Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA IN

INGEGNERIA CIVILE, CHIMICA, AMBIENTALE E DEI MATERIALI

Ciclo 36

**Settore Concorsuale:** 08/A1 - IDRAULICA, IDROLOGIA, COSTRUZIONI IDRAULICHE E MARITTIME

**Settore Scientifico Disciplinare:** ICAR/01 - IDRAULICA

# DECODING COMPLEX FLOW PHENOMENA VIA DYNAMIC MODE DECOMPOSITION

**Presentata da:** Giulia Libero

**Coordinatore Dottorato**

Alessandro Tugnoli

**Supervisore**

Valentina Ciriello

**Co-supervisore**

Daniel Tartakovsky

**Esame finale anno 2024**

# Abstract

In this thesis, the viability of the Dynamic Mode Decomposition (DMD) as a technique to analyze and model complex dynamic real-world systems is presented. This method derives, directly from data, computationally efficient reduced-order models (ROMs) which can replace too onerous or unavailable high-fidelity physics-based models. Optimizations and extensions to the standard implementation of the methodology are proposed, investigating diverse case studies related to the decoding of complex flow phenomena. The flexibility of this data-driven technique allows its application to high-fidelity fluid dynamics simulations, as well as time series of real systems observations. The resulting ROMs are tested against two tasks: (i) reduction of the storage requirements of high-fidelity simulations or observations; (ii) interpolation and extrapolation of missing data. The capabilities of DMD can also be exploited to alleviate the cost of onerous studies that require many simulations, such as uncertainty quantification analysis, especially when dealing with complex high-dimensional systems. In this context, a novel approach to address parameter variability issues when modeling systems with space and time-variant response is proposed. Specifically, DMD is merged with another model-reduction technique, namely the Polynomial Chaos Expansion, for uncertainty quantification purposes. Useful guidelines for DMD deployment result from the study, together with the demonstration of its potential to ease diagnosis and scenario analysis when complex flow processes are involved.

**Alma Mater Studiorum - Università di Bologna**

PhD in Civil, Chemical, Environmental, and Materials Engineering

# Decoding complex flow phenomena via Dynamic Mode Decomposition

A dissertation submitted for the degree of
*Doctor of Philosophy*

Candidate:

**Giulia Libero**

Advisor:

**Prof. Valentina Ciriello**

PhD Coordinator:

**Prof. Alessandro Tugnoli**

Co-Advisor:

**Prof. Daniel M. Tartakovsky**

2023

*“Je suis la nature sans la pouvoir saisir, et puis, cette rivière qui baisse, remonte, un jour verte, puis jaune, tantôt à sec, et qui demain sera un torrent.”*

*Claude Monet*

# Contents

# Abstract

In this thesis, the viability of the Dynamic Mode Decomposition (DMD) as a technique to analyze and model complex dynamic real-world systems is presented. This method derives, directly from data, computationally efficient reduced-order models (ROMs) which can replace too onerous or unavailable high-fidelity physics-based models. Optimizations and extensions to the standard implementation of the methodology are proposed, investigating diverse case studies related to the decoding of complex flow phenomena. The flexibility of this data-driven technique allows its application to high-fidelity fluid dynamics simulations, as well as time series of real systems observations. The resulting ROMs are tested against two tasks: (i) reduction of the storage requirements of high-fidelity simulations or observations; (ii) interpolation and extrapolation of missing data. The capabilities of DMD can also be exploited to alleviate the cost of onerous studies that require many simulations, such as uncertainty quantification analysis, especially when dealing with complex high-dimensional systems. In this context, a novel approach to address parameter variability issues when modeling systems with space and time-variant response is proposed. Specifically, DMD is merged with another model-reduction technique, namely the Polynomial Chaos Expansion, for uncertainty quantification purposes. Useful guidelines for DMD deployment result from the study, together with the demonstration of its potential to ease diagnosis and scenario analysis when complex flow processes are involved.

# 1 | Introduction

The world around us constantly evolves, in an endless chain of causes and effects, and humans are asked to face new challenges to adapt and secure a liveable and sustainable future for all within this complex system.

Water resources management is a central issue in this framework. Although approximately 70% of our planet is made up of water, the majority of this resource is stored in the oceans, and the remaining freshwater supply is not adequately distributed in space and time to meet all current and foreseeable water demands (Gourbesville, 2008). This has evident impacts on agricultural productivity and energy production, limiting economic development and having undeniable social implications. The lack of adequate clean water is moreover a constraint on human health as well as on the maintenance of a clean environment and healthy ecosystems (Cosgrove and Loucks, 2015).

Climate change now occurring makes this picture even more alarming. As described in the latest report of the IPCC (2023b), widespread adverse impacts and related losses and damages to nature and people demonstrate that climate change is an evident threat to human well-being and planetary health. Food and water security (see Simpson and Jewitt (2019) for details on the Water-Energy-Food Nexus concept) have already been affected by global warming, changing precipitation patterns, reduction of cryospheric elements, and greater frequency and intensity of climatic extremes. Adverse impacts have been reported on human physical and mental health, livelihoods, and infrastructure in urban settings. Furthermore, economic and societal impacts are affecting people and systems disproportionately across sectors and regions, contributing to humanitarian crises and conflicts.

As any further delay can result in compounding and cascading risks across sectors and regions, global actions are urgently necessary (IPCC, 2023b). Specifically, responding to climate change involves a two-pronged approach: on one hand, we need to reduce emissions of and stabilize the levels of greenhouse gases in the atmosphere (mitigation) (IPCC, 2014b); while on the other, we have to design strategies to adapt to present and future impacts of climate change (adaptation) (IPCC, 2014a). At the same time, all the strategies we define should be carefully conceived to guarantee equality and sustainability worldwide.

Thus, the first goal we need to pursue is limiting human impact on climate and environment, to avoid an escalation of negative effects (mitigation). Over the past century, as the global population quadrupled and economic production increased about 20-fold, to meet the growing demand for resources and food we have intensively exploited the environment causing depletion of resources, contamination, and impacting climate (Steer, 2014). Past patterns of energy and natural resource use, agricultural practices, and urbanization proved largely unsustainable. Now we need to reduce the human footprint (Mu et al., 2022) and move to a more sustainable model of development. To this aim, innovative strategies and technologies are designed and tested to reduce resource consumption and limit emissions or production of contaminant substances. Notable efforts involve the energy sector, where research focuses both on the sustainable exploitation of resources (Hilborn et al., 1995) and the design of technologies to achieve carbon neutrality and global greenhouse gas emissions reduction (Wang et al., 2021; Fankhauser et al., 2021).

As we are already committed to some level of climate change, the second focal point is the development of strategies and technologies to adapt to existing and future negative impacts (Herman et al., 2020; Füssel, 2007). Adaptation seeks to facilitate adjustment to actual or expected climate and its effects and moderate or avoid harm or exploit beneficial opportunities. It's not only necessary to adjust our way of living to thrive in a new environment, but we also need to enhance our resilience. This means being able to anticipate and manage stress and shocks and recover from their impacts quickly and effectively. In the past, we have disrupted and overallocated river flow regimes, overdrawn groundwater aquifers, and polluted many water bodies. This amplifies our vulnerability to extreme weather events, such as drought and floods. As the intensity and frequency of these phenomena increase due to climate change (Swain et al., 2020), we need to improve the resilience of infrastructures and water management strategies to defend ourselves from the negative consequences. Appropriate actions should be designed to protect urban areas against flood damage and at the same time to guarantee water supplies even during long and intense drought periods (Baldassarre et al., 2017). For this reason, the civil and environmental engineering sector is carefully focused on the observation of present and past occurrences, to derive predictions of future events and draw adaptation solutions. Measurements of phenomena are collected and experimental simulations are performed to investigate systems behavior and evaluate alternative solutions (Herman et al., 2020; Hallegatte, 2009). The challenges we need to face are not limited to the quantitative availability of water but also refer to the quality of the resource. Contamination, indeed, further exacerbates the critical conditions in which water bodies find themselves and deeply influences the health of the surrounding ecosystems, reflecting on food security and human health (Lu et al., 2015). To limit the negative effects, such as disease spreading and loss in food production, and guarantee a safe, sustainable, and prosperous future, effective remediation interventions to preserve and restore water quality are being designed. To derive robust adaptation strategies growing attention is being paid also to the interactions between water and ecological systems (Rinaldo and Rodriguez-Iturbe, 2022; D'Odorico et al., 2010). Modeling and interpretation of these connections involve many diverse sectors, ranging from chemistry to biology, from hydrology to sociology. Since the warming climate is already producing effects on human body function (Patz et al., 2005; Mora et al., 2022), and the freshwater shortage is closely linked to malnutrition and poor hygiene conditions (Lin et al., 2022; Howard et al., 2016), even the medical sector is involved. The focus in this case is on the diagnosis of dysfunctions, and on the development of cheap and efficient treatments for widely and quickly spreading diseases.

The third key element in the definition of efficient and sustainable responses to global water challenges is ensuring universal and equitable access to water for all (WSH, 2014; Cetrulo et al., 2020). Climate change is expected to emphasize the differences in water availability around the globe, majorly affecting developing countries, while the global population continues its growth (Cappelli et al., 2021). To avoid humanitarian crises and conflicts (Kreamer, 2012; Nandalal and Simonovic, 2003), policies and actions should fight these disproportions and inequalities, encouraging integrated water resources management at all levels, and transboundary cooperation.

