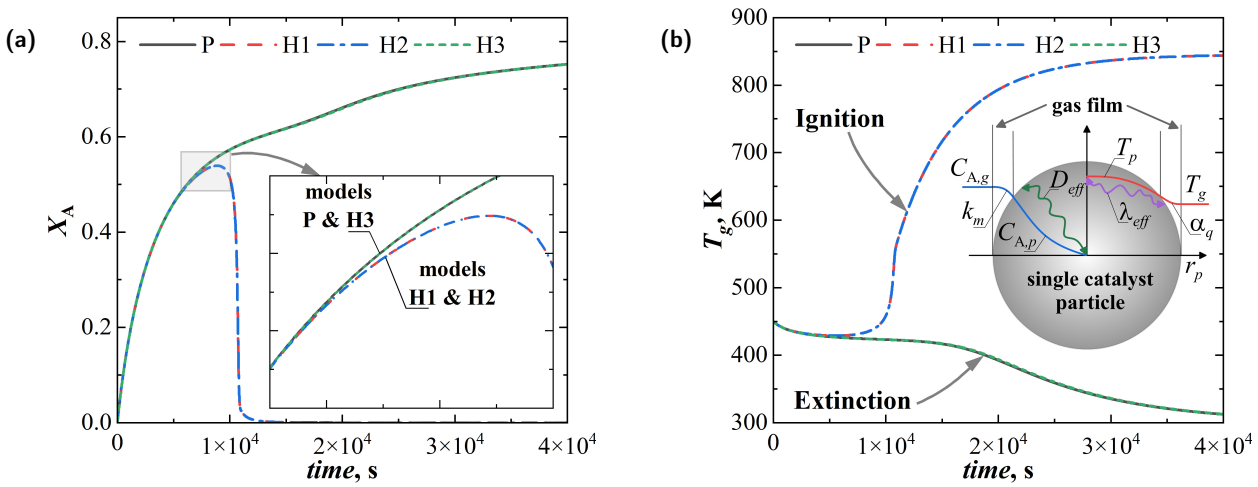

Figure 3. Evolution of dimensionless concentration of reactant A (a) and gas temperature in the dense phase of catalytic BFBR during reactor start-up. Results obtained for a single first-order irreversible catalytic exothermic reaction with $l_f = 5.06$, $k_0 = 5 \cdot 10^6$ 1/s and $\Delta H = -6 \cdot 10^5$ kJ/kmol using models of different complexity; based on Bizon (2021).

2019). This confirms the notion that the practical application of models simplified with the conceptual approach must be preceded by a detailed comparative analysis with more complex models, in order to determine the extent of their applicability (Bizon, 2021; Fernandes and Lona, 2002; Froment, 1974; Rout and Jakobsen, 2015).

Despite the well-established knowledge regarding the modeling and simulation of catalytic reactors, given the emerging nowadays technologies associated with the power-to-gas or more generally power-to-X concept (Oztruk and Dincer, 2021; Palys and Daoutidis, 2022), the development of more and more computationally efficient numerical models of such apparatuses is still of particular interest. The core of power-to-X concept is hydrogen production via electrolysis using surplus electric energy from renewable sources. Hydrogen, often referred as to "green," can then be utilized directly as a fuel or further converted, for example, into other fuels (Oztruk and Dincer, 2021), including methane, methanol, or dimethyl ether (DME) via hydrogenation of carbon dioxide. The latter can be obtained from a variety of sources, including facilities to capture carbon dioxide from industrial waste gases (Palys and Daoutidis, 2022). Such reutilization of carbon dioxide in a wide spectrum of power-to-X processes renders it no longer seen as waste, but as a valuable chemical compound (Styring et al., 2021). The aforementioned chemical processes are typically conducted in catalytic fixed-bed, moving-bed, or fluidized-bed reactors. In addition, very often, due to the increasing need for process integration and intensification (Moioli, 2022), they are conducted in special types of these apparatuses: multifunctional (e.g., direct synthesis of DME, also known as one-pot synthesis) or adsorptive reactors (e.g., sorption-enhanced methanation or methanol synthesis) (Dhoke et al., 2021; Kurzina et al., 2017). A characteristic feature of both solutions is their potential to integrate individual functionalities, i.e., two different catalysts in the case of direct DME synthesis or a catalyst and adsorbent in the case of sorption-assisted processes, both at the apparatus level and at the level of a single particle. In the first scenario, the reactor packing is composed of a physical mixture of different particles, while in the second scenario one deals with so-called multifunctional or hybrid particles, which may have a core-shell structure (Bizon et al., 2019). Thus, the distribution of diverse types of active centers can then be an additional design parameter (Sánchez-Contador et al., 2019; Bizon et al., 2019). As a result, when optimizing the distributions of active centers in the particle, it is not possible to ignore the intra-particle distributions of state variables. Another problem that arises in sorption-enhanced processes conducted in fixed-bed reactors independently on the mode of integration of catalyst and sorbent (i.e., reactor- or particle-level) is the dynamic nature of the operation of such apparatuses and the need for periodic regeneration of the bed. This not only requires more time-consuming dynamic simulations, but also optimization at the level of the entire installation consisting usually of several reactors operating in parallel, that is in reaction and regeneration mode (Bermejo-López et al., 2022). While in industrial practice, steady-state operation of the plant is usually desired, and most apparatuses are designed in relation to this most economical and safe mode of operation, in certain applications, the dynamic operation may be more advantageous or even inevitable (Fischer and Freund, 2020). The latter case is also closely related to power-to-X technologies, as large fluctuations in the renewable power supply are transferred to chemical reactors. This inevitably results in the need to simulate the dynamics and dynamic optimization – both being computationally expensive – of systems that have so far been analyzed only in terms of steady state.

Distributed parameter and multi-scale character, frequently dynamic behavior, and the almost always nonlinear nature of

the models of chemical processes, make it still challenging to solve them using classical discretization methods (Bremer et al., 2017). Thus, it is desirable to directly transform the original infinite-dimensional problem into its low-dimensional approximation or to reduce a large set (i.e., a finite yet high-dimensional model) of algebraic equations or ODEs, resulting, for example, from the approximation of spatial differential operators by finite difference schemes, to its low-dimensional truncated form. There are a variety of methods for performing both, but the discussion that follows focuses on one of the most widely used, namely the combination of proper orthogonal decomposition (POD) and Galërkin projection method (Hesthaven et al., 2007; Holmes and Lumley, 1996). This choice is dictated by the fact that POD is closely related to data mining, as it is classified along with machine learning-based techniques in the group of feature extraction methodologies. In fact the so-called basis functions used in the POD-Galërkin method to transform a full-order model (FOM), i.e. the finite yet high-dimensional model, into a reduced-order model (ROM) are often referred to as empirical functions or modes.

The POD method offers the ability to determine a set of optimal empirical basis functions from a set of observations, derived from a numerical simulation or experiment, representative of the spatio-temporal or spatio-parametric complexity of the system under study (Holmes and Lumley, 1996). The method originated from the works of Pearson (1901) and was proposed independently by several researchers, specifically Kosambi (1943), Loève (1945) and Karhunen (1947). The empirical POD functions are optimal in the sense of the L^2 norm, which makes them perform much better in combination with the Galërkin projection method than classical orthonormal bases such as Legendre or Chebyshev polynomials. Among other things, they allow to alleviate the Gibbs effect typical of approximations with polynomial functions (Hesthaven et al., 2007).

To provide a synthesized description of the POD-Galërkin technique, let us directly consider a large system, i.e., FOM, of nonlinear algebraic equations or ODEs resulting from the discretization of a parameter-dependent steady-state boundary value problem (BVP) or PDE describing the dynamics of a distributed-parameter system, both in one spatial variable. These systems can be written in a matrix form, respectively, as (Bizon and Continillo, 2021; Cutillo et al., 2023):

$$\mathbf{A}\mathbf{y}(\mu) + \mathbf{F}\left(\mathbf{y}(\mu)\right) = 0 \tag{1}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{y}(t) = \mathbf{A}\mathbf{y}(t) + \mathbf{F}\left(\mathbf{y}(t)\right), \quad \mathbf{y}(0) = \mathbf{y}_{\text{init}} \tag{2}$$

where $\mathbf{y} = [y_1, y_2, \ldots, y_N]^T$ is a state variable vector and N is the number of discrete nodes employed for the discretization. The orthonormal POD basis $\mathbf{\Phi} = [\varphi_1, \varphi_2, \ldots, \varphi_N]^T$ can be determined by solving the following eigenvalue problem:

$$\mathbf{C}\mathbf{\Phi} = \mathbf{\Lambda}\,\mathbf{\Phi} \quad \text{where} \quad \mathbf{C} = \langle \mathbf{Y} \cdot \mathbf{Y}^T \rangle \tag{3}$$

where $\mathbf{Y}$ is a matrix constructed of M observations of the variable $\mathbf{y}$ in the parameter space μ or over time t, $\mathbf{C}$ is an autocorrelation matrix with the angular brackets denoting parameter- or time-averaging. These observations, often referred to as "snapshots," are usually derived from numerical simulations performed using the FOM, much less often from experiments (Ma et al., 2002). When the number of snapshots M is smaller than the number of discrete nodes N it is more convenient to use a modification of the Eq. (3) proposed by Sirovich (1987).

Using the determined POD basis, the state variable vector $\mathbf{y}$ can then be represented in truncated form as $\tilde{\mathbf{y}} \approx \mathbf{\Phi}_K \mathbf{c}_K$ where $K \ll N$ is an approximation order. Substitution of $\tilde{\mathbf{y}}$, respectively, into Eq. (1) and Eq. (2) followed by the Galërkin projection onto the basis yields the following systems of equations, referred to as reduced-order models (ROMs):

$$\mathbf{\Phi}_K^T \mathbf{A} \mathbf{\Phi}_K \mathbf{c}_K + \mathbf{\Phi}_K^T \mathbf{F}\left(\mathbf{\Phi}_K \mathbf{c}_K\right) = 0 \tag{4}$$

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{c}_K = \mathbf{\Phi}_K^T \mathbf{A} \mathbf{\Phi}_K \mathbf{c}_K + \mathbf{\Phi}_K^T \mathbf{F}\left(\mathbf{\Phi}_K \mathbf{c}_K\right), \\ &\mathbf{c}_K \mathbf{y}(0) = \mathbf{\Phi}_K \mathbf{y}_{\text{init}} \end{aligned} \tag{5}$$

More details concerning how to select the truncation order K, and how to collect snapshots can be found in earlier works (Bizon and Continillo, 2021; Bizon et al., 2012; Cutillo et al., 2023). Nevertheless, it is worth signaling here that, similarly to typical data-driven models, (e.g., black-box models based on neural networks) POD-based ROMs generally do not accurately predict the solution, or even completely fail also qualitatively, when used far from the conditions under which snapshots were collected. In recent decades, various sampling policies have been proposed, ranging from sampling of chaotic trajectories (Bizon et al., 2012; Kerschen et al., 2003), using randomly collected snapshots (Segala and Naseradinmousavi, 2017) or – in reference to dynamic systems – combining solutions obtained using different model parameters and initial conditions (Bizon et al., 2008; Graham and Kevrekidis, 1996).

The optimal nature of the POD-Galërkin method coming from the fact that it is based on data containing information about the behavior of the system resulted in its widespread use for ROM construction. Although the method originated in the 1940s, increased interest in it only began in the 1990s because of the widespread use of computers. The technique found its application in many areas, including chemical reactor engineering (Park and Cho, 1996), including their control (Padhi and Balakrishnan, 2003), fluid dynamics (Sirisup and Karniadakis, 2005) or chemical kinetics (Danby and Echekki, 2005). POD was also widely used to extract features from massive experimental data, including velocity fields (Druault et al., 2005; Fogleman et al., 2004) or luminosity fields (Bizon et al., 2010), in particular regarding internal combustion engine processes.

Describing all possible modifications of the POD-Galërkin method and its applications developed over the years is vir-

Table 1. A selection of some recent applications of the POD-Galërkin method for model-order reduction in chemical engineering and processing, and related areas.

Reference	Problem under study	Applied variant of POD	Objective
Bremer et al. (2017)	Two-dimensional dynamic model of a catalytic wall reactor for carbon dioxide methanation.	POD with discrete empirical interpolation method (DEIM) for treatment of nonlinearities.	Super-reduction of the model nonlinearities and reactor start-up simulation.
Yang and Armaou (2018)	Reaction-diffusion process taking place along the surface of a catalytic rod, and Kuramoto-Sivashinsky equation.	Adaptive POD (APOD) and discrete adaptive POD (DAPOD), i.e. incremental method with snapshots and modes updating.	System control with the feedback linearization.
Jo et al. (2019)	Steam reformer for polymer electrolyte membrane fuel cell (PEMFC).	Gappy POD, i.e., method managing incomplete data sets.	Determination of optimal sensor placement on steam reformers for PEMFC.
Li et al. (2020)	Fracture-dominated flow in two-dimensional porous media.	POD based on Sirovich (1987) approach.	Steady-state simulation of flow in complex fracture using ROM.
Siddiqui et al. (2020)	Wake dynamics behind 5-MW offshore reference wind turbine of National Renewable Energy Laboratory (NREL).	POD based on Sirovich (1987) approach.	Evaluation of the prediction capabilities of POD-based ROM applied to the dynamics of a full-scale wind turbine.
Sun et al. (2020)	Steady-state and transient neutron transport in a nuclear reactor system.	POD based on Sirovich (1987) approach.	Development of a fast and accurate technique for resolution of the neutron transport equation.
Masoudi and McPhee (2021)	Electrochemistry-based lithium-ion battery model.	Standard POD approach, which accounts for the sampling appropriate for control tasks.	Physics-based reduction of battery model for control-oriented problems used in real-time applications.
Li et al., (2022a)	Non-reactive gas-solid flow in fluidized bed described by Eulerian-Lagrangian approach.	Lanczos POD (LPOD) that permits to deal with extremely large correlation matrices (Eq. (3)).	Reduction of computational cost of simulation of industrial-scale powder processes.
Cutillo et al. (2023)	Two-dimensional reaction-diffusion system describing self-ignition of a stockpile of solid fuel.	POD-DEIM with innovative sampling based on k-means clustering.	Determination of global POD basis for parametric studies.
Pergam and Briesen (2023)	Compressible filter-aid cake filtration processes.	Global POD involving the use of a combination of snapshots obtained for more values of the parameter of interest.	Increasing the computational efficiency of a complex mathematical model.

tually impossible; yet it is worth examining the publications of recent years. A selection of relevant model reduction studies is summarized in Table 1. Their topics (here limited to chemical engineering and processing, and related issues), as well as the problems addressed, confirm the enormous versatility of the method. Moreover, as mentioned earlier, the POD method itself, as a feature extraction technique, has also found application in the analysis of experimental or numerical data. While this issue goes beyond the topic of this study, it is worth mentioning that new applications are constantly emerging in this area as well. Some examples include recent works on the characterization of flow dynamics of two-phase pipe flow via identification of coherent structures (Viggiano et al., 2018), analysis of the acoustic near field of a ducted axial fan (Moghadam et al., 2020) or identification of dominating mechanisms in powder mixing (Li et al., 2022b).

To illustrate the performance of the POD-Galërkin method both in terms of the savings in computation time and its high approximation accuracy, the results obtained for a non-isothermal fixed-bed catalytic reactor operating under steady-state conditions are provided in Fig. 4. Further details on the problem presented here can be found in (Bizon and Continillo, 2021). The examined process involves a gas-solid catalyzed chemical process consisting of two exothermic reversible chemical reactions A $\leftrightarrow$ B $\leftrightarrow$ C, with C being the desired product. Each of the two steps is catalyzed by a different type of active centers, integrated in bifunctional catalyst particle. The objective of the study was to optimize the fractions of each type of catalyst (f_1, f_2) within the particle, with the yield of product C with respect to reactant A, $Y_{CA,r}$, chosen as the objective function. The problem was solved using a two-scale approach, involving a one-dimensional model of a spherically symmetric catalyst particle coupled with a one-dimensional model of gas flowing through a packed bed. To build the POD basis in the first step, the so-called FOM was built using the finite difference method, and then resolved for various values of f_1 and f_2 to collect snapshots. The number

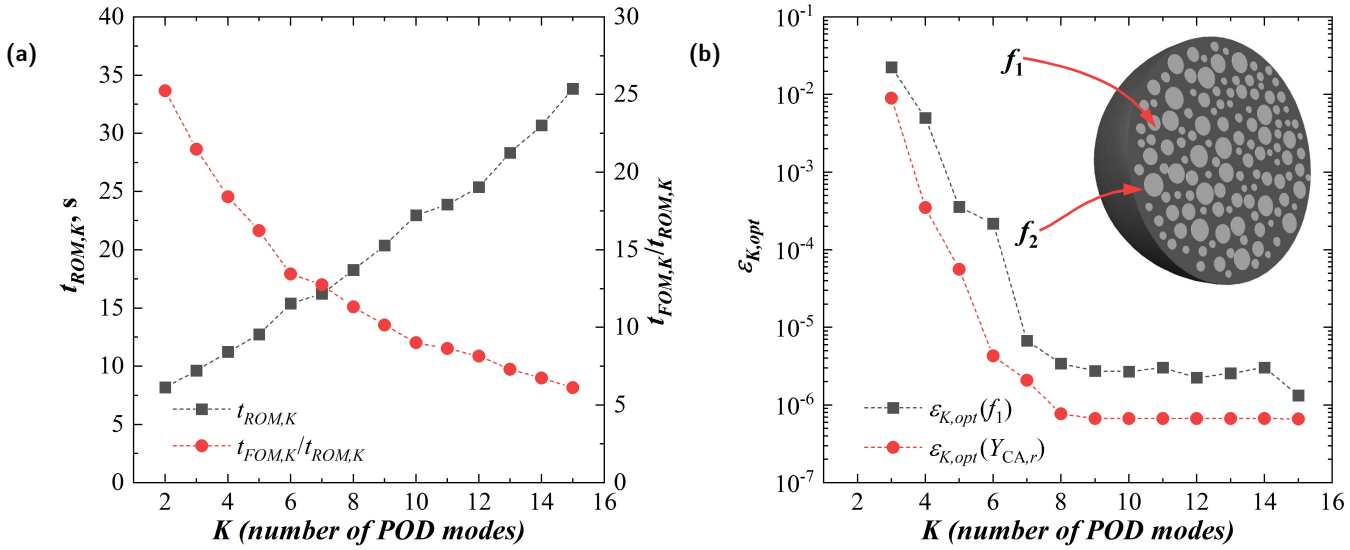

Figure 4. Performance of POD-based ROM in terms of computational time (a) and absolute error in the optimal solution approximation (b). Results concern a catalytic fixed-bed reactor integrating two types of active centers, with f_1 (design variable) being the particle fraction occupied by the active sites catalyzing first step of the process and $Y_{CA,r}$ (cost function) – the yield of product C with respect to reactant A; based on Bizon and Continillo (2021).

of nodes used for the particle and reactor model discretization was, respectively, $N_p = 101$ and $N_r = 1001$, resulting in the final $N_{tot} = 408000$ of algebraic equations. Since in case of multi-scale model it is convenient to apply the reduction methodology to the smaller scale, here the catalyst particle model, in the second step the latter was reduced via POD-Galërkin approach.

The resulting simulation time of the ROM was, depending on the adopted truncation order, K, from several to several dozen times shorter than the simulation time of the FOM (Fig. 4a). This reduction was possible while maintaining the accuracy of the computations, as confirmed by the error values (Fig. 4b) in the approximation of the optimal solution, reported here both for the design variable and the objective function. It should be mentioned at this point that the problem studied here was one-dimensional, hence the time savings obtained may not be very impressive. For problems in a two- or three-dimensional space, the performance of the method is far more remarkable: for example, for the problem studied in (Cutillo et al., 2023) and concerning a two-dimensional reaction-diffusion system describing self-ignition of a stockpile of solid fuel, the speed-up of ROM with respect of FOM was of the order of 1000. Nevertheless, to obtain such performance the POD-Galërkin methodology was additionally combined with an discrete empirical interpolation method (DEIM).

The analysis of Eq. (4) and Eq. (5) shows that despite the decrease of the number of equations from N to K, the nonlinear terms are evaluated in the original N-dimensional space. This, in turn, implies that for highly nonlinear systems such as those generally encountered in chemical and process engineering, the time savings from using POD-Galërkin are not such as would be desirable. This problem can be addressed by implementing the aforementioned DEIM, developed by Chaturantabut and Sorensen (2010). It allows to determine a relatively small number of grid nodes, on which the nonlinearities of the model are then evaluated. The DEIM technique consists of two steps: deriving an extra POD basis from spatio-temporal or spatio-parametric representations of the nonlinear term and then determining, based on a dedicated algorithm (Bizon, 2017; Chaturantabut and Sorensen, 2010; Cutillo et al., 2023), a small number of the grid nodes, referred to as interpolation indices, at which the value of this term is evaluated in the reduced model. Because of the adoption of two POD bases for the construction of the ROM within POD-DEIM-Galërkin methodology, the technique is sometimes called super-reduction.

Let us assume that $\mathbf{\Phi} = [\psi_1, \psi_2, \ldots \psi_N]^T$ is such a POD basis determined by solving Eq. (3) and using as snapshot set of representations of nonlinear term which is present in the generic FOMs given by Eq. (1) and Eq. (2). Then the nonlinear function, $\mathbf{F}$, can be approximated using $J \ll N$ basis functions as $\mathbf{F}(\mathbf{y}) \approx \mathbf{\Phi}_J \mathbf{q}_J$ where:

$$\mathbf{q}_J = \left(\mathbf{P}^T \mathbf{\Phi}_J\right)^{-1} \mathbf{P}^T \mathbf{F}(\mathbf{y}) \tag{6}$$

where $\mathbf{P} \in R^{N \times J}$ is a matrix of interpolation indices identified following the algorithm proposed in (Chaturantabut and Sorensen, 2010). Introduction of the above formula in Eq. (4) and Eq. (5) yields, respectively, POD-DEIM-Galërkin reduced ROMs:

$$\mathbf{\Phi}_K^T \mathbf{A} \mathbf{\Phi}_K \mathbf{c}_K + \mathbf{\Phi}_K^T \mathbf{\Phi}_J \left(\mathbf{P}^T \mathbf{\Phi}_J\right)^{-1} \mathbf{F}\left(\mathbf{P}^T \mathbf{\Phi}_K \mathbf{c}_K\right) = 0 \tag{7}$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{c}_K &= \mathbf{\Phi}_K^T\mathbf{A}\mathbf{\Phi}_K\mathbf{c}_K \\ &+ \mathbf{\Phi}_K^T\mathbf{\Phi}_J\left(\mathbf{P}^T\mathbf{\Phi}_J\right)^{-1}\mathbf{F}\left(\mathbf{P}^T\mathbf{\Phi}_K\mathbf{c}_K\right) \\ \mathbf{c}_K\mathbf{y}(0) &= \mathbf{\Phi}_K\mathbf{y}_{\text{init}}\end{aligned} \tag{8}$$

where $\mathbf{P}^T\mathbf{\Phi}_K\mathbf{c}_K \in R^J$ (note that in Eq. (4) and Eq. (5) $\mathbf{\Phi}_K\mathbf{c}_K \in R^N$, and $J \ll N$).