The design and development of efficient technologies and strategies to support all these urgent actions require us to deal with a highly intricate framework of acting variables and phenomena (Leavesley, 1994). Specifically, we are commonly asked to model dynamical real-world systems, which are usually highly complex to interpret, generally nonlinear, and with a multi-scale behavior in both space and time (Phan et al., 2021). The dense network of interactions between different phenomena and variables further complicates the tasks, introducing uncertainty and variability in the framework (Herman et al., 2020; Hallegatte, 2009). For this reason, the direct use of physical laws or governing equations to reproduce complex dynamical systems' behavior is often an elusive task.

Nowadays, the vast amount of data collected in all scientific fields, either directly through observation or using simulations, together with the development of efficient techniques to handle them, offers new opportunities for analyzing dynamic systems. In many research sectors, ranging from the medical (Pegolotti et al., 2023) to the environmental one (Song and Tartakovsky, 2021), data-driven models are emerging as powerful tools to overcome computational problems and manage uncertainty when complexity and high dimensions make traditional modeling impracticable (Tartakovsky et al., 2020; Vespignani, 2011). The primary goal of data-driven modeling is building efficient and computationally cheap models for future state prediction, but also interpreting and identifying physical insights on the phenomena directly from data.

Computational efficiency is a crucial feature in complex system modeling, especially when large numbers of simulations are required to solve time-dependent problems or to explore the impacts of parameter variability on the model response (e.g., uncertainty quantification analysis). Order-reduction methods stand out for their specific ability to reduce the computational burden in case of high-dimensional problems (Kutz et al., 2016). These techniques typically capture the key features of the underlying dynamics of a complex system directly from high-fidelity simulations or observational data and then employ these elements to derive an efficient reduced-order model (ROM) of the system state (Pegolotti et al., 2021; Ciriello et al., 2019; Oladyshkin et al., 2012). The specific model-reduction technique can be selected to, e.g., relate the key governing parameters in input and the model response in output (like response surface methods) or to approximate dynamic system behavior with simple but accurate mathematical relationships. Its simple form reduces the computational cost and allows the generation of a large number of simulations while preserving the high-fidelity representation of the phenomena (Focaccia et al., 2021; Ciriello et al., 2017, 2012; Oladyshkin et al., 2012).

The Dynamic Mode Decomposition (DMD) method is a promising data-driven technique based on Singular Value Decomposition (SVD) for the generation of ROMs of complex dynamical systems

(Schmid et al., 2010). It originated in the fluid dynamics community as a technique to identify spatio-temporal coherent structures from high-dimensional data (Rowley et al., 2009; Schmid et al., 2010; Schmid, 2010; Bagheri, 2013; Pan et al., 2021; Liu et al., 2021; Ohh and Spedding, 2022). In this field, many optimizations have been developed to adapt the algorithm to several different contexts and applications (Kutz et al., 2016; Schmid, 2022). Later, its data-based, equation-free, and flexible structure encouraged its spreading in very different environments (Brunton et al., 2016; Proctor and Eckhoff, 2015; Mann and Kutz, 2016; Erichson et al., 2016). The method seeks to compute, by means of the SVD, the best-fit linear operator which relates time-shifted snapshots of the state of the system. The DMD-based ROM generation is purely based on data, equally valid for experimental and numerical datasets. Hence, major advantages can derive from its application when traditional physics-based models are too complex to design and employ, but a large number of observations of the system evolution in time are available. Physical interpretability of the structure of DMD-based models, guaranteed by the connection of DMD with Koopman operator theory (Rowley et al., 2009; Mezić, 2013), is a further advantage. Indeed, DMD-based models can be used not only as faster and cheaper ROMs but their structure can also be interpreted to discover patterns and physical features in the phenomena.

In this research, the employment of the DMD method with respect to the task of modeling complex flow phenomena is presented. The use of DMD is promising in the diagnosis of complex hydraulic systems ranging from water resources flow and transport in natural domains to fluid mechanics applications. Thanks to its flexibility, DMD can be applied to high-fidelity fluid dynamics simulations, as well as time series of real systems observations, to obtain computationally efficient ROMs. These can replace too complex or unavailable high-fidelity physics-based models when onerous scenario and uncertainty analysis are needed, reducing the computational cost and speeding up the development of appropriate solutions. Specifically, we present the use of DMD-based ROMs to accomplish two tasks: (i) reduction of the storage requirements of high-fidelity simulations or observations; and (ii) interpolation and extrapolation of missing data. To demonstrate the usefulness of the technique with respect to these objectives, two different innovative applications of DMD in the fluid dynamic context are presented. To the best of our knowledge, we are the first to apply an extended version of the DMD algorithm to 3D simulations of a complex dynamic system, and then to a time series of satellite global observations. Moreover, we propose a novel approach to address parameter variability issues when modeling systems with space and time-variant responses. Here, DMD is exploited to alleviate the calibration cost of the procedure, and hence make onerous analysis, such as uncertainty quantification, practicable.

The Thesis is organized as follows. In Chapter 2 we introduce the mathematical definition of dynamical systems and we present the DMD method. We describe the standard algorithm and an extended version adapted to deal with possible problem inhomogeneity. We propose an optimized implementation of the algorithms and we discuss the key role played by the selection of an appropriate SVD truncation rank to guarantee efficiency, accuracy, and generalizability of the model. The second part of the Chapter is devoted to the description of another model reduction technique that is a response surface method: the Polynomial Chaos Expansion (PCE); we describe the algorithm, prove its usefulness by briefly presenting results from the first of our papers, and conclude by discussing the efficiency limits of PCE when applied to dynamical systems. This introduction to the PCE method serves as a base for the development of an integrated DMD-PCE framework for the robust reduction of models with time-variant response that is presented in Chapter 5. A common thread guides us from this methodological description to the different applications involving complex flow phenomena proposed in the following chapters.

In Chapter 3 we investigate the viability of DMD as a method to reduce the computational cost and storage/memory requirements related to fluid dynamics simulations. Applying the extended DMD (xDMD) algorithm for the first time to a set of three-dimensional cardiovascular simulations, we test the efficacy and robustness of the method and explore the influence of SVD truncation on accuracy. Given an optimal low-rank truncation, which balances accuracy, generalizability, and computational saving, the xDMD proves to be a successful tool for data compression and reconstruction of the system's dynamics. This could help speed up the modeling phase, with positive impacts on the diagnosis and the optimization of intervention choice.

In Chapter 4 we analyze the ability of DMD to derive accurate ROMs directly from high-dimensional datasets of measurements/observations of the real system behavior. Differently from the previous application, here no high-fidelity model is available, and heavenly-spaced simulations are replaced by time series of observations, which may present gaps and suffer from measurement errors and biases. Specifically, the analysis is performed on a novel application of DMD to satellite data concerning water distribution on Earth. We test the performances of the standard (sDMD) and the extended (xDMD) algorithms on the task of reconstruction and extrapolation of data, demonstrating that DMD could be an innovative technique to decode precious satellite mission products.

In Chapter 5 we propose the combination of the capabilities of DMD and PCE into an integrated novel approach to address parameter variability issues when modeling systems with space-time variant response. The DMD method is a powerful tool to reduce the computational cost of dynamic systems modeling. Still, it is not suitable to explore the system response in the space of variability of parameters. On the other hand, the PCE technique is widely employed to address parameters' variability issues, but the huge number of simulations required for training inhibits its use in case of time-space variant problems. Here, we propose a DMD-informed PCE, where part of the high-fidelity simulations required to calibrate the polynomial model is replaced by low-cost DMD reconstructions. The novel approach is tested against a scenario of a non-linear 2D multiphase flow in a heterogeneous media, considering different replacement rates and employing either the standard (sDMD) or the extended (xDMD) version of the method for the interpolation task. Thanks to its computational efficiency and accuracy the proposed framework could ease onerous studies such as uncertainty quantification analysis.

Remarks from each application are collected at the end of each chapter and then reinterpreted in a set of general conclusions and future perspectives in Chapter 6 which closes the thesis.

# 2 | Methods

Dynamical systems allow us to represent the evolution of a quantity of interest (QoI) over time, as a consequence of unsteady phenomena. However, complexity, non-linearity, and the multi-scale behavior in space and time, often prevent the systematic use of physically-based high-fidelity models (HFMs) to investigate the behavior of the QoI. Nowadays, the large amount of data available and the development of efficient techniques to handle these data offer new opportunities to analyze and understand complex dynamical systems. Data-driven methods, in particular, are emerging as powerful tools to derive physical insights on the phenomena directly from data and build efficient surrogate models for future state prediction. Computational efficiency is crucial when large numbers of simulations are required to explore how the variability of the input parameters impacts the model response (e.g., uncertainty quantification analysis). Hence, order-reduction techniques can be employed to reduce the computational burden of the analysis in case of high-dimensional problems.

In this chapter, we present the Dynamic Mode Decomposition (DMD) method, a promising data-driven technique based on Singular Value Decomposition (SVD) for the generation of reduced-order models (ROMs) of complex dynamical systems. DMD is a data-based, equation-free method whose flexible structure encourages its use in different fields, primarily where traditional physically-based models are too complex to employ and a large amount of data describing the temporal and spatial evolution of the system is available. We describe the algorithm and introduce optimizations for its implementation. We also propose criteria for appropriate matrix truncation and metrics to explore the method performances.

Besides the DMD, the featured method of the whole project, a second technique is presented: the Polynomial Chaos Expansion (PCE). As it will be discussed in Chapter 5, in fact, the attitude of DMD can be combined with the capabilities of PCE to approximate the system's response in the space of variability of the parameters to obtain a new robust framework for the reduction of high-dimensional models with time-variant response.

The chapter structure is as follows: in Section 2.1 we introduce the mathematical definition of dynamical systems; Section 2.2 is devoted to presenting the DMD, describing a standard and an extended algorithm to implement the method, and finally discussing performances and proposing optimizations; in Section 2.3 the focus is on the PCE technique, with the description of the algorithm and proof of its usefulness through the results from a paper of ours, and the discussion of its limits when applied to dynamical systems.

## 2.1 Dynamical Systems

Consider a dynamical system

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{f}(\mathbf{x}(t), t; \mu), \tag{2.1}$$

where $\mathbf{x}(t)$ is a vector representing the state of the system at time t, $\mu$ contains parameters of the system, and $\mathbf{f}(\mathbf{x}, \cdot)$ represents the dynamics.

In general, $\mathbf{x}(t) \in \mathbf{M}$ is an n-dimensional state that lives on a smooth manifold $\mathbf{M}$, and $\mathbf{f}$ is an element of the tangent bundle $\mathbf{TM}$ of $\mathbf{M}$ so that $\mathbf{f}(\mathbf{x}(t)) \in \mathbf{T}_{\mathbf{x}(t)}\mathbf{M}$. However, we typically consider the simpler case where $\mathbf{x}(t) \in \mathbf{M} = \mathbb{R}^N$ is a vector of the system state discretized on a numerical mesh of $N$ elements, and $\mathbf{f}$ is a Lipschitz continuous function, guaranteeing existence and uniqueness of solutions to Equation 2.1 (Brunton and Kutz, 2019).

The evolution in time of the system state can also be described in a discrete-time form:

$$\mathbf{x}_k = \mathbf{F}(\mathbf{x}_{k-1}). \tag{2.2}$$

This formulation is more general than the continuous one in Equation 2.1, encompassing discontinuous and hybrid systems as well, and is usually more appropriate and natural to be used when considering experimental data.

Sampling the trajectory of Equation 2.1 in time, discrete-time dynamics may be induced from the continuous-time dynamics, so that $\mathbf{x}_k = \mathbf{x}(k\Delta t)$. The discrete-time propagator $\mathbf{F}$, also known as *flow map*, is now parameterized by the time step $\Delta t$:

$$\mathbf{x}_k = \mathbf{F}_{\Delta t}(\mathbf{x}_{k-1}). \tag{2.3}$$

In this formulation, $\mathbf{x}_k$ and $\mathbf{x}_{k-1}$ can be interpreted as snapshots of the system taken at a distance in time equal to the time step $\Delta t$. These vectors contain the values of the QoI measured or simulated at several points in the space domain. The evolution of the variable from time $t_{k-1}$ to $t_k$ in each point is described by $\mathbf{F}_{\Delta t}$.

## 2.2 Dynamic Mode Decomposition (DMD)

DMD originated in the fluid dynamics community, where Schmid et al. (2010) first proposed this method to identify spatiotemporal coherent structures from high-dimensional data. The method seeks to compute the best-fit linear operator to relate time-shifted snapshots of the system's state to model the complex system behavior directly from the data.

Based on the well-known SVD, DMD is classified as a typical order-reduction technique. The selection of an appropriate truncation rank for the SVD allows us to control the dimension of the model and balance accuracy and computational cost. The result is a computationally efficient mathematical representation of the complex system directly inferred from data, commonly referred to as ROM. A ROM typically captures the key features of the system's dynamics directly from high-fidelity simulations or measurements and acts as a surrogate to get predictions at a negligible computational cost. Its simple mathematical form allows for the generation of a large number of simulations while preserving the high-fidelity representation of the phenomena.

As explored in Rowley et al. (2009) and Mezić (2013) soon after its first development, the linear DMD-based ROM can be viewed, given an appropriate selection of observables, as a finite-dimensional linear approximation of the linear, infinite-dimensional Koopman operator (Koopman, 1931). That guarantees the possibility of physically interpreting the structure of the model, which, as a consequence, can serve not only as an efficient model to reproduce the system behavior but also as a tool to discover dominant features of the phenomena.

Similarly to the Proper Orthogonal Decomposition (POD), based on the information contained in the SVD, the DMD can identify spatiotemporal patterns that can be analyzed to decode the system behavior and then recombined to reconstruct and predict system states. However, while POD provides a hierarchy of features (i.e., modes) that only considers spatial correlation and energy content, ignoring any temporal information, DMD provides spatial modes together with the description of their evolution over time, in terms of frequency, oscillations, and growth or decay rate (i.e., time dynamics). Hence, this method can be seen as a combination of the favorable aspects of the POD/SVD technique and the Fast Fourier Transform for temporal frequency identification.

The interpretability of the model structure makes DMD different from many other data-driven methods, such as Neural Networks, which are typical ”black-box” models. In this case, the ROM is useful not only for the prediction or reconstruction of the evolution of a QoI in space and time but also to provide insights into the structure of the phenomena, and patterns that can be related to physical influencing factors. The connection offered by DMD between nonlinear dynamical systems and a linear, interpretable, and computationally straightforward framework has raised considerable interest in this method.

The primary field where DMD has been largely employed since its origin is fluid dynamics (Rowley et al., 2009; Schmid et al., 2010; Schmid, 2010; Bagheri, 2013; Pan et al., 2021; Liu et al., 2021; Ohh and Spedding, 2022). Many optimizations have been developed to adapt the algorithm to several different contexts and applications in this field, as discussed in the reviews by Kutz et al. (2016) and more recently by Schmid (2022). Nevertheless, its equation-free and flexible structure is encouraging the spreading in very different fields, from neuroscience (Brunton et al., 2016) to epidemiology (Proctor and Eckhoff, 2015), from financial trading (Mann and Kutz, 2016) to video processing (Erichson et al., 2016).

DMD is purely based on data, equally valid for experimental and numerical datasets, and doesn’t require knowledge of the governing equations. The main advantages derive from its application when traditional physics-based models are too complex to design and employ, and a large amount of data describing the temporal and spatial evolution of the system is available. In these cases, DMD-based models are used not only as cheaper surrogates but their structure can also be investigated to discover patterns and

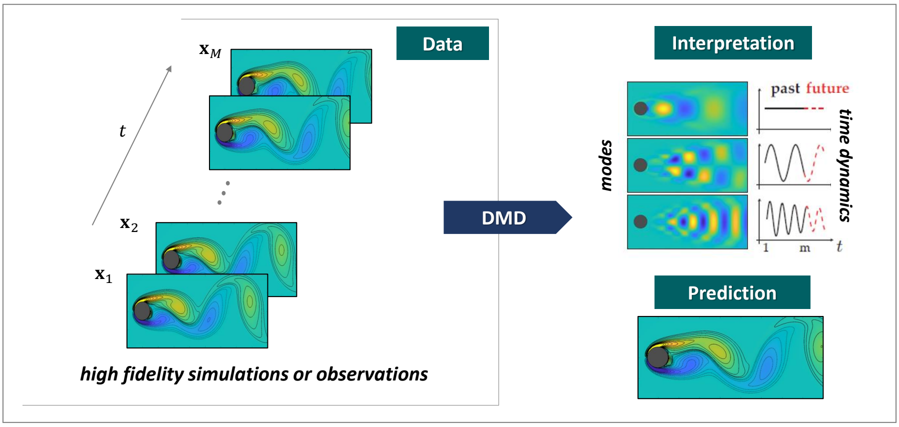

**Figure 2.1:** Dynamic Mode Decomposition to interpret and predict complex dynamical systems' behavior. The method derives from high-fidelity simulations or observations a reduced-order model, whose structure can be decomposed into spatial patterns (modes) and respective temporal evolution (time dynamics), which can be physically interpreted and recombined to predict future states of the system.

physical features in the phenomena.

### 2.2.1 Standard DMD Algorithm

Among different versions of the original standard DMD (sDMD) formulation, the exact DMD framework developed by Tu et al. (2014) is commonly considered the most general and rigorous.