The DEIM algorithm selects interpolation indexes such that the nonlinear term approximation error is minimized. Figure 5 shows its outcome using as an example a dynamic model of a single non-isothermal catalyst particle, being and extension of the problem examined in (Bizon, 2017). If a single irreversible chemical reaction A → B takes place in the catalyst, the following nonlinear expression emerges in the model equations:

$$r_{\mathrm{A}}\left(C_{\mathrm{A}}, T\right) = k_0 \exp\left(-E/RT\right) C_{\mathrm{A}} \tag{9}$$

which, after the introduction of classically defined dimensionless variables and the Thiele modulus, Φ, takes the form:

$$r_{\mathrm{A}}(\beta,\theta) = \Phi^2 \frac{\exp(-\gamma/\theta)\beta}{\exp(-\gamma)} \tag{10}$$

where β, θ and γ are, respectively, dimensionless concentration, dimensionless temperature and dimensionless activation energy. A set of spatio-temporal representations of the nonlinear term obtained from FOM simulation is presented in Fig. 5a (negative values result from the minus sign in the equation preceding the source term), while Fig. 5b shows the first four POD bases determined using as snapshots these representations. In addition, the locations of first 10 interpolation indices determined using DEIM are marked with circles in Fig. 5b, where larger symbols refer to the first 5 of them, additionally numbered in order of relevance. It can be observed that more interpolation indices are located nearby the catalyst surface, where the values of nonlinear function are characterized by larger gradients (Fig. 5a).

Based on the above, it can be stated that projection-based techniques combined with POD methodology provide undoubtedly an extremely attractive solution for building fast yet faithful models. The variety of modifications of this methodology that have emerged over the years, the different problems that can be solved using them – from steady-state and transient simulations, optimization or determination of sensor locations – make the data-enhanced but still constitutive equation-based ROMs a powerful simulation tool. In addition to computational efficiency and accuracy, the advantage of ROMs based on the POD-Galërkin approach is the relatively easy numerical implementation, as evidenced by the rather simple matrix notations presented in this section. Nevertheless, in situations when a first-principle model is not available, or is simply too complex, or when even greater computational efficiency is needed, yet another type of tool is required, namely surrogate or hybrid models, sometimes also referred to as equation-free models.

3. SURROGATE AND HYBRID MODELS – BLACK-BOX AND GRAY-BOX MODELS

Although the concepts of surrogate models, also known as metamodels, emulators, substitute or black-box models, and of hybrid models, also known as gray-box models (Sansana et al., 2021; Williams and Cremaschi, 2021) have gained popularity in the scientific literature relatively recently, in essence they are not entirely new tools. In fact, both are based on more or less sophisticated input-output data-approximation techniques, which have been present in the field of chemical and process engineering modeling for many decades. How-

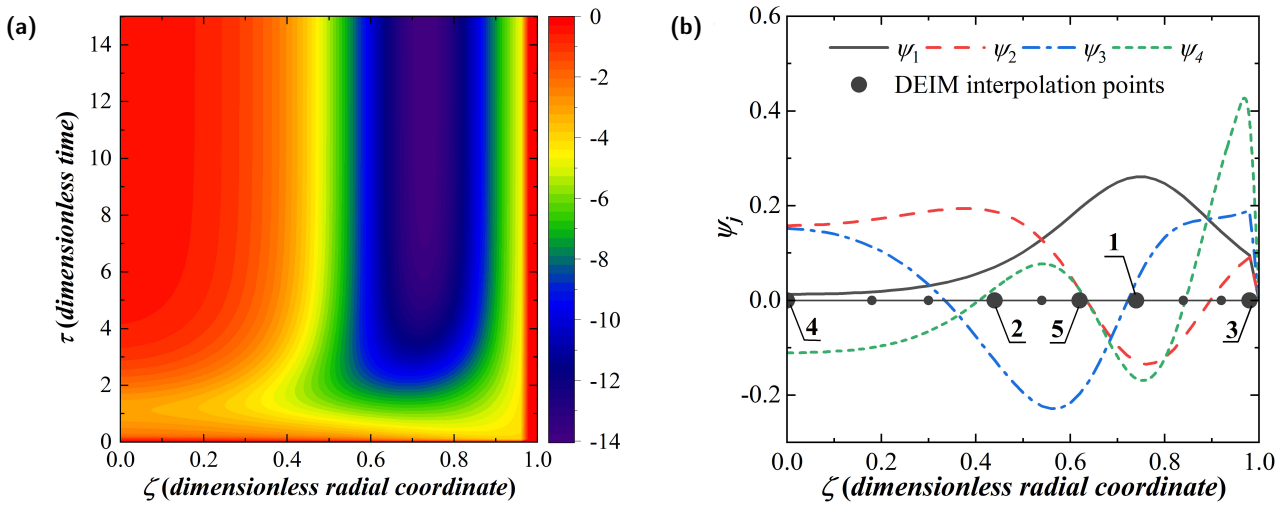

Figure 5. Spatiotemporal set of evaluations of the nonlinearity of the single catalyst particle model (a) and corresponding POD modes with the first 10 interpolation indices (•) determined using DEIM (b) for $\gamma = 20$, $\delta = 0.2$, $\Phi = 2$, Le $= 0.1$ and $k_0 = 1/\mathrm{s}$; based on Bizon (2017) and extended here to non-isothermal case.

ever, a growing interest in the last two decades in approximation and approximation-based optimization methods accompanied by development of new metamodeling techniques has only contributed to the emergence and widespread adoption of these concepts (Wang and Shan, 2007). This has resulted, in the first step, into advancement and dissemination of techniques based on machine learning, such as neural networks. The development of a data-driven framework for knowledge extraction and model construction has been further triggered by the increasing capabilities to collect, store and process massive sets of data. Finally, the fourth industrial revolution, widespread digitalization and the smart manufacturing paradigm have brought surrogate and hybrid models to the forefront of many practical industrial applications (Sansana et al., 2021; Yang et al., 2020).

It is difficult to provide a precise definition of both the surrogate model and the hybrid model; the definitions given by different authors vary significantly, and sometimes these terms are used even interchangeably (Sansana et al., 2021). Likely, the most common definition of a surrogate model is that it is a substitute model allowing input data to be mapped into output data if the actual relationship between them is either unknown or computationally expensive to evaluate (Williams and Cremaschi, 2021). The latter case indicates that surrogate models, which are basically simpler representations of more complex models, do not necessarily need to be derived using experimental data. In fact, according to some reports, one way to design a surrogate model is by generating data points through numerical simulations of a complex mechanistic model based on first principles, e.g. with the aid of computational fluid dynamics (CFD), and then employ them to built a data-driven substitute (Sansana et al., 2021). At the same time, according to the commonly accepted classification of hybrid models, the latter approach is basically equivalent to a hybrid model, generally consisting of a combination of first-principle model (FPM) and data-driven model (DDM), in this case having a serial arrangement (Fig. 6b, Sansana et al., 2021; Yang et al., 2020).

In hybrid gray-box modeling, DDMs are usually complementary to FPMs, used to deal with mechanisms that are difficult to describe using first-principles knowledge, or that are unfeasible to include into a mechanistic model due to the high computational effort required. Examples of such mechanisms include not fully understood complex (bio-)chemical reaction kinetics (Rogers et., 2023), or the evolution of particle size distributions (PSD) in systems described by population balance equations, such as the crystallization process (Makrygiorgos et al., 2020), which is generally characterized by numerous uncertain parameters. More examples are provided further in Table 2.

Depending on the balance maintained in the final hybrid model between mechanistic (white-box model) and data-driven (black-box model) components, there exists a broad spectrum of shades of gray of such models (Rogers et., 2023). The white and black box components can also be assembled in different manners; the basic arrangements are shown in Fig. 6. A serial arrangement in which a black-box model (Fig. 6a) is followed by a white-box model (S-BW) is typically used when there is sufficient input-output data that can be incorporated into a nonparametric black-box model, but the detailed knowledge of the process mechanism is not available, for instance, when the kinetics of a chemical reaction or transport properties (e.g. diffusion coefficients) are unknown (Zendehboudi et al., 2018). The output of the black-box model is then fed to the mechanistic white-box model, making the whole hybrid model more reliable. The reverse arrangement, i.e., first a white-box model and then a black-box model (S-WB, Fig. 6b), is primarily applicable when some but yet limited data are available from a mechanistic model, which is very time-consuming to simulate. In such a case, the results obtained from the mechanistic model can be used to train the black-box model (Zendehboudi et al., 2018). The parallel structure (P, Fig. 6c) is primarily used when a mechanistic model is provided, but its predictive capabilities are restricted due to the limited description of certain effects (Sansana et al., 2021) or when some elements present in the model reduce computational efficiency, such as nonlinearities (Zendehboudi et al., 2018). In both cases, the model performance can be improved by embedding a properly trained data-driven submodel. In addition to the basic arrangements outlined here, there exist other mixed structures; more details can be found among others in (Zendehboudi et al., 2018).

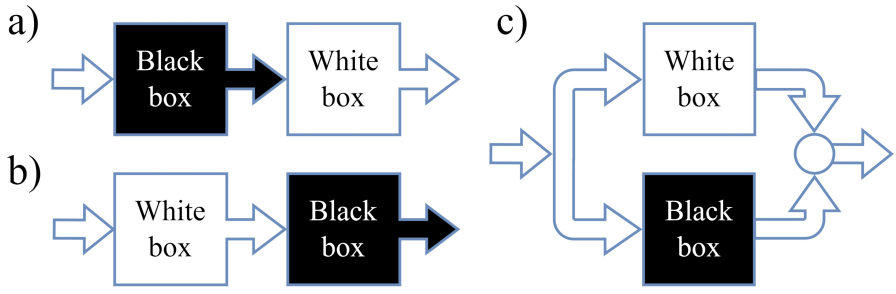

Figure 6. Basic structures of hybrid models; based on Yang et al. (2020) and Sansana et al., (2021): (a) S-BW – serial structure with black- followed by white-box model; (b) S-WB – serial structure with white- followed by black-box model; (c) P – parallel structure.

Table 2. A selection of some recent applications of surrogate and hybrid modeling in chemical engineering and processing, and related areas.

Reference	Problem under study	Applied data-driven methodology	Objective
Ali et al. (2018)	Single mixed refrigerant process of natural gas liquefaction.	Surrogate model; approximation via radial basis functions (RBF) involving thin slab splines trained on data generated with Aspen HYSYS (AspenTech).	Addressing the computational burden associated with the modeling and optimization of the process under study.
Babanezhad et al. (2020)	Thermal and hydrodynamic behavior of co-current bubble column gas-liquid reactor.	Hybrid model; differential evolution fuzzy interference system (DEFIS) applied to learn flow patterns based on data determined using CFD.	Understanding and prediction of the flow patterns in bubble column reactors.
Carranza-Abaid et al. (2020)	Vapor-liquid equilibrium (VLE) behavior of a mixture containing CO_2, H_2O and monoethanolamine (MEA).	Surrogate model; feedforward artificial neural network (ANN) with a single hidden layer, trained using data generated with a semi-empirical model.	Development of a technique integrating computational efficiency with the robustness of rigorous models; validated against the eNRTL method with the Peng-Robinson equation.
Makrygiorgos et al. (2020)	Two-dimensional population balance equation (PBE) of batch cooling crystallization of ibuprofen.	Surrogate model; sparse polynomial chaos and kriging methods applied to numerical results obtained from PBE solved by finite volume method.	Quantification of uncertainty in problems with a large number of uncertain inputs; probabilistic uncertainty characterized in the study 20 parameters related to growth, dissolution and physico-chemical properties.
Krippl et al. (2021)	Single-pass tangential flow filtration (SPTFF) for continuous biomanufacturing.	Hybrid model; ANN trained using experimental data from a membrane operating in TFF mode using bovine serum albumin (BSA) as a model protein.	Performance prediction of serial and parallel SPTFF with up to three membrane cassettes oriented at efficient process development and model predictive control (MPC).
Jayaweera et al. (2022)	Heat and mass transfer in a full-scale hybrid-draft cooling tower and an induced draft cooling tower.	Hybrid model; single layer feedforward ANN featuring RBF activation functions used to estimate volumetric mass transfer coefficient.	Development of a model with high predictive accuracy capable of being implemented in real time to optimize fan-generated air flow rates and to monitor the thermal performance of a full-scale cooling tower.
McKay et al. (2022)	Fed-batch acid crystallization process.	Hybrid model; submodel based on subspace identification (SID) derived from simulated data batches using singular value decomposition (SVD).	Using a potentially challenging first-principles model in a way suited to MPC implementation.
Pinto et al. (2022)	Stirred fed-batch tank bioreactors for *Escherichia coli* recombination and *Pichia pastoris* cultivation.	Hybrid model; shallow and deep feedforward ANNs with *tanh* and rectified linear unit (ReLU) activation functions trained with synthetic (*Escherichia coli*) and pilot (*Pichia pastoris*) data sets.	Systematic enhancement of the generalization of deep hybrid models versus their shallow equivalents combined with exploration of different learning techniques.
Matsunami et al. (2023)	Dissolution of paracetamol tablets produced through dry and wet granulation methods.	Surrogate model; Weibull model fitted to input-output data generated using a mechanistic model implemented in gPROMS (Siemens PSE) with Latin hypercube sampling (LHS).	Identification of critical input parameters in the tablet fabrication process that affect dissolution behavior, targeted towards efficient design of tablet manufacturing.
Xing et al. (2023)	Trickle bed (TB) and packed bubble column (PBC) reactors for CO_2 capture.	Surrogate model; response surface methodology (RSM) and extended adaptive hybrid functions (E-AHF) trained using data generated from mechanistic model.	Rapid design and optimization of TB and PBC reactors oriented at improved CO_2 capture and reduced energy and water consumption.

There is a wide selection of techniques for constructing data-driven models; very broadly, they can be divided into interpolation (e.g., kriging) or non-interpolation (e.g., neural networks) methods. Among the most popular data-driven methodologies used to construct both surrogate models and black-box components of hybrid models are (Bárkányi et al., 2021; Bhosekar and Ierapetritou, 2018; Bradley et al., 2022; Sansana et al., 2021; Williams and Cremaschi, 2021; Yang et al., 2020; Zendehboudi et al., 2018):

- Artificial neural networks (ANN) consisting of an input and output layer and a number of information processing units, referred to as neurons, embedded in the so-called hidden layer, connected by information flows; ANNs are considered universal approximators as they do not require any structural knowledge regarding the modeled system (Sansana et al., 2021).
- Support vector machines (SVM) that is, supervised learning techniques that can be used for classification, regression and outlier detection; SVM methodology involves mapping the input data into a high-dimensional space followed by determination of the optimal hyperplane, which divides the data into classes (Sansana et al., 2021; Zendehboudi et al., 2018).
- Fuzzy logic (FL), allowing to implement infinite-value logic to deal with systems characterized by so-called "non-crisp" boundaries; FL applies primarily to control, decision-making and classification (Zendehboudi et al., 2018).
- Radial basis functions (RBF) generating interpolation surrogate in the form of a linear combination of local univariate functions using selected distance measures from a point to determined centers (Bárkányi et al., 2021); RBF functions are also very often used as activation functions in hidden layers of ANNs, known as RBF networks (RBFN) (Williams and Cremaschi, 2021).
- Kriging, originating from geostatistics and involving the interpolation of discrete data points based on a statistical approach utilizing a weighted variance that minimizes the error between actual and estimated values to determine a so-called response surface model (Bárkányi et al., 2021).
- Polynomial functions, which are the simplest and most commonly used substitute models in engineering calculations, although they are only suitable for describing simple relationships among variables (Bárkányi et al., 2021).

Perhaps the most extensively used surrogate models nowadays are ANNs. They can feature a variety of architecture and employ different activation functions; a summary of the most common designs, along with their advantages and disadvantages, and potential engineering applications, can be found among others in (Zendehboudi et al., 2018). A summary of the advantages and disadvantages of the other metamodeling techniques mentioned above can be found in (Williams and Cremaschi, 2021). More details on the theory behind the techniques used for surrogate and hybrid modeling, along with applications in the field of chemical and process engineering, can be found in numerous extensive review papers (Bárkányi et al., 2021; Bhosekar and Ierapetritou, 2018; Sansana et al., 2021; Williams and Cremaschi, 2021; Yang et al., 2020; Zendehboudi et al., 2018). To emphasize the versatile nature of these techniques and the different scales of problems to which they are applied, illustrative examples of some recent applications involving both surrogate and hybrid modeling are listed in Table 2.

The brief summary of applications provided in Table 2 confirms that one of the most popular metamodeling techniques are ANNs, particularly those incorporating RBFs as hidden layer activation functions, i.e. RBFN. For a shallow feedforward RBFN consisting of an input layer, a single hidden layer, and an output layer, the output of the network, y, for a given input x can be written as (Bizon et al., 2014; Broomhead and Lowe, 1988):

$$y = \sum_{i=1}^{N} \omega_i R_i(x) + \omega_0 \tag{11}$$

where ω_i are the weights associated with connections between neurons, ω_0 is the bias term, R_i are the RBF activation functions, whereas N is the number of neurons within the hidden layer. The radial basis function can be further expressed as:

$$R_i(x) = \varphi\left(\|x - c_i\|\right) \tag{12}$$

with φ denoting the radial function that substantiates the nonlinear nature of the model, while c_i is the so-called center. There exists a variety of radial functions, but the most widely used is the so-called Gaussian function defined as follows:

$$\varphi(r) = \exp\left(-r^2/\sigma^2\right) \tag{13}$$

Symbol r in Eq. (13) denotes the Euclidian distance between vector x and center c_i, and σ is the spread parameter.

To illustrate the operation of RBFNs, representative results demonstrating their performance in converting input signals into output signals are shown in Fig. 7. Figure 7a shows the outcome of using a neural network as a soft sensor. The soft sensor is a variation of the surrogate model developed to measure required, for instance, for diagnostic reasons, signal using a low-cost alternative to the actual real sensor. Soft sensors are also used when certain variables cannot be measured directly, or the measurement is performed only at a low sampling rate or offline (Ferreira et al., 2022). The results shown in Fig. 7a concern the implementation of RBFN as a soft sensor for the diagnosis of a three-cylinder real compression-ignition engine; further details can be found in (Bizon et al., 2014; Bizon et al., 2015). Due to the observed correlation between in-cylinder pressure and engine block vibration, an ANN was constructed using an acceleration signal measured with a low-cost accelerometer and in-cylinder pressure measured with a piezoelectric pressure transducer as input and output, respectively. A shallow feedforward RBFN

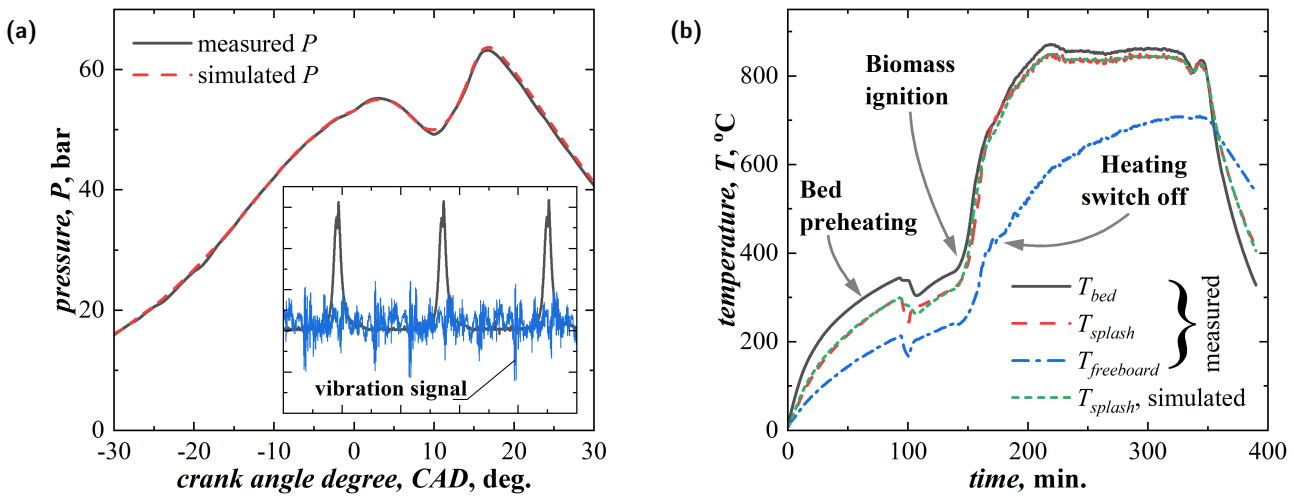

Figure 7. Performance of RBFN implemented as a soft sensor to monitor the in-cylinder pressure in the internal combustion engine (a) and as a submodel to predict the temperature in the splashing zone of the fluidized-bed combustor (b); based, respectively, on Bizon et al. (2014) and Marra et al. (2015).

with 25 neurons in the hidden layer proved to approximate the in-cylinder pressure evolution very well, despite the large noise inherent to the acceleration signal (Fig. 7a). It is worth adding that due to the simple structure consisting of radial functions and weights, that can be generally collected into an array, such a network can be very easily implemented for real-time online operation. In particular, in the case presented here, the network trained in the Matlab (MathWorks) environment using the *newrb* function was exported to the CONCERTO environment, which is an integral part of AVL IndiCom platform (AVL List GmbH), i.e. advanced combustion analysis and data acquisition software for internal combustion engines. As a result, the performance of the designed soft sensor was also validated successfully online.

Another example, shown in Fig. 7b, involves using RBFN as a surrogate submodel embedded in a mechanistic model of a biomass-fueled combined heat and power (CHP) system developed within the project MEGARIS: Micro Electric Generator from Alternative Renewable energy Innovative Stirling engine (Angrisani et al., 2013; Marra et al., 2015). The system consisted of a fluidized-bed combustor (FBC) with a Stirling engine heat transfer head immersed in the bed. To keep the model relatively simple, the CHP system was represented by a series of interconnected blocks described using lumped-parameter models. To describe the dense phase of the fluidized bed operating in bubbling regime, an ideally mixed reactor model was employed; the same approach was adopted to describe the freeboard zone above the bed. However, the intermediate zone, the so-called splashing zone, cannot be described using the ideal reactor model. It also cannot be neglected in the entire system model, as the temperature within the splashing zone is crucial for the correct assessment of energy fluxes to and from the bed. To address this problem, the RBFN was designed and trained using experimental data collected during start-up (the stage of electrical heating of the bed and subsequent ignition of the biomass) and during steady-state operation of the apparatus. Given the large number of different variables measured during the experimental study, the construction of the final network was preceded by a sensitivity analysis. It provides an opportunity to evaluate input variables in regard to the significance of their contribution to the output variable and to identify irrelevant variables. In the final network design, the temperature of the dense phase of the bed, T_{bed}, and the temperature of the freeboard zone, $T_{\text{freeboard}}$, were used as input variables. The results shown in Fig. 7b confirm that RBFN having only 3 neurons in the hidden layer is able to describe the temperature in the splashing zone, T_{splash}, quite accurately.