DMD is inherently data-driven hence, the first step concerns the collection and management of data. Suppose to have a set of $M$ pairs of snapshots representing the state of the system in $N$ points in space as it evolves in time. Each pair $m = 1, 2, \ldots, M$ can be denoted by $\{\mathbf{x}(t_k), \mathbf{x}(t'_k)\}^m$, where $t_k$ is the $k$-th instant in the observed temporal interval of system evolution, $t'_k = t_k + \Delta t$, and $\Delta t$ is a fixed time step for all the pairs. Each snapshot can be arranged into a vector and two time-shifted matrices, $\mathbf{X} \in \mathbb{R}^{N\times M}$ and $\mathbf{X}' \in \mathbb{R}^{N\times M}$, can be built such that the first element $\mathbf{x}(t_k)$ of each pair becomes a column of $\mathbf{X}$, while the second element $\mathbf{x}(t'_k)$ becomes a column of $\mathbf{X}'$:

$$\begin{aligned}
\mathbf{X} &= \begin{bmatrix} | & | & & | \\ \mathbf{x}(t_1) & \mathbf{x}(t_2) & \dots & \mathbf{x}(t_M) \\ | & | & & | \end{bmatrix}, \\
\mathbf{X}' &= \begin{bmatrix} | & | & & | \\ \mathbf{x}(t'_1) & \mathbf{x}(t'_2) & \dots & \mathbf{x}(t'_M) \\ | & | & & | \end{bmatrix},
\end{aligned} \tag{2.4}$$

In the original formulations provided by Schmid et al. (2010) and Rowley et al. (2009), the sampling in time was assumed to be uniform, hence $t_k = k\Delta t$ and $t'_k = t_k + \Delta t = t_{k+1}$. In this case, the notation can be simplified in $\mathbf{x}_k = \mathbf{x}(k\Delta t)$. The two matrices $\mathbf{X}$ and $\mathbf{X}'$, in this particular case, are formed with a collection of sequential snapshots, evenly spaced in time and therefore the first $m-1$ columns of $\mathbf{X}'$ can be obtained by simply shifting $\mathbf{X}$ by one time step:

$$\begin{aligned}
\mathbf{X} &= \begin{bmatrix} | & | & & | \\ \mathbf{x}_0 & \mathbf{x}_1 & \dots & \mathbf{x}_{M-1} \\ | & | & & | \end{bmatrix}, \\
\mathbf{X}' &= \begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_M \\ | & | & & | \end{bmatrix},
\end{aligned} \tag{2.5}$$

In the standard version of the DMD algorithm (sDMD), the relationship between $\mathbf{X}$ and $\mathbf{X}'$ is approximated with the best-fit linear operator $\mathbf{A}$ as follows:

$$\mathbf{X}' \approx \mathbf{A}\mathbf{X}. \tag{2.6}$$

If we assume uniform time sampling, the temporal evolution of the variable $\mathbf{x}(t)$ is approximated by a linear model:

$$\mathbf{x}_{k+1} \approx \mathbf{A}\mathbf{x}_k. \tag{2.7}$$

The best-fit operator $\mathbf{A}$ establishes a linear dynamical system that best advances snapshot measurements forward in time. It is mathematically defined as:

$$\mathbf{A} = \arg\min_{\mathbf{A}} \parallel \mathbf{X}' - \mathbf{A}\mathbf{X} \parallel_F = \mathbf{X}'\mathbf{X}^\dagger \in \mathbb{R}^{N\times N} \tag{2.8}$$

where $\parallel \cdot \parallel_F$ is the Frobenius norm and $\dagger$ denotes the pseudo-inverse.

In a typical application, the number of snapshot pairs is much smaller than the number of elements in each snapshot (sampling points), $M \ll N$, so that the rank of $\mathbf{A}$ is at most $M$. Anyway, computing $\mathbf{A}$ through Equation 2.8 is generally onerous. Alternatively, the pseudo-inverse $\mathbf{X}^\dagger$ in Equation 2.8 is computed via the truncated SVD of $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, with rank $r < M$. It follows:

$$\mathbf{A} \approx \mathbf{X}'\mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^\top, \tag{2.9}$$

where $\mathbf{U} \in \mathbb{R}^{N\times r}$, $\mathbf{\Sigma} \in \mathbb{R}^{r\times r}$, $\mathbf{V} \in \mathbb{R}^{M\times r}$. If $r$ is smaller than the number of nonzero singular values (i.e., the rank of $\mathbf{X}$), then the truncated SVD is a proxy of $\mathbf{X}$.

Possibly, the problem dimensionality can be further reduced by projecting $\mathbf{A}$ onto the leading singular vectors of $\mathbf{X}$ (POD modes), thus resulting in a new matrix $\tilde{\mathbf{A}} \in \mathbb{C}^{r\times r}$:

$$\tilde{\mathbf{A}} = \mathbf{U}^\top\mathbf{A}\mathbf{U} = \mathbf{U}^\top\mathbf{X}'\mathbf{V}\mathbf{\Sigma}^{-1} \tag{2.10}$$

The matrix $\tilde{\mathbf{A}}$ defines a low-dimensional linear model of the dynamical system on POD coordinates

$$\tilde{\mathbf{x}}_{k+1} \approx \tilde{\mathbf{A}}\tilde{\mathbf{x}}_k, \tag{2.11}$$

from which it is possible to reconstruct the high-dimensional state $\mathbf{x}_{k+1}$ as:

$$\mathbf{x}_{k+1} = \mathbf{U}\tilde{\mathbf{x}}_{k+1} = \mathbf{U}\left(\tilde{\mathbf{A}}\tilde{\mathbf{x}}_k\right). \tag{2.12}$$

Note that computing the spectral decomposition of $\tilde{\mathbf{A}}$:

$$\tilde{\mathbf{A}}\mathbf{\Phi}_{\tilde{\mathbf{A}}} = \mathbf{\Phi}_{\tilde{\mathbf{A}}}\mathbf{\Lambda} \tag{2.13}$$

provides the DMD eigenvalues $\mathbf{\Lambda}$ of matrix $\mathbf{A}$ which corresponds to eigenvalues of matrix $\tilde{\mathbf{A}}$; while the DMD modes, i.e. the eigenvectors of $\mathbf{A}$, can be computed from the eigenvectors $\mathbf{\Phi}_{\tilde{\mathbf{A}}}$ of $\tilde{\mathbf{A}}$ as follows: $\mathbf{\Phi}_{\mathbf{A}} = \mathbf{X}'\mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{\Phi}_{\tilde{\mathbf{A}}}$.

While the matrix $\mathbf{A}$ can be directly used as a ROM of the system behavior to reconstruct and/or predict the system state at given time steps, the eigenvalues and eigenvectors resulting from its spectral decomposition allow the identification of spatial and temporal patterns in the data. These structures can be re-combined to reproduce and/or predict future states, but more importantly, they can be studied to unveil the physics of the underlying phenomena.

### 2.2.2 Extended DMD Algorithm

One of the main advantages of DMD is its simple, flexible, and equation-free structure. For this reason, several methodological innovations have been included to optimize and adapt the standard formulation to different problems.

Here, we briefly present the development of an extended version of the classical DMD algorithm proposed by Lu and Tartakovsky (2021) to adapt the method to inhomogeneous conditions.

To allow for a problem's inhomogeneity, the generalized DMD (gDMD) algorithm adds a bias term $\mathbf{b}_g \in \mathbb{R}^N$ (Qin et al., 2019) to the standard formulation described by Equation 2.7, such that

$$\mathbf{x}_{k+1} \approx \mathbf{A}_g \mathbf{x}_k + \mathbf{b}_g. \tag{2.14}$$

Here, $[\mathbf{A}_g \ \mathbf{b}_g] = \mathbf{X}'\tilde{\mathbf{X}}^{\dagger} \in \mathbb{R}^{N \times N+1}$, where $\tilde{\mathbf{X}}^{\top} = [\mathbf{X} \ \mathbf{1}]$ and $\tilde{\mathbf{X}} \in \mathbb{R}^{N+1 \times M}$. As in the sDMD algorithm, the SVD of the matrix $\tilde{\mathbf{X}} \approx \mathbf{U}_g \mathbf{\Sigma}_g \mathbf{V}_g^{\top}$ is employed to reduce the computational cost of the best-fit linear operator estimation:

$$[\mathbf{A}_g \ \mathbf{b}_g] \approx \mathbf{X}'\mathbf{V}_g \mathbf{\Sigma}_g^{-1} \mathbf{U}_g^{\top}, \tag{2.15}$$

where $\mathbf{U}_g \in \mathbb{R}^{N+1 \times r}, \mathbf{\Sigma}_g \in \mathbb{R}^{r \times r}, \mathbf{V}_g \in \mathbb{R}^{M \times r}$.

The extended DMD (xDMD) approach (Lu and Tartakovsky, 2021) endows gDMD (Equation 2.14) with a residual-learning idea (Chen and Xiu, 2021). It approximates the relationship between $\mathbf{Y} = \mathbf{X}' - \mathbf{X}$ and $\mathbf{X}$ such that

$$\mathbf{y}_{k+1} = \mathbf{B}_x \mathbf{x}_k + \mathbf{b}_x. \tag{2.16}$$

Here, $[\mathbf{B}_x \ \mathbf{b}_x] = \mathbf{Y}\tilde{\mathbf{X}}^\dagger \in \mathbb{R}^{N \times N+1}$, and $\tilde{\mathbf{X}}^\top \in \mathbb{R}^{N+1 \times M}$ is defined as before. For computational saving, the best-fit linear operator is obtained through the SVD of the matrix $\tilde{\mathbf{X}}$ as

$$[\mathbf{B}_x \ \mathbf{b}_x] \approx \mathbf{Y}\mathbf{V}_g \mathbf{\Sigma}_g^{-1} \mathbf{U}_g^\top. \tag{2.17}$$

The matrices $\mathbf{A}$ (Equation 2.9) and $[\mathbf{B}_x \mathbf{b}_x]$ (Equation 2.17) can be equally employed as ROMs of the system state evolution, if this is described by nonlinear PDEs, whose solution is confined in $\mathcal{H} \subseteq \mathbb{R}^N$ (to satisfy the assumptions in Lemma 2.1 in Lu and Tartakovsky (2021)).