Interest in data-driven models has undoubtedly increased over the past few years. This is confirmed by the constant emergence of new metamodeling methods as well as new applications, often non-trivial, as briefly characterized here. In the field of chemical and process engineering, as well as in related fields, these big data-based approaches are now offering the possibility of novel, improved solutions to old problems and tackling problems that were previously unsolvable. Even more, these techniques, sometimes developed years ago and already well-established, are now increasingly being coupled with real objects, in the form of so-called digital twins. The latter are undoubtedly a tool that will soon transform the industry into a more adaptive, resilient, and intelligent one.

4. DIGITAL TWINS – A NOVEL USE OF COMMON TOOLS

According to most literature sources (Singh et al., 2021), the concept of the digital twin (DT) emerged in 2002 at the University of Michigan in relation to product lifecycle manage-

ment (PLM). The model developed there by Michael Grieves, initially referred to as the "mirrored spaces model" involved three elements, namely real space, virtual space and the flow of information between them. It is precisely this bidirectional transfer of information between the digital twin, i.e. the numerical model, and the simulated real world asset, i.e. the physical twin, that distinguishes this design from the classical numerical model (Fig. 8). Nowadays, in the era of the pursuit towards widespread digitalization and smart industry, the concept of the digital twin has gained a prominent presence in the scientific literature. This is confirmed by the number of publications in this area, which has recently been growing each year much faster than, for example, the number of publications on surrogate or hybrid modeling (Fig. 1b). This is to some extent the result of overusing the DT term and also employing it to refer to models that do not meet the exact definition of a digital twin.

What makes a model to be a digital twin is essentially the following (Wright and Davidson, 2020):

- it is the model of an existing real object;
- there is an evolving set of data concerning the object, transferred to the model;
- the model is dynamically updated/improved based on the transferred data.

As a result, a digital version of a physical asset can analyze its performance, simulate its behavior under certain conditions, helping, for instance, to improve the process or prevent its failure. Thus, the primary difference between the conventional data-driven model and DT is the tight coupling and real-time interaction with its real physical twin. Thanks to these features, DTs can be used for real-time monitoring, maintenance, optimization, design or remote access. (Singh et al., 2021).

Digital twins may basically rely on various types of models and submodels that are able to reflect the behavior of the physical object being twinned. Despite the coupling between twins, it is desirable to use mechanistic models at least locally because of their greater reliability and ability to extrapolate. The contribution of data-driven elements in DT is largely dependent on the time-scale of the process and the associated time that might be devoted to simulating the model, but also to its measurement-based update. In short: the entire DT should be sufficiently physics-based, accurate and fast for its intended applications (Singh et al., 2021).

In addition to utilizing various metamodeling techniques developed over the past few decades, the operation of digital twins benefits greatly on a range of other tools and techniques relatively well established in the field of chemical engineering, and more generally in industrial practice, including among others advanced control strategies (Kestering et al., 2023). Worth mentioning are also operator training simulators (OTS), whose origins date back to the late 1970s (Bose et al., 2023). OTS systems, that is, virtual plants used to develop the capabilities of plant personnel, including handling plant abnormalities and failures, are indeed sometimes referred to as "early stage" DTs (Appl et al., 2020). The latter is due to the fact that, in general, OTSs are used for practical training of plant operators without the need to perform the actual process, so there is no online pairing of the virtual replica with the real process, thus no risk on impairment of equipment components. The considerable success of OTSs in industrial practice, which has resulted in a number of applications (mainly in petrochemical and energy sectors) over the years and the rapid development of commercial software (Patle et al., 2014) for developing such training systems, makes them an attractive starting point for the construction of actual DTs. In fact, according to de Beer and Depew (2021), OTS represents one of the last phases of the DT design lifecycle, which essentially consists of conceptual engineering, front-end engineering and design, detailed engineering, commissioning and start-up, and ultimately DT operation. In the penultimate phase of the commissioning of the digital twin, a tool is provided that can be used for operator training. Therefore, DTs not only rely on a variety of already existing methodologies, but also supply a tool that eliminates the investment in a separate OTS.

Once the DT is available, another important aspect is to keep it up to date by performing proper maintenance throughout its operational lifecycle. This is essential to ensure correct

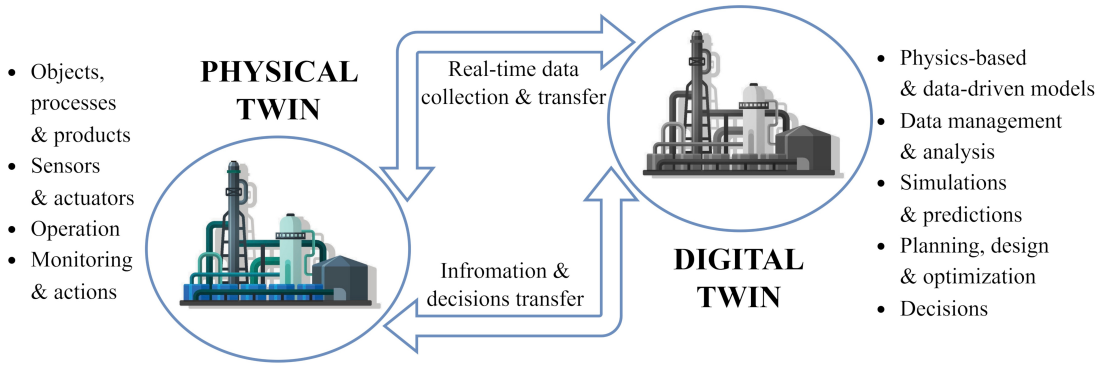

Figure 8. Constituent elements and information flow between physical and digital twin.

predictions of the digital replica. One of the challenges to be faced in this regard is the proper management of data originating from the physical twin. A systematic literature review concerning this issue can be found, among others, in the work of Correia et al. (2023). A data management system that is an integral part of a digital twin (Fig. 8) must, first and foremost, deal with data of a heterogeneous character. This heterogeneity, i.e. diversity of sources that are generating data within physical twin, renders it crucial to properly transform, integrate and match them so that they can be utilized to update, with suitable machine learning algorithms, data-driven submodels of DTs (Correia et al., 2023). Another aspect that is closely related to DT maintenance during its whole operational lifecycle is the quality of the data itself (Chen et al., 2023). The raw sensor streams should be cleaned using statistical methods or deviation/anomaly detection methods prior to being used as an input to simulation tools that constitute the elements of the DT (Correia et al., 2023). It is also important to be aware of the degradation of the hardware, such as the sensors on which the simulation, prediction, action and maintenance activities of the DT are based (Studer et al., 2021). Since the topic is really vast, it is difficult to address all aspects here, and the aforementioned concerns are intended to underline that the development of efficient and reliable DTs still involves many challenges, including those of a scientific nature.

Although the concept of DT is usually associated with the prediction of operation and control of entire plants, e.g. brewery (Koulouris et al., 2021), or processes, e.g. food supply chain (Shoji et al., 2022), the tool can also be applied to smaller parts of engineering problems. This is confirmed by a compilation of some recent applications, summarized in Table 3.

The overview presented in Table 3 indicates not only the variety of DT applications but also identifies some tools that can

Table 3. A selection of some recent applications of digital twin (DT) in chemical engineering and processing, and related areas.

Reference	Problem under study	Models used and measurements performed	Objective of the study
Koulouris et al. (2021)	Brewery process consisting of brewhouse, fermentation and bottling sections.	Block model developed in SuperPro Designer (Intelligen, Inc.) for multi-product batch and semi-continuous manufacturing facilities combined with simulated data.	Calculation of process cycle time and improvement of plant efficiency by identifying and eliminating bottlenecks, and carrying out an economic analysis.
Wang et al. (2021)	Particle breakage during milling using impact pin mill.	Population balance model developed and customized using gPROMS (Siemens PSE) platform integrated with nano- and particle scale data.	Development of alternative measures of model-based design for the milling process by integrating complementary information at different scales.
Lunev et al. (2022)	Thermal performance of AlSi7Mg open-cell metal foams.	Finite element analysis (FEA) based on digitalized with the aid of laser flash analysis (LFA) and X-ray computed tomography (CT) samples.	Elucidation of the nature of heat conduction in metallic foams and the feasibility of using LFA for characterization of such materials.
Shoji et al. (2022)	Temperature-based monitoring of postharvest fruits within supply chain.	Two-dimensional axisymmetric energy equations of fruits developed in COMSOL Multiphysics (COMSOL Inc.) combined with air temperature datasets.	Optimization of the cold supply chain with the aid of DT oriented at maximization of shelf life and uniform product quality.
van Rooij et al. (2021)	Degradation of membrane in a reverse osmosis (RO) pressure vessel due to biofouling.	Mathematical model of RO vessel with eight membrane elements with real measurement-based parameter estimation.	Development of a decision support system (DSS) for membrane restoration strategies.
Yu et al. (2022)	Thermal system of 660 MW ultra-supercritical double reheat power plant.	Gray-box model combining basic mass and energy balances with historical operation data.	Development of a new thermal system simulation concept and method to support the advancement of DT for thermal power plants.
Galeazzi et al. (2023)	Industrial amine scrubbing process from the exhausted oil refinery.	Steady-state model developed in Aspen HYSYS (AspenTech) relying on design of experiment (DoE) concept and validated against real plant data.	Proposal of a framework for automated surrogate model development to overcome numerical limitations in conventional industrial process simulations.
McLaughlin et al. (2023)	Hierarchical CO_2 electrolyzer gas diffusion electrode.	Simulations with GeoDict (Math2Market GmbH) combined with focused ion beam (FIB), scanning electron microscopy (SEM) and X-ray imaging.	Optimization of the electrode structure oriented at higher conversion efficiencies and lifetime, and lower cost.

Table 3 continued.

Reference	Problem under study	Models used and measurements performed	Objective of the study
Procacci et al. (2023)	Semi-industrial moderate or intense low oxygen dilution (MILD) combust-ion furnace operating under varying conditions and fuel charges (from pure H_2 to pure CH_4).	POD-based ROM built using data from three-dimensional CFD simulations with real time temperature update towards the experimental measurements.	Use of a sparse sensing framework to develop DT able to determine optimal sensor placement and predict three-dimensional temperature field using few measurements.
Wang et al. (2023)	Scheduling of laminar cooling water supply system for hot rolling mills.	Fast neural model of water consumption based on online sequential extreme learning machine (OS-ELM) trained using historical production data.	Prediction of water consumption trends, reduction of water and power consumption, and improvement of safety and reliability of the pump station operation.

be employed for practical implementation of digital twins, e.g. gPROMS (Siemens PSE) or Aspen HYSYS (Aspen Tech). However, it should be noted that as of today, the capabilities of available engineering software are still rather limited as the flow of information they provide is usually unidirectional. Therefore, digital transformation of industry definitely requires also improved, next generation process simulators (de Beer and Depew, 2021).

5. SUMMARY

The modern times can undoubtedly be described as the times of big data. The methodology outlined in this study, which is, naturally, much more extensive, shows that over the past decades in chemical and process engineering, as in virtually any field, there has been a gradual shift from mechanistic to semi- or fully data-driven models. The former, obviously, are continuously utilized and further developed, and thanks to ever-increasing computational power, numerical solutions for more and more complex multi-scale and multi-physics models are now available. Approaches based on big data today make it possible to solve many problems that were previously impossible to address. In this context, in addition to complex problems with uncertain parameters and unknown mechanisms, surrogate and hybrid models that enable real-time computation, which are essential mainly for process control, deserve special attention. Their close coupling and real-time upgrading based on measurement data, which contributed to the development of digital twins, is also a necessary step on the road to smart manufacturing and to the digitalization of the chemical industry, and beyond. However, in this era of big data, big theory must not be ignored or even forgotten.

While Aris and Froment, quoted at the very beginning, looked forward to further advancement of science, the development that is now taking place is becoming somewhat hazardous. Particularly noteworthy here is the criticism presented several years ago in the journal Philosophical Transactions A by Peter Coveney (Coveney et al., 2016), a professor of physical chemistry and honorary professor of computer science, in the paper entitled "Big data need big theory too". Referring to philosopher Francis Bacon, whose works influenced the scientific revolution that took place in the 16^{th} and 17^{th} centuries, Coveney emphasizes that modern science emerged in the 17^{th} century as a combination of observation and reason, and he argues that the often blind collection of (sometimes unreliable!) data makes the era of big data a modern version of pre-17^{th} century thinking. Moreover, many proponents of big data believe that soon there will be no need for theory to understand many phenomena, and, according to Coveney, even in relatively simple scientific disciplines, machine learning and big data are "intoxicating" to many researchers.

Therefore, in this journey from mechanistic to data-driven models, it is important to keep in mind that the extremely versatile tools available nowadays should first and foremost complement and enrich existing theory and make it more powerful, rather than replace what is well-known and well-established. In developing data-driven models, more emphasis should be placed now on the reliability and curation of the data. In addition, it should be remembered that the predictive capabilities of such models are limited due to the difficulty of extrapolation. Eventually, it has be recognized that not all correlations between data provide evidence of relationships that exist between them, in terms of underlying physical phenomena.

SYMBOLS

$\mathbf{A}$	constant coefficient matrix arising from discretization of the infinite-dimensional model equation, $\mathbf{A} \in R^{N\times N}$
c_i	i^{th} center of RBFN
C_A	concentration of reactant A, kmol/m^3
$\mathbf{c}_K$	vector of POD coefficients, $\mathbf{c}_K \in R^K$
$\mathbf{C}$	POD autocorrelation matrix, $\mathbf{C} \in R^{N\times N}$
D_{eff}	effective diffusion coefficient in porous particle, m^2/s
E	activation energy of chemical reaction, kJ/kmol
f_1, f_2	volume fraction of particle occupied by active sites catalyzing, respectively, first and second step of a generic process A $\leftrightarrow$ B $\leftrightarrow$ C

$\mathbf{F}$ vector of nonlinear functions, $\mathbf{F}: R^N \rightarrow R^N$
ΔH enthalpy of chemical reaction, kJ/kmol
J truncation order of a nonlinear term within POD-DEIM procedure
k_0 frequency coefficient in the Arrhenius equation, 1/s
k_m mass transfer coefficient, m/s
l_f fluidization ratio, $l_f = u/u_{mf}$ where u and u_{mf} denote, respectively, gas velocity and minimum fluidization velocity
Le Lewis number, $\mathrm{Le} = \lambda_{\mathrm{eff}}/(\rho_p c_p D_{\mathrm{eff}})$, where ρ_p and c_p denote, respectively, density of the catalyst particle and its specific heat
K POD truncation order of a state variable
M number of solutions (snapshots) used to construct the POD basis
N number of grid points utilized in the construction of FOM
P pressure, bar
$\mathbf{P}$ matrix of interpolation indices within POD-DEIM procedure, $\mathbf{P} \in R^{N\times J}$
$\mathbf{q}_J$ vector of POD coefficients of the nonlinear term, $\mathbf{q}_J \in R^J$
r Euclidian distance between input x to RBFN and center c_i
r_{A} chemical reaction rate with respect to reactant A, kmol/(m^3·s)
r_p particle radius, m
R universal gas constant, kJ/(kmol·K)
R_i radial basis function
t time, s
T temperature, K or °C
y arbitrary state variable or arbitrary output of RFBN
$\mathbf{y}$ arbitrary vector state variable, $\mathbf{y} \in R^N$
$\tilde{\mathbf{y}}$ truncated state variable vector, $\tilde{\mathbf{y}} \in R^N$
$Y_{\mathrm{CA},r}$ yield of product C with respect to reactant A for generic system of chemical reactions A ↔ B ↔ C evaluated at the reactor level
$\mathbf{Y}$ matrix of solutions (snapshots) used to construct the POD basis, $\mathbf{Y} \in R^{N\times M}$
x arbitrary input to RFBN
X_{A} dimensionless concentration of reactant A, $X_{\mathrm{A}} = C_{\mathrm{A}}/C_{\mathrm{A,ref}}$

Greek symbols

α_q heat transfer coefficient, kW/(m^2·K)
β dimensionless concentration within the catalyst particle, $\beta = C/C_{\mathrm{ref}}$
γ dimensionless activation energy, $\gamma = E/(RT_{\mathrm{ref}})$
δ dimensionless parameter related to heat of chemical reaction, $\delta = C_{\mathrm{A,ref}}(-\Delta H)D_{\mathrm{eff}}/(\lambda_{\mathrm{eff}}T_{\mathrm{ref}})$
$\varepsilon_{K,\mathrm{opt}}$ relative absolute error of the K^{th} order approximation of the optimal solution
ζ dimensionless coordinate in the catalyst pellet, $\zeta = x/r_p$ where x and r_p denote, respectively, radial coordinate and particle radius
θ dimensionless temperature within the catalyst particle, $\theta = T/T_{\mathrm{ref}}$
λ_{eff} effective heat transfer coefficient in the catalyst particle, kW/(m·K)
Λ diagonal matrix containing eigenvalues of the POD modes, $\mathbf{\Lambda} \in R^{N\times N}$
μ arbitrary parameter of the model equation
σ spread parameter of RBFN
τ dimensionless time
φ Thiele modulus of the first-order irreversible reaction, $\Phi = r_p\sqrt{k}/D_{\mathrm{eff}}$ where $k = k_0\exp(-E/RT)$
$\mathbf{\Phi}$ POD basis, $\mathbf{\Phi} \in R^{N\times N}$
$\mathbf{\Phi}_K$ truncated POD basis, $\mathbf{\Phi}_K \in R^{N\times K}$
ψ POD basis function of the nonlinear term
$\mathbf{\Psi}$ POD basis of the nonlinear term, $\mathbf{\Psi} \in R^{N\times N}$
$\mathbf{\Psi}_J$ truncated POD basis of the nonlinear term, $\mathbf{\Psi}_J \in R^{N\times J}$
ω_0 bias of RBFN
ω_i weight associated with the i^{th} connection between neurons within RBFN

Subscripts

g gas phase
init initial conditions
p particle
r reactor
ref reference value
tot total

Abbreviations

ANN artificial neural network
CAD crank angle degree
DEIM discrete empirical interpolation method
FOM full-order model
ODE ordinary differential equation
OTS operator training simulator
PDE partial differential equation
POD proper orthogonal decomposition
RBF radial basis function
RBFN radial basis function network
ROM reduced-order model

REFERENCES

Ali W., Khan M.S., Qyyum M.A., Lee M., 2018. Surrogate-assisted modeling and optimization of a natural-gas liquefaction plant. *Comput. Chem. Eng.*, 118, 132–142. DOI: 10.1016/j.compchemeng.2018.08.003.

Angrisani G., Bizon K., Chirone R., Continillo G., Fusco G., Lombardi S., Marra F.S., Miccio F., Roselli C., Sasso M., Solimene R., Tariello F., Urciolo M., 2013. Development of a new concept solar-biomass cogeneration system. *Energy Convers. Manage.*, 75, 552–560. DOI: 10.1016/j.enconman.2013.06.042.

Appl C., Moser A., Baganz F., Hass V.C., 2020. Digital twins for bioprocess control strategy development and realization, In: Herwig C., Pörtner R., Möller J. (Eds.), *Digital Twins. Advances in Biochemical Engineering/Biotechnology*, 177, 63–94. Springer, Cham. DOI: 10.1007/10_2020_151.

Aris R., 1976. How to get the most out of an equation without really trying. *Educ. Chem. Eng.*, 10, 114–124.

Aris R., 1993. Ends and beginnings in the mathematical modelling of chemical engineering systems. *Chem. Eng. Sci.*, 48, 2507–2517. DOI: 10.1016/0009-2509(93)80263-P.

Babanezhad M., Behroyan I., Nakhjiri A.T., Marjani A., Rezakazemi M., Shirazian S. 2020. High-performance hybrid modeling chemical reactors using differential evolution based fuzzy inference system. *Sci. Rep.*, 10, 21304. DOI: 10.1038/s41598-020-78277-3.

Bárkányi Á, Chován T., Németh S., Abonyi J., 2021. Modelling for digital twins – potential role of surrogate models. *Processes*, 9, 476. DOI: 10.3390/pr9030476.

Bermejo-López A., Pereda-Ayo B., González-Marcos J.A., González-Velasco J.R., 2022. Simulation-based optimization of cycle timing for CO_2 capture and hydrogenation with dual function catalyst. *Catal. Today*, 394–396, 314–324. DOI: 10.1016/j.cattod.2021.08.023.

Bhosekar A., Ierapetritou M., 2018. Advances in surrogate based modeling, feasibility analysis, and optimization: A review. *Comput. Chem. Eng.*, 108, 250–267. DOI: 10.1016/j.compchemeng.2017.09.017.

Biegler L.R., Lang Y.-D., Lin W., 2014. Multi-scale optimization for process systems engineering. *Comput. Chem. Eng.*, 60, 17–30. DOI: 10.1016/j.compchemeng.2013.07.009.

Bizon K., 2016. Autothermicity, multiplicity, yield and selectivity of catalytic processes in a polytropic fluidized bed reactor. *Chem. Eng. J.*, 288, 834–844. DOI: 10.1016/j.cej.2015.12.061.

Bizon K., 2017. Assessment of a POD method for the dynamical analysis of a catalyst pellet with simultaneous chemical reaction, adsorption and diffusion: Uniform temperature case. *Comput. Chem. Eng.*, 97, 259–270. DOI: 10.1016/j.compchemeng.2016.11.009.

Bizon K., 2021. Application of pseudohomogeneous and heterogeneous models in assessing the behavior of a fluidized-bed catalytic reactor. *Energies*, 14, 208. DOI: 10.3390/en14010208.

Bizon K., Continillo G., 2012. Reduced order modelling of chemical reactors with recycle by means of POD-penalty method. *Comput. Chem. Eng.*, 39, 22–32. DOI: 10.1016/j.compchemeng.2011.10.001.

Bizon K., Continillo G., 2021. Efficient optimization of a multifunctional catalytic fixed-bed reactor via reduced-order modeling approach. *Chem. Eng. Res. Des.*, 165, 214–229. DOI: 10.1016/j.cherd.2020.11.007.

Bizon K., Continillo G., Berezowski M., Smuła-Ostaszewska J., 2012. Optimal model reduction by empirical spectral methods via sampling of chaotic orbits. *Physica D: Nonlinear Phenom.*, 241, 1441–1449. DOI: 10.1016/j.physd.2012.05.004.