### 2.2.3 Optimized DMD Implementation

Both the sDMD and xDMD algorithms have a simple mathematical formulation. However, as any application of practical significance involves high-dimensional datasets, careful code implementation is crucial to guarantee efficiency and accuracy.

More robust and accurate results can be achieved if chained multiplications of matrices are broken into convenient groups using parentheses to dictate the order of the operations. Thus, here we propose an efficient computational strategy to derive predictions from Equation 2.16:

$$\mathbf{y}_{k+1} = \mathbf{C}_x \mathbf{D}_x, \tag{2.18}$$

where $\mathbf{C}_x = \mathbf{Y}\mathbf{V}_g \Sigma_g^{-1} \in \mathbb{R}^{N \times r}$, and $\mathbf{D}_x = \mathbf{U}_g^\top \begin{bmatrix} \mathbf{x}_k \\ 1 \end{bmatrix} \in \mathbb{R}^{r \times 1}$.

A similar computational approach can be conveniently adopted to derive prediction from Equation 2.7:

$$\mathbf{x}_{k+1} = \mathbf{C}_s \mathbf{D}_s, \tag{2.19}$$

where $\mathbf{C}_s = \mathbf{Y}\mathbf{V}\Sigma^{-1} \in \mathbb{R}^{N\times r}$, and $\mathbf{D}_s = \mathbf{U}^\top \mathbf{x}_k \in \mathbb{R}^{r\times 1}$. Thanks to this convenient implementation fewer operations and less intermediate memory usage are required and the consequence is a speed-up in computations.

### 2.2.4 Truncation Rank in DMD Models

As described for all the different DMD formulations, the core of the algorithm is the computation of the pseudoinverse (Equation 2.8) via a truncated SVD of the original matrix of data $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, with rank $r < M$. This step allows for a reduction in computational costs and guarantees the applicability of the method. However, the efficiency and accuracy of the resulting ROM depend on the selected truncation rank $r$. In practice, choosing this rank is one of the most important and subjective steps in computing DMD, and in dimensionality reduction in general.

Here, we propose a series of truncation criteria that can be applied and compared, to find an optimal balance between computational efficiency and accuracy.

One possible option is neglecting a number of features that contain a small percentage of the total information according to the cumulative energy curve associated with the diagonal elements $\sigma_k$ of the diagonal matrix $\mathbf{\Sigma}$. We can use, e.g., a truncation rank $r = r_{90}$ corresponding to the number of diagonal elements of $\mathbf{\Sigma}$ accounting for 90% of the cumulative energy in the SVD of $\mathbf{X}$.

An alternative method is setting a threshold $\zeta$ to be applied on the order of magnitude of $\sigma_k$ as follows:

$$r^* = \min(n) : \sigma_n \leq \zeta \sum_{k=0}^{M-1} \sigma_k. \tag{2.20}$$

A third option could be using functions from the libraries of the employed coding software. The function `rank`($\mathbf{Z}$) of `MATLAB` returns the number of linearly independent columns in the matrix $\mathbf{Z}$. This rank is computed as the number of singular values that are larger than a tolerance. The tolerance is set by default as:

$$\texttt{max(size((}\mathbf{Z}\texttt{)))*eps(norm((}\mathbf{Z}\texttt{)))}, \tag{2.21}$$

but a different value can be specified in `rank(Z,tol)`.

### 2.2.5 Performances of DMD Algorithms

To evaluate the suitability of the different algorithms and guide the selection of the optimal truncation rank, the snapshots predicted by the DMD-based ROMs can be compared with the high-fidelity reference data.

The difference between the state of the system predicted by the generic DMD model $\mathcal{L}$ and the real snapshot at time $t_k$ is defined as

$$\mathbf{d}_k^{\mathcal{L}} = \mathbf{x}_k^{\mathcal{L}} - \mathbf{x}_k, \tag{2.22}$$

where $\mathbf{x}_k$ is the true solution induced by the flow map $\mathbf{F}_{\Delta t}$ at time $t_k$, and $\mathbf{x}_k^{\mathcal{L}}$ its DMD approximation.

The performances of the different DMD versions can be assessed and compared in terms of the relative error defined as:

$$\varepsilon_k^{\mathcal{L}} = \frac{\|\mathbf{d}_k^{\mathcal{L}}\|^2}{\|\mathbf{x}_k\|^2}, \tag{2.23}$$

where $\|\cdot\|$ denotes the vector 2-norm.

## 2.3 Polynomial Chaos Expansion (PCE)

Among the techniques for the construction of ROMs, the Polynomial Chaos Expansion (PCE) is an efficient tool to derive, directly from high-fidelity simulations, an approximation of the model response surface in the form of a polynomial series in the space of variability of the governing parameters (Sudret, 2008; Ghanem and Spanos, 1991; Wiener, 1938).

This series is given by the truncation to a finite order of an expansion made of multi-variate polynomials (orthonormal with respect to the joint pdf of the parameters), multiplied by deterministic coefficients that embed the space-time dependence of the response (Xiu and Karniadakis, 2002). The form of the multivariate orthogonal polynomials depends on the input variables' probability distribution function, while the coefficients need to be computed using high-fidelity data.

The efficiency of the method depends on the minimization of the number of high-fidelity simulations required for the calibration. Calibration can be accomplished through a non-intrusive regression-based method (Sudret, 2008; Webster et al., 1996) which requires the solution of a minimization problem on an optimum set of regression points provided by the Probabilistic Collocation Method (PCM). These points correspond to combinations of values properly selected to span the entire space of variability of the input parameters.

The resulting polynomial is a ROM, built to replace the complex and computationally onerous HFM. The PCE can capture, based on a limited number of runs of the HFM, the essential features of the underlying dynamics for the specific QoI and provide a surrogate to simulate at a reduced computational cost the system behavior for different input scenarios. The simple polynomial form of the surrogate guarantees a negligible computational cost for each simulation, compared to physics-based simulations performed with the HFM. The drastic reduction of the computational cost allows us to perform in-depth analyses (otherwise unfeasible on the HFM) such as global sensitivity analysis and Monte Carlo simulations (Ciriello et al., 2019; Focaccia et al., 2021). The result is an increased and improved ability to diagnose the systems' behavior, as influenced by different phenomena and parameters affected by variability and uncertainty.

One major field of application of PCE-based surrogates is hydrogeology (Oladyshkin et al., 2012; Ciriello et al., 2012; Riva et al., 2015; Ciriello et al., 2019), where the PCE has been successfully employed to build ROMs of many flow and transport problems, accelerating uncertainty quantification and other onerous analyses, otherwise computationally prohibitive if performed directly on a high-fidelity model (Ciriello et al., 2017; Ciriello and de Barros, 2020).

In recent years, PCE-based surrogates have also been used in different research fields, such as cardiovascular research. Here, PCE can be employed to ease uncertainty quantification analysis and support the identification of relevant information to assist the clinical practice (Sankaran and Marsden, 2011; Schiavazzi et al., 2015; Quicken et al., 2016).

A cardiovascular application of this surrogate modeling technique is illustrated in Appendix A. Here, the PCE method is employed to derive surrogate models of the dynamics of selected quantities of interest that describe the left ventricle function. The resulting surrogates are directly employable in clinical analysis, can be used to perform global sensitivity analysis at a negligible computational cost and provide insights on the impact of different diagnostic metrics on the left ventricle function. The biofluid application, reported in this thesis to briefly illustrate the use and potentialities of the PCE technique, is the object of our work *Surrogate models provide new insights on metrics based on blood flow for the assessment of left ventricular function* (Collia et al., 2022). For a more detailed discussion of the methods and the results please refer to the paper.

### 2.3.1 The PCE algorithm

Let $f(\cdot)$ be a HFM predicting a QoI $\mathbf{x}$ (output) depending on a vector of governing parameters $\mathbf{p}$ (inputs), whose variability can be reproduced by modeling them as independent random variables.

If the variance of the QoI $\sigma^2_{\mathbf{x}}$ is finite, the PCE may be applied to generate an approximation $\hat{\mathbf{x}}$ of the response surface in the random parameters space in terms of a polynomial series as follows (Wiener, 1938; Ghanem and Spanos, 1991; Xiu and Karniadakis, 2002):

$$\mathbf{x} = f(\mathbf{p}) \rightarrow \qquad \hat{\mathbf{x}} = \sum_{j=0}^{P-1} \mathrm{a}_j \Psi_j(\mathbf{p}), \qquad P = \frac{(N_{par} + q)!}{N_{par}! q!}, \quad (2.24)$$

where $\Psi_j$ are multivariate polynomials that constitute an orthonormal basis with respect to the joint PDF of $\mathbf{p}$, and coefficients $\mathrm{a}_j$ are the deterministic coordinates of the spectral decomposition. The number of terms of the series $P$ depends on the number of parameters $N_{par} = \dim \mathbf{p}$ and the maximum degree of the expansion $q$.