Bizon K., Continillo G., Lombardi S., Mancaruso E., Vaglieco B.M, 2014. ANN-based virtual sensor for on-line prediction of in-cylinder pressure in a Diesel engine. *Comput. Aided Chem. Eng.*, 33, 763–768. DOI: 10.1016/B978-0-444-63456-6.50128-9.

Bizon K., Continillo G., Mancaruso E., Merola S.S., Vaglieco B.M., 2010. POD-based analysis of combustion images in optically accessible engines. *Combust. Flame*, 157, 632–640. DOI: 10.1016/j.combustflame.2009.12.013.

Bizon K., Continillo G., Mancaruso E., Vaglieco B.M., 2015. Application of RBF neural networks for real-time pressure prediction in a Diesel engine. *Int. J. Eng. Technol.*, 4, 497–508. DOI: 10.14419/ijet.v4i4.4927.

Bizon K., Continillo G., Russo L., Smuła J., 2008. On POD reduced models of tubular reactor with periodic regimes. *Comput. Chem. Eng.*, 32, 1305–1315. DOI: 10.1016/j.compchemeng.2007.06.004.

Bizon K., Skrzypek-Markiewicz K., Pędzich D., Reczek N., 2019. Intensification of catalytic processes through the pellet structuring: Steady-state properties of a bifunctional catalyst pellet applied to generic chemical reactions and the direct synthesis of DME. *Catalysts*, 9, 1020. DOI: 10.3390/catal9121020.

Bose A., Podmore R., Spanel U., 2023. Challenges in operator training: avoiding blackouts in the evolving power grid. *IEEE Power Energy Mag.*, 21, 61–69. DOI: 10.1109/MPE.2023.3247045.

Bradley W., Kim J., Kilwein Z., Blakely L., Eydenberg M., Jalvin J., Laird C., Boukovala F., 2022. Perspectives on the integration between first-principles and data-driven modeling. *Comput. Chem. Eng.*, 166, 107898. DOI: 10.1016/j.compchemeng.2022.107898.

Bremer J., Goyal P., Feng L., Benner P., Sundmacher K., 2017. POD-DEIM for efficient reduction of a dynamic 2D catalytic reactor model. *Comput. Chem. Eng.*, 106, 777–784. DOI: 10.1016/j.compchemeng.2017.02.032.

Broomhead D.S., Lowe D., 1988. Multivariable functional interpolation and adaptive networks. *Complex Syst.*, 2, 321–355.

Carranza-Abaid A., Svendsen H.F., Jakobsen J.P., 2020. Surrogate modelling of VLE: Integrating machine learning with thermodynamic constraints. *Chem. Eng. Sci.: X*, 8, 100080. DOI: 10.1016/j.cesx.2020.100080.

Chakrabarty A., Mannan S., Cagin T., 2016. *Multiscale modeling for process safety applications*. Butterworth-Heinemann, Oxford.

Chaturantabut S., Sorensen D.C., 2010. Non-linear model reduction via discrete empirical interpolation. *SIAM J. Sci. Comput.*, 32, 2737–2764. DOI: 10.1137/090766498.

Chen X., Kurdve M., Johansson B., Despeisse M., 2023. Enabling the twin transitions: Digital technologies support environmental sustainability through lean principles. *Sustain. Prod. Consum.*, 38, 13–27. DOI: 10.1016/j.spc.2023.03.020.

Correia J.B., Abel M., Becker K., 2023. Data management in digital twins: a systematic literature review. *Knowl. Inf. Syst.*, 65, 3165–3196. DOI: 10.1007/s10115-023-01870-1.

Coveney P.V., Dougherty E.R., Highfield R.R., 2016. Big data need big theory too. *Phil. Trans. R. Soc. A*, 374, 20160153. DOI: 10.1098/rsta.2016.0153.

Cutillo E.A., Petito G., Bizon K., Continillo G., 2023. Analysis of an innovative sampling strategy based on *k*-means clustering algorithm for POD and POD-DEIM reduced order models of a 2-D reaction-diffusion system. *Combust. Theory Model*, 27, 508–535. DOI: 10.1080/13647830.2023.2174451.

Danby S.J., Echekki, T., 2005. Proper orthogonal decomposition analysis of autoignition simulation data of nonhomogeneous hydrogen-air mixtures. *Combust. Flame*, 144, 126–138. DOI: 10.1016/j.combustflame.2005.06.014.

de Beer J., Depew C., 2021. The role of process engineering in the digital transformation. *Comput. Chem. Eng.*, 154, 107423. DOI: 10.1016/j.compchemeng.2021.107423.

Dhoke C., Zaabout A., Cloete S., Amini S., 2021. Review on reactor configurations for adsorption-based CO_2 capture. *Ind. Eng. Res.*, 60, 3779–3798. DOI: 10.1021/acs.iecr.0c04547.

Druault P., Guibert P., Alizon F., 2005. Use of proper orthogonal decomposition for time interpolation from PIV data. *Exp. Fluids*, 39, 1009–1023. DOI: 10.1007/s00348-005-0035-3.

Fernandes F.A.N., Lona L.M.F., 2002. Heterogeneous modeling of fluidized bed polymerization reactors. Influence of mass diffusion into the polymer particle. *Comput. Chem. Eng.*, 26, 841–848. DOI: 10.1016/S0098-1354(02)00010-8.

Ferreira J., Pedemonte M., Torres A.I., 2022. Development of a machine learning-based soft sensor for an oil refinery's distillation column. *Comput. Chem. Eng.*, 161, 107756. DOI: 10.1016/j.compchemeng.2022.107756.

Fischer K.L., Freund H., 2020. On the optimal design of load flexible fixed bed reactors: Integration of dynamics into the design problem. *Chem. Eng. J.*, 393, 124722. DOI: 10.1016/j.cej.2020.124722.

Fogleman M., Lumley J., Rempfer D., Haworth D., 2004. Application of the proper orthogonal decomposition to datasets of internal combustion engine flows. *J. Turbul.*, 5, N23. DOI: 10.1088/1468-5248/5/1/023.

Froment G.F., 1974. Analysis and design of fixed bed catalytic reactors, In: Bischoff K.B. (Ed.), *Chemical Reaction Engineering*. American Chemical Society, Washington, D.C., 1–55. DOI: 10.1021/ba-1972-0109.ch001.

Galeazzi A., Prifti K., Cortellini C., Di Pretoro A., Gallo F., Manenti F., 2023. Development of a surrogate model of an amine scrubbing digital twin using machine learning methods. *Comput. Chem. Eng.*, 174, 108252. DOI: 10.1016/j.compchemeng.2023.108252.

Graham M.D., Kevrekidis I.G., 1996. Alternative approaches to the Karhunen-Loève decomposition for model reduction and data analysis. *Comput. Chem. Eng.*, 20, 495–506. DOI: 10.1016/0098-1354(95)00040-2.

Hesthaven J.S., Gottlieb S., Gottlieb D., 2007. *Spectral methods for time-dependent problems*. Cambridge University Press, Cambridge. DOI: 10.1017/CBO9780511618352.

Holmes P., Lumley J.L., 1996. *Turbulence, coherent structures, dynamical systems and symmetry*. Cambridge University Press, Cambridge.

Hou C.K.J.., Behdinan. K., 2022. Dimensionality reduction in surrogate modeling: A review of combined methods. *Data Sci. Eng.*, 7, 402–427. DOI: 10.1007/s41019-022-00193-5.

Jayaweera C., Groot N., Meul S., Verliefde A., Nopens I., Hitsov I., 2022. Development of a hybrid model for reliably predicting the thermal performance of direct contact countercurrent cooling towers. *Int. J. Heat Mass Transf.*, 197, 123336. DOI: 10.1016/j.ijheatmasstransfer.2022.123336.

Jo T., Koo B., Kim H., Lee D., Yoon J.Y., 2019. Effective sensor placement in a steam reformer using gappy proper orthogonal decomposition. *Appl. Therm. Eng.*, 154, 419–432. DOI: 10.1016/j.applthermaleng.2019.03.089.

Karhunen K., 1947. Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae*, 37, 3–79.

Kerschen G., Feeny B.F., Golinval J.-C., 2003. On the exploitation of chaos to build reduced-order models. *Comput. Methods Appl. Mech. Eng.*, 192, 1785–1795. DOI: 10.1016/S0045-7825(03)00206-8.

Kestering D., Agbleze S., Bispo H., Lima F.V., 2023. Model predictive control of power plant cycling using Industry 4.0 infrastructure. *Digital Chem. Eng.*, 7, 100090. DOI: 10.1016/j.dche.2023.100090.

Kosambi D., 1943. Statistics in function space. *J. Ind. Math. Soc.*, 7, 76–88.

Koulouris A., Misailidis N., Petrides D., 2021. Applications of process and digital twin models for production simulation and scheduling in the manufacturing of food ingredients and products. *Food Bioprod. Process.*, 126, 317–333. DOI: 10.1016/j.fbp.2021.01.016.

Krippl M., Kargl T., Duerkop M., Dürauer A., 2021. Hybrid modeling reduces experimental effort to predict performance of serial and parallel single-pass tangential flow filtration. *Sep. Purif. Technol.*, 276, 119277. DOI: 10.1016/j.seppur.2021.119277.

Kurtz L.A., Smith R.E., Parks C.L., Boney L.R., 1978. A comparison of the method of lines to finite difference techniques in solving time-dependent partial differential equations. *Comput. Fluids*, 6, 49–70. DOI: 10.1016/0045-7930(78)90007-5.

Kurzina I.A., Reshetinikov S.I., Karakchieva N.I., Kurina L.N., 2017. Direct synthesis of dimethyl ether from synthesis gas: Experimental study and mathematical modeling. *Chem. Eng. J.*, 329, 135–141. DOI: 10.1016/j.cej.2017.04.132.

Li S., Duan G., Sakai M., 2022a. Development of a reduced-order model for large-scale Eulerian-Lagrangian simulations. *Adv. Powder. Technol.*, 33, 103632. DOI: 10.1016/j.apt.2022.103632.

Li T., Duan G., Sakai M., 2022b. POD-based identification approach for powder mixing mechanism in Eulerian-Lagrangian simulations. *Adv. Powder Technol.*, 33, 103364. DOI: 10.1016/j.apt.2021.11.016.

Li T., Han D., Yu B., Li J., Sun D., 2020. Study on a POD reduced-order model for steady-state flows in fractured porous media. *Int. Commun. Heat Mass Transfer*, 112, 104489. DOI: 10.1016/j.icheatmasstransfer.2020.104489.

Loève M., 1945. Fonctions aléatoires du second ordre. *C. R. Acad. Sci.*, 220, 469.

Lunev A., Lauerer A., Zborovskii V., Léonard F., 2022. Digital twin of a laser flash experiment helps to assess the thermal performance of metal foams. *Int. J. Therm. Sci.*, 181, 107743. DOI: 10.1016/j.ijthermalsci.2022.107743.

Ma X., Karniadakis G.E., Park H., Gharib M., 2002. DPIV/T-driven convective heat transfer simulation. *Int. J. Heat Mass Transfer*, 45, 3517–3527. DOI: 10.1016/S0017-9310(02)00071-6.

Makrygiorgos G., Maggioni G.M., Mesbah A., 2020. Surrogate modeling for fast uncertainty quantification: Application to 2D population balance models. *Comput. Chem. Eng.*, 138, 106814. DOI: 10.1016/j.compchemeng.2020.106814.

Marra F.S., Bizon K., Continillo G., 2015. Fast predictive model for design and optimization of a low emissions biomass fueled

CHP system. *23rd European Biomass Conference and Exhibition*. Vienna, Austria, 1–4 June 2015.