To guarantee optimum convergence rate in the case of non-Gaussian processes, the family of the multivariate polynomials $\Psi_j$ is selected, based on the PDF of the inputs, from a variety of possible polynomial families, according to the generalized PCE scheme introduced by Xiu and Karniadakis (2002).

The PCE coefficients $\mathrm{a}_j$ are computed through a non-intrusive regression-based approach. According to this method, the variance of the residual $\varepsilon = |\hat{\mathbf{x}} - \mathbf{x}|$ is minimized with respect to the vector of coefficients $\mathbf{a}$ (Sudret, 2008) as follows:

$$\mathbf{a} = Argmin \frac{1}{N_{RP}} \sum_{i=1}^{N_{RP}} \left\{ \mathbf{x} - \sum_{j=0}^{P-1} \mathrm{a}_j \Psi_j(\mathbf{p}) \right\}^2, \quad (2.25)$$

where $N_{RP}$ is the number of regression points. This is done on a limited dataset of high-fidelity simulations generated for an optimal set of values of $\mathbf{p}$ provided by the Probabilistic Collocation Method (PCM) (Webster et al., 1996). These values are named *collocation points* and they are selected based on the same arguments adopted for integral estimation through Gaussian quadrature; specifically, the PCM employs the roots of the polynomial of one order higher than $q$ to assure proper sampling of the region associated with the largest probability in the distributions of the input parameters (Webster et al., 1996). The efficiency of the PCE method resides in this procedure: the less the number of collocation points ($N_{RP} = N_{CP}$) the more the effectiveness of the PCE. Note that $N_{CP} \geq P$, i.e. it increases with $N_{par}$ and $q$ (Ciriello et al., 2012).

### 2.3.2 PCE Efficiency and Dynamical Systems

If we approximate the behavior of a dynamical system that evolves both in space and time, i.e. $f(\mathbf{s}, t, \mathbf{p})$, the space-time dependence of the response $\mathbf{x}$ is included in the PCE coefficients $\mathbf{a}$:

$$\hat{\mathbf{x}}\left(\mathbf{s}, t\right) = \sum_{j=0}^{P-1} \mathrm{a}_j\left(\mathbf{s}, t\right) \Psi_j(\mathbf{p}). \tag{2.26}$$

In this case, a PCE model has to be built at each point $(\mathbf{s}, t)$ of interest, which means that the coefficients should be computed based on high-fidelity simulations of $\mathbf{x}$ generated at the same space-time locations for each value of $\mathbf{p}$ provided by the PCM.

It follows that if we have a complex dynamical process, solved numerically by means of a HFM over a grid of $N$ elements and for $M$ time instants, we need to compute our QoI $\mathbf{x}$ for each point in space and time in $N_{CP}$ differently parameterized scenarios (Ciriello and de Barros, 2020).

This results in $N \times M \times N_{CP}$ evaluations of $\mathbf{x}$. Using the PCE technique to evaluate how complex dynamical systems respond to the variability of input parameters becomes unfeasible, limited by the prohibitive computational cost related to the high-fidelity simulations required for the calibration.

A proper combination of the PCE with other techniques to build ROMs can help overcome these difficulties. A possible option is reducing the computational burden related to the generation of the training simulations, through the replacement of the high-fidelity data with data provided by interpolation with the DMD. In Chapter 5 we propose a robust framework based on the combination of DMD and PCE techniques for the reduction of models with time-variant response.

# 3 | DMD for Order Reduction in High-Fidelity Simulations

Dynamic Mode Decomposition (DMD) is a powerful method to extract the key underlying dynamics of a complex system directly from high-fidelity data and recombine them into a reduced-order model (ROM) of the system behavior. Thanks to their efficiency, DMD-based ROMs can provide a viable solution to reduce the computational cost and storage requirements related to fluid dynamics simulations.

In this chapter, we investigate the efficacy and robustness of the extended DMD (xDMD) version, testing its ability to derive ROMs from high-fidelity flow simulations. The method is applied to a set of three-dimensional cardiovascular simulations.

We explore the ROMs' performances in representation, interpolation, and extrapolation, and we identify relationships between the accuracy and the truncation rank applied to the Singular Value Decomposition (SVD). Different truncation criteria are proposed and tested to investigate the impact of this choice, which underpins DMD and other approaches to ROM construction, on the efficiency of the method. The innovation is to relate the truncation rank to the singular values of the original flow problem so that a priori guidelines are established for the xDMD use.

The study demonstrates that, given an optimal low-rank truncation, able to balance accuracy, generalizability, and computational saving, the xDMD can be successfully employed as a means of data compression and reconstruction of the system's dynamics. This method could help speed up the modeling phase, with positive impacts on the diagnosis and the optimization of intervention choice.

The chapter is organized as follows: in Section 3.1 we introduce the concept of high-fidelity models' (HFMs) reduction; in Section 3.2 we present the test case, describing high-fidelity data generation and the structure of the analysis; Section 3.3 is devoted to the discussion of results; while a set of final remarks is reported in Section 3.4.

The study presented in this chapter is related to the paper G. Libero, A. Chiofalo, V. Ciriello, D.M. Tartakovsky, *Dynamic mode decomposition for order reduction in high-fidelity cardiovascular simulations* expected to be submitted soon to Physics of Fluids.

## 3.1 High-Fidelity Model Reduction

HFMs are commonly developed to reproduce the behavior of specific systems of interest accurately. They provide detailed and reliable descriptions of complex interactions among different phenomena and variables, and they are mainly physics-based. However, the complexity of reality poses limits to the applicability of these models. On one hand, the wide framework of variables considered in the model requires a detailed characterization, which is not always feasible, and often limits model validity to case-specific domains. On the other hand, the time and resources required to run simulations of such complex structures are not compensated by the rapid and constant growth in computational power.

Model reduction techniques are emerging as powerful tools to derive models with a lower computational burden, suitable for a wider and more flexible description of how the world around us evolves. A prove is the vast literature concerning the use of these instruments to model complex flow and transport processes in fields as diverse as geosciences and biomedicine (Lu and Tartakovsky, 2020; Ciriello and de Barros, 2020; Dutta et al., 2021; Callaham et al., 2022; Karbasian and Vermeire, 2022; Collia et al., 2022; Pegolotti et al., 2021). These techniques can derive, from a limited number of high-fidelity simulations, computationally efficient mathematical representations of the relationships between the governing variables and the quantity of interest. ROMs can reconstruct the leading dynamics of the system and provide simulations of its state at a low computational cost. Thus, a large number of simulations can be performed, allowing uncertainty quantification and other onerous analyses that would be prohibitive if performed directly on an HFM (Focaccia et al., 2021; Ciriello et al., 2017, 2012; Oladyshkin et al., 2012). The key feature of a ROM is efficiency, defined as a balance between the accuracy and the computational cost related to the generation of the high-fidelity data required for the model's construction (Ciriello et al., 2017).

In all the methods for ROM generation that rely on the Singular Value Decomposition (SVD), such as the Dynamic Mode Decomposition (DMD), the efficiency is governed by the selection of the truncation rank. The DMD method aims at extracting, through the SVD of a set of system states' simulations, the spatiotemporal structures characterizing the system behavior (Kutz et al., 2016). The truncation applied to the SVD governs the identification of the dominant structures, which converge in the construction of the DMD-based ROM. The complexity of the resulting model depends on the number of structures recombined to reproduce the underlying processes. As a result, the truncation rank selection impacts the computational demand of the ROM but also controls the degree of order reduction and therefore accuracy. Truncating at higher levels we maintain a larger number of informative features from the original data. This may lead to better performances in the calibration phase when the ROM reproduces the same training data. Though, the error may increase for systems states out of the train set and the costs of running the model could be prohibitive. Low-rank truncation, instead, prevents overfitting and also cuts the computational expense. Despite its importance, the selection of an appropriate rank is typically done via experimentation, rendering the methods' implementation subjective (Lu and Tartakovsky, 2021). The definition of a-priori guidelines for the selection of an optimal truncation rank could be crucial to the success of DMD.

In this chapter, we analyze how the error of a DMD-based ROM is linked to the truncation rank in SVD, which, in turn, is linked to the singular values of the problem. We test the accuracy of the approximation in terms of both reconstruction of snapshots employed for the training phase, and extrapolation or interpolation of system states not included in the train set. Different truncation ranks are applied and the resulting errors are compared to define optimal truncation criteria for the use of DMD as a viable method for both data compression and reconstruction of the system's dynamics.

## 3.2 Data and Application

### 3.2.1 3D Cardiovascular Modeling

Numerical simulations of fluid dynamics have gained considerable attention during the last decades as a valuable quantitative tool for diagnosis and scenario analysis in the medical sector. In particular, by modeling blood flow in the cardiovascular system we can get key insights for the identification of long and short-term pathologies, but we can also simulate hypothetical scenarios to evaluate diverse intervention choices (Pegolotti et al., 2021; Bao et al., 2014; Collia et al., 2022; Pedrizzetti et al., 2014). Blood dynamics are typically modeled through Navier Stokes equations, discretized by numerical methods, such as the finite element method. This often leads to HFMs, which are incompatible with clinical practice due to the time and resources required to run simulations.