Masoudi R., McPhee J., 2021. Application of Karhunen-Loéve decomposition and piecewise linearization to a physics-based battery model. *Electrochim. Acta*, 365, 137093. DOI: 10.1016/j.electacta.2020.137093.

Matsunami K., Miura T., Yaginuma K., Tanabe S., Badr S., Sugiyama H., 2023. Surrogate modeling of dissolution behavior toward efficient design of tablet manufacturing process. *Comput. Chem. Eng.*, 171. 108141. DOI: 10.1016/j.compchemeng.2023.108141.

McKay A., Ghosh D., Zhu L., Xi L., Mhaskar P., 2022. A novel linear hybrid model predictive control design: application to a fed batch crystallization process. *Digital Chem. Eng.*, 3, 100033. DOI: 10.1016/j.dche.2022.100033.

McLaughlin D., Bierling A., Mayerhöfer B., Schmid G., Thiele S., 2023. Digital twin of a hierarchical CO_2 electrolyzer gas diffusion electrode. *Adv. Funct. Mater.*, 33, 2212462. DOI: 10.1002/adfm.202212462.

McLean K.A.P., McAuley K.B., 2012. Mathematical modelling of chemical processes – obtaining the best model predictions and parameter estimates using identifiability and estimability procedures. *Can. J. Chem. Eng.*, 90, 351–366. DOI: 10.1002/cjce.20660.

Moghadam S.M.A., Loosen S.L., Meinke M., Schröder W., 2020. Reduced-order analysis of the acoustic near field of a ducted axial fan. *Int. J. Heat Fluid Flow*, 85, 108657. DOI: 10.1016/j.ijheatfluidflow.2020.108657.

Moioli E., 2022. Process intensification and energy transition: A necessary coupling? *Chem. Eng. Process. Process Intensif.*, 179, 109097. DOI: 10.1016/j.cep.2022.109097.

Oztruk M., Dincer I., 2021. A comprehensive review on power-to-gas with hydrogen options for cleaner applications. *Int. J. Hydrogen Energy*, 46, 31511–31522. DOI: 10.1016/ j.ijhyd ene. 2021.07.066.

Padhi R., Balakrishnan S.N., 2003. Proper orthogonal decomposition based optimal neurocontrol synthesis of a chemical reactor process using approximate dynamic programming. *Neural Networks*, 16, 719–728. DOI: 10.1016/S0893-6080(03)00131-X.

Palys M.J., Daoutidis P., 2022. Power-to-X: A review and perspective. *Comput. Chem. Eng.*, 165, 107948. DOI: 10.1016/j.compchemeng.2022.107948.

Park H.M., Cho D.H., 1996. Low dimensional modelling of flow reactors. *Int. J. Heat Mass Transfer*, 39, 3311–3323. DOI: 10.1016/0017-9310(96)00038-5.

Patle D.S., Ahmad Z., Rangaiah G.P., 2014. Operator training simulators in the chemical industry: review, issues, and future directions. *Rev. Chem. Eng.* 30, 199–216. DOI: 10.1515/revce-2013-0027.

Pearson K., 1901. On lines and planes of closest fit to systems of points in space. *Philos. Mag.*, 2, 559–572.

Pergam P., Briesen H., 2023. Reduced order modeling for compressible cake filtration process using proper orthogonal decomposition. *Comput. Chem. Eng.*, 171, 108165. DOI: 10.1016/j.compchemeng.2023.108165.

Pinto J., Mestre M., Ramos J., Costa R.S., Striedner G.S., Oliveira R., 2022. A general deep hybrid model for bioreactors systems: Combining first principles with deep neural networks. *Comput. Chem. Eng.*, 165, 107952. DOI: 10.1016/j.compchemeng.2022.107952.

Procacci A., Amaduzzi R., Coussement A., Parente A., 2023. Adaptive digital twins of combustion systems using sparse sensing strategies. *Proc. Combust. Inst.*, 39, 4257–4266. DOI: 10.1016/j.proci.2022.07.029.

Rogers A.W., Song Z., Ramon F.V., Jing K., Zhang D., 2023. Investigation 'greyness' of hybrid model for bioprocess predictive modeling. *Biochem. Eng. J.*, 190, 108761. DOI: 10.1016/j.bej.2022.108761.

Rout K.R., Jakobsen H.A., 2015. A numerical study of fixed bed reactor modelling for steam methane reforming process. *Can. J. Chem. Eng.*, 93, 1222–1238. DOI: 10.1002/cjce.22202.

Sánchez-Contador M., Ateka A., Ibáñez, M., Bilbao, J., Aguayo A.T., 2019. Influence of the operating conditions on the behavior and deactivation of a CuO-ZnO-ZrO_2@SAPO-11 core-shell-like catalyst in the direct synthesis of DME. *Renew. Energy*, 138, 585–597. DOI: 10.1016/j.renene.2019.01.093.

Sansana J., Joswiak M.N., Castillo I., Wang Z., Rendall R., Chiang L.H., Reis M.S., 2021. Recent trends on hybrid modeling for Industry 4.0. *Comput. Chem. Eng.*, 151, 107365. DOI: 10.1016/j.compchemeng.2021.107365.

Segala D.B., Naseradinmousavi P., 2017. On the inclusion of time derivatives of state variables for parametric model order reduction for a beam on a nonlinear foundation. *J. Dyn. Sys. Meas. Control.*, 139, 081009. DOI: 10.1115/1.4035759.

Shoji K., Schudel S., Onwude D., Shirvastava C., Defraeye T., 2022. Mapping the postharvest life on imported fruits from packhouse to retail stores using physics-based digital twins. *Resour. Conserv. Recycl.*, 176, 105914. DOI: 10.1016/j.resconrec.2021.105914.

Siddiqui M.S., Latif S.T.M., Saeed M., Rahman M., Badar A.W., Hasan S.M., 2020. Reduced order model of offshore wind turbine wake by proper orthogonal decomposition. *Int. J. Heat Fluid Flow*, 82, 108554. DOI: 10.1016/j.ijheatfluidflow.2020.108554.

Singh M., Fuenmayor E., Hinchy E.P., Qiao Y., Murray N., Devine D., 2021. Digital twin: Origin to future. *Appl. Syst. Innov.*, 4, 36. DOI: 10.3390/asi4020036.

Sirisup S., Karniadakis G.E., 2005. Stability and accuracy of periodic flow solutions obtained by a POD-penalty method. *Physica D: Nonlinear Phenom.*, 202, 218–237. DOI: 10.1016/j.physd.2005.02.006.

Sirovich L., 1987. Turbulence and the dynamics of coherent structures. I. Coherent structures. *Quart. Appl. Math.*, 45, 561–590. DOI: 10.1090/qam/910462.

Studer S., Bui T.B., Drescher C., Hanuschkin A., Winkler L., Peters S., Müller K.-R., 2021. Towards CRISP-ML(Q): A machine learning process model with quality assurance methodology. *Mach. Learn. Knowl. Extr.*, 3, 392–413. DOI: 10.3390/make3020020.

Styring P., Duckworth E.L., Platt E.G., 2021. Synthetic fuels in a transport transition: fuels to prevent a transport

underclass. *Front. Energy Res.*, 9, 707867. DOI: 10.3389/fenrg.2021.707867.

Sun Y., Yang J., Wang Y., Li Z., Ma Y., 2020. A POD reduced-order model for resolving the neutron transport problem of nuclear reactor. *Ann. Nucl. Energy*, 149, 107799. DOI: 10.1016/j.anucene.2020.107799.

Sun Z., West D.H., Balakotaiah V., 2019. Bifurcation analysis of catalytic partial oxidations in laboratory-scale packed-bed reactors with heat exchange. *Chem. Eng. J.*, 377, 119765. DOI: 10.1016/j.cej.2018.08.151.

Swischuk R., Mainini L., Peherstorfer B., Willcox K., 2018. Projection-based model reduction: Formulations for physics-based machine learning. *Comput. Fluids*, 179, 704–717. DOI: 10.1016/j.compfluid.2018.07.021.

van Rooij F., Scarf P., Do P., 2021. Planning the restoration of membranes in RO desalination using a digital twin. *Desalination*, 519, 115214. DOI: 10.1016/j.desal.2021.115214.

Viggiano B., Skjæraasen O., Schümann H., Tutkun M., Cal R.B., 2018. Characterization of flow dynamics and re duce d-order description of experimental two-phase pipe flow. *Int. J. Multiph. Flow*, 105, 91–101. DOI: 10.1016/j.ijmultiphaseflow.2018.03.020.

Villadsen J., Michelsen M.L., 1978. *Solution of differential equation models by polynomial approximation*. Prentice Hall, Englewood Cliffs.

Vinokur M., 1989. An analysis of finite-difference and finite-volume formulations of conservation laws. *J. Comput. Phys.*, 81, 1–52. DOI: 10.1016/0021-9991(89)90063-6.

von Stosch M., Oliveira R., Peres J., de Azevedo S.F., 2014. Hybrid semiparametric modeling in process systems engineering: Past, present and future. *Comput. Chem. Eng.*, 60, 86–101. DOI: 10.1016/j.compchemeng.2013.08.008.

Wang F., Song Y., Liu C., He A., Qiang T., 2023. Multi-objective optimal scheduling of laminar cooling water supply system for hot rolling mills driven by digital twin for energy-saving. *J. Process Control*, 122, 134–146. DOI: 10.1016/j.jprocont.2023.01.004.

Wang G.G., Shan S., 2007. Review on metamodeling techniques in support of engineering design optimization. *J. Mech. Des.*, 129, 370–380. DOI: 10.1115/1.2429697.

Wang L.G., Ge R., Chen X., Zhou R., Chen H.-M., 2021. Multi-scale digital twin for particle breakage in milling: From nanoindentation to population balance model. *Powder Technol.*, 386, 247–261. DOI: 10.1016/j.powtec.2021.03.005.

Williams B., Cremaschi S., 2021. Selection of surrogate modeling techniques for surface approximation and surrogate-based optimization. *Chem. Eng. Res. Des.*, 170, 76–89. DOI: 10.1016/j.cherd.2021.03.028.

Wright L., Davidson S., 2020. How to tell the difference between a model and a digital twin. *Adv. Model. Simul. Eng. Sci.*, 7, 13. DOI: 10.1186/s40323-020-00147-4.

Xing L., Jiang H., Wang S., Pinfield V.J., Xuan J., 2023. Data-driven surrogate modelling and multi-variable optimization of trickle bed and packed bubble column reactors for CO_2 capture via enhanced weathering. *Chem. Eng. J.*, 454, 139997. DOI: 10.1016/j.cej.2022.139997.

Yang M., Armaou A., 2018. Revisiting APOD accuracy for nonlinear control of transport reaction process: A spatially discrete approach. *Chem. Eng. Sci.*, 181, 146-158. DOI: 10.1016/j.ces.2017.12.037.

Yang S., Navarathna P., Ghosh S., Bequette B.W., 2020. Hybrid modeling in the era of smart manufacturing. *Comput. Chem. Eng.*, 140, 106874. DOI: 10.1016/j.compchemeng.2020.106874.

Yu J., Liu P. Li Z., 2022. Data reconciliation-based simulation on thermal power plants for performance estimation and digital twin development. *Comput. Chem. Eng.*, 168, 108063. DOI: 10.1016/j.compchemeng.2022.108063.

Zendehboudi S., Rezaei N., Lohi A., 2018. Applications of hybrid models in chemical, petroleum, and energy systems: A systematic review. *Appl. Energy*, 228, 2539–2566. DOI: 10.1016/j.apenergy.2018.06.051.