Here we apply the extended version of DMD (xDMD) to prove its usefulness in this context and investigate its settings and efficiency. The xDMD method could help speed up the modeling phase, with positive impacts on the diagnosis and the optimization of intervention choice. The crucial advantage is the possibility to learn the dynamics directly from a small number of high-fidelity simulations. We test the method on a set of three-dimensional (3D) cardiovascular simulations of blood flow in a patient-specific aorta.

High-fidelity data are generated with the `SimVascular` software (*simvascular.github.io*), which provides a complete pipeline, from medical image data segmentation to patient-specific blood flow simulations (Updegrove et al., 2016). We employ this fully open-source software package to solve 3D incompressible Navier-Stokes equations describing blood flow in a patient-specific aorta with a homogeneous Dirichlet boundary condition (rigid wall).

The reference geometry is selected from the Vascular Model Repository (*www.vascularmodel.com*), a library of patient-specific cardiovascular models developed on volumetric image datasets and relevant physiologic data (Wilson et al., 2013). The quantity of interest is the velocity magnitude, for which 1929 snapshots are collected over 7.7 s, covering about 12 pulsations. Each snapshot contains the values of the velocity simulated for the $N = 343352$ elements of the triangular mesh employed to discretize the spatial domain. The size of the dataset is $\approx 50$ Gb.

On this dataset, we calibrate xDMD-based ROMs truncated at different ranks and we evaluate their ability to reconstruct snapshots from the training set (representation regime). Results from this analysis can prove the viability of using xDMD as an efficient technique for data compression. We also test the performance of the ROMs to predict extra-snapshots, referring to time steps located in between (interpolation regime) and after (extrapolation regime) the ones employed for training. The impact of the truncation rank selection is investigated in all three regimes. We test the three truncation criteria presented in Section 2.2.4, to identify useful a-priori guidelines to get an optimal balance of accuracy, generalizability, and computational saving. In particular, $r_{90}$ corresponds to the minimum number of singular values accounting for 90% of the cumulative energy in the SVD of $\tilde{\mathbf{X}}$; $r^*$ is identified setting $\zeta = 10^{-5}$ in Equation 2.20; while $r = \texttt{rank}(\tilde{\mathbf{X}})$ is computed with the standard `MATLAB` function.

## 3.3 Results and Discussion

### 3.3.1 Representation Error and Data Compression

We employ xDMD to generate ROMs based on the entire dataset of snapshots of the velocity magnitude simulated by the HFM and test their ability to reproduce these training data. Given the 1929 high-fidelity snapshots, we build the xDMD training matrices as described in Sections 2.2.1 and 2.2.2: $\tilde{\mathbf{X}} \in \mathbb{R}^{N+1\times M}$ and $\mathbf{Y} \in \mathbb{R}^{N\times M}$ with $M = 1928$ and $N = 343352$. This exercise quantifies the representation error of xDMD. The application of different truncation ranks to the SVD produces a sequence of ROMs.

The accuracy of xDMD is explored at low-rank truncations, which are relied upon to identify dominant spatiotemporal structures in the data to reconstruct the high-fidelity simulations. This is also relevant for the evaluation of xDMD effectiveness for data compression and storage.

We first investigate the structure of the dataset of high-fidelity simulations used to train the ROMs. We perform the SVD of the matrix $\tilde{\mathbf{X}}$ (see Sections 2.2.1 and 2.2.2 for matrix construction from data) and we analyze the distribution of the information among the singular values. Panels (a) and (b) of Figure 3.1 represent the singular values and the associated cumulative energy, respectively. As the singular values rapidly decrease to zero, the cumulative energy quickly reaches the maximum value, becoming approximately equal to 1 already at $r^*$. This indicates that a limited number of dominant modes, captured by $r^*$, are sufficient to describe and reproduce the system states with high accuracy. As a result, the remaining features ($n > 404$) are interpreted as noise and, for the purpose of data compression, can be neglected.

Given these insights on the informative content of the training set, we investigate the influence of the truncation rank on the ability of the method to reconstruct training snapshots. Panel (c) of Figure 3.1 shows the representation error, computed with Equation (2.23) and averaged over all the time steps, for several ROMs truncated at different ranks $r$. As expected, by increasing the truncation rank $r$ we decrease the representation error. The lowest error is measured when $r = \texttt{rank}(\tilde{\mathbf{X}})$, i.e. in the absence of truncation, and it is equal to $1.4 \cdot 10^{-14}$. However, high accuracy is already reached for relatively low $r$: setting $r = r^*$ or $r = r_{90}$ leads to errors of $2.96 \cdot 10^{-5}$ or $1.49 \cdot 10^{-1}$, respectively.

To elucidate further the effects of the truncation rank on the accuracy of xDMD-based ROMs, in Figure 3.2 we compare the original data (Panel (a)) with the corresponding reconstructed snapshots provided by the ROMs truncated at $r_{90}$ (Panel (b)) and $r^*$ (Panel (c)). Even if $r_{90}$ returns a slightly worse approximation,

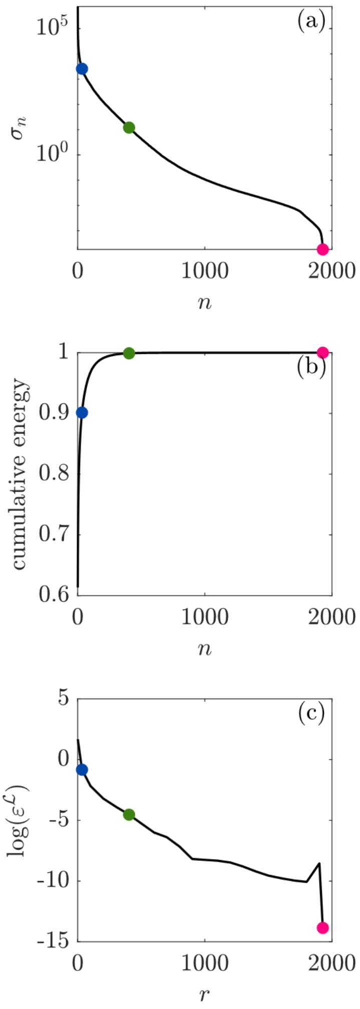

**Figure 3.1:** (a) Singular values and (b) cumulative energy associated with the SVD of $\tilde{\mathbf{X}}$ when all ($N = 343352$, $M = 1928$) data are used to train the ROMs, both plotted as a function of the singular values number $n$. (c) Representation error (averaged over the time instants) plotted as a function of the truncation rank $r$ of the SVD. In all panels, the blue, green, and red dots correspond to $r = r_{90}$, $r = r^*$, and $r = \texttt{rank}(\tilde{\mathbf{X}})$, respectively. In this example, $r_{90} = 34$, $r^* = 404$, while $\texttt{rank}(\tilde{\mathbf{X}}) = M = 1928$ results in no truncation.

**Figure 3.2:** Magnitude of the flow velocity $u$ in the aorta, as predicted by (a) direct numerical simulations, and xDMD-based ROMs with truncation ranks (b) $r_{90}$, and (c) $r^*$. The velocity is plotted at times $k = 643$, $k = 1286$, and $k = 1929$ in the first, second, and third rows, respectively.

**Table 3.1:** Data considered for representation, interpolation, and extrapolation tests, with the respective values of the three truncation criteria.

| Case | Operation | Train | Test | $r = r_{90}$ | $r = r^*$ | $\mathtt{rank}(\tilde{\mathbf{X}})$ |
|---|---|---|---|---|---|---|
| 0 | representation | $k = 1 : 1929$ | $k = 1 : 1929$ | 34 | 404 | 1928 |
| 1 | interpolation | $k = 1 : 2 : 1929$ | $k = 2 : 2 : 1929 - 1$ | 31 | 281 | 964 |
| 2 | interpolation | $k = 1 : 3 : 1929$ | $k = 2 : 3 : 1929 - 2$ | 31 | 279 | 642 |
| 3 | extrapolation | $k = 1 : 1 : 965$ | $k = 966 : 1 : 1929$ | 41 | 418 | 963 |

both ROMs can capture the salient features of the flow, reproducing the general velocity patterns. Hence, xDMD proves to be a suitable method to interpret and reproduce 3D cardiovascular simulations. It follows that xDMD-based ROMs can be successfully employed to replace high-fidelity simulations with compressed reconstructions. The rate of compression depends on the truncation applied to the SVD, which can be selected based on the accuracy required by the application.

To test the method's robustness, we train ROMs on datasets with missing spatial data. This reduces the dimension of the dataset but also generates a loss in spatial continuity. Specifically, $N/10$ and $N/100$ elements are randomly selected from the original mesh of size $N$ to obtain two reduced-size datasets. Representation error of the resulting ROMs, trained on all $M = 1928$ temporal snapshots, is shown in Figure 3.3, as a function of the truncation rank. Note that values of $r = r_{90}$, $r = r^*$, and $r = \mathtt{rank}(\tilde{\mathbf{X}})$ are the same for all the three cases. For low-rank ROMs, considering only the dominant spatiotemporal structures of the underlying flow, the data loss does not affect the accuracy. The effects of the missing spatial data increase with $r$, causing the error to reach tens of orders of magnitude for $r = \mathtt{rank}(\tilde{\mathbf{X}})$.

Another facet of xDMD's robustness is its sensitivity to the number of temporal snapshots available for training. Figure 3.4 shows the time-averaged representation error of ROMs trained on $N/100$ velocity measurements and 200 snapshots, as a function of the truncation rank. The 200 snapshots are selected from the full dataset of size 1929 using either the first 200 images or every $4^{\text{th}}$

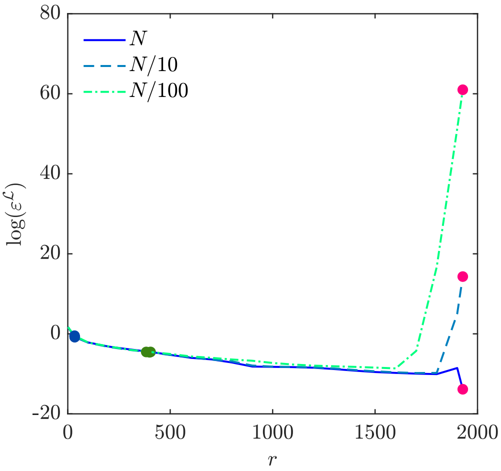

**Figure 3.3:** Dependence of time-averaged representation error of ROMs on the SVD truncation rank $r$. The ROMs are alternatively trained on the data in all $N$ pixels and on the data at randomly selected $N/10$ and $N/100$ pixels; in all three cases, using $M = 1928$ snapshots. The blue, green and red dots correspond to $r = r_{90}$, $r = r^*$ and $r = \mathtt{rank}(\tilde{\mathbf{X}})$, respectively.

or every 8th image. Both the reduction of the number of training snapshots, compared to the case in Figure 3.1, and the use of different time steps between the snapshots, do not affect the ROM's accuracy. Hence, we can state that the xDMD algorithm is robust and able to provide a good approximation of nonlinear flow phenomena.

### 3.3.2 Interpolation and Extrapolation Errors

The typical application of ROMs is predicting the values of a quantity of interest at space-time points wherein the output of high-fidelity simulations is unavailable. Therefore, in the second part of our study, we test the performance of our xDMD-based ROMs in both the interpolation and extrapolation regimes. Dimensions of

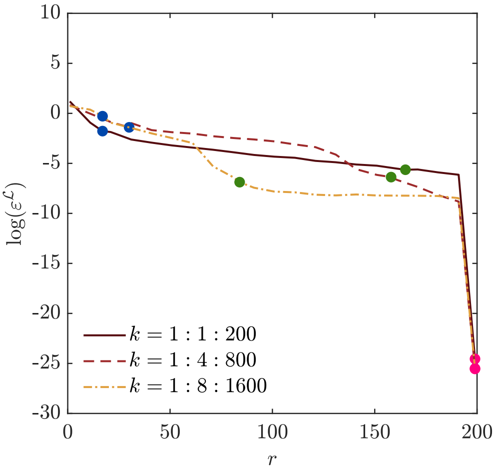

**Figure 3.4:** Dependence of time-averaged representation error of ROMs on the SVD truncation rank $r$. The ROMs are alternatively trained on the first 200 snapshots and on the 200 snapshots selected with time intervals 4 or 8; in all three cases, using $N/100$ pixels. The blue, green and red dots correspond to $r = r_{90}$, $r = r^*$ and $r = \texttt{rank}(\tilde{\mathbf{X}})$, respectively.

the training and testing matrices for all the cases considered and corresponding truncation ranks are reported in Table 3.1.

We start by constructing three ROMs associated with the truncation ranks $r = \texttt{rank}(\tilde{\mathbf{X}})$, $r = r^*$ and $r = r_{90}$, and trained on half of the snapshots (Case 1 in Table 3.1). Predictions are carried out for the missing half of time steps, and interpolation errors are computed through Equation (2.23) at interpolated instants $k$ not considered during calibration. The errors associated with the three ROMs and computed for each interpolated time step are plotted in Panel (a) of Figure 3.5. While truncating at $r = \texttt{rank}(\tilde{\mathbf{X}})$ leads to an anomalous trend of the error increasing in time, the ROMs truncated at $r = r^*$ and $r = r_{90}$ assure high accuracy and stability. When we reduce the size of the training set (Case 2 in Table 3.1) the three different ROMs (Panel (b) of Figure 3.5) present aligned error picks and similar periodicity. The magnitude of the errors is similar, with $r = r^*$ providing a slightly smaller error.

The capability of xDMD in predicting future values of the quantity of interest is evaluated by training the ROMs on the first half of the dataset and testing the performance on the remaining half (Case 3 in Table 3.1). The representation (training) and extrapolation (test) errors are shown in Figure 3.6. The ROM truncated at $r = \texttt{rank}(\tilde{\mathrm{X}})$ exhibits a trend that increases with time. We also notice a significant discontinuity between representation and extrapolation regimes that reveals the overfitting of the model on training data. More stable performances are associated with the ROMs truncated at $r = r^*$ and $r = r_{90}$: the error of the first ROM varies in $10^{-4} - 10^{0}$, while the error of the second one is in $10^{-6} - 10^{-3}$, with less visible but aligned negative picks and similar periodicity. These results provide actionable indicators for the truncation rank selection and clarify the role played by the non-dominant modes. When all the modes are included in the training phase, $r = \texttt{rank}(\tilde{\mathbf{X}})$, the ROM suffers from overfitting and loses accuracy in interpolation and extrapolation; this is true regardless of how many data are used for training. In addition, the loss in accuracy is difficult to predict given the unstable behavior of the

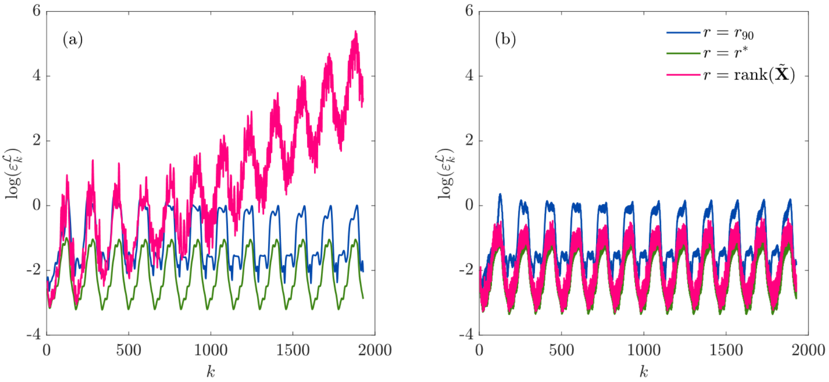

**Figure 3.5:** Interpolation errors for (a) Case 1 and (b) Case 2 in Table 3.1. In each plot different lines correspond to the ROMs with different truncation ranks: $r = r_{90}$, $r = r^*$ and $r = \mathtt{rank}(\tilde{\mathbf{X}})$.

error. Hence, the use of low-rank truncation not only aligns with a ROM's purpose (identification of the dominant modes) but also increases the ROM's prediction reliability at space-time locations where high-fidelity data are not available.

To provide a local view of the accuracy of the ROMs, in Figure 3.7 and Figure 3.8 we compare the velocities simulated by the HFM in 4 points of the geometry, with the corresponding series of values provided by the ROMs truncated at $r = r_{90}$ and $r = r^*$, respectively. Panel (a) and (b) refer to the first and second aorta's cross-section, where we pick one point near the wall (blue) and one near the center of the aorta (red). As expected, the ROM truncated at rank $r = r^*$ (Figure 3.8) has high accuracy both in representation and extrapolation for all the points considered. The ROM truncated at $r = r_{90}$ (Figure 3.7), instead, adequately reproduces the overall system state but loses accuracy, especially in extrapolation as time increases. The ROM's performance is not affected by the selection of the points near the wall or in the middle of the aorta.

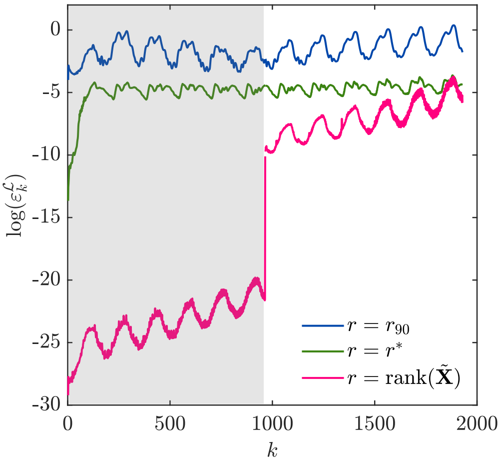

**Figure 3.6:** Representation (shaded) and extrapolation (not shaded) errors for Case 3 in Table 3.1. In each plot different lines correspond to the ROMs with different truncation ranks: $r = r_{90}$, $r = r^*$ and $r = \texttt{rank}(\tilde{\mathbf{X}})$.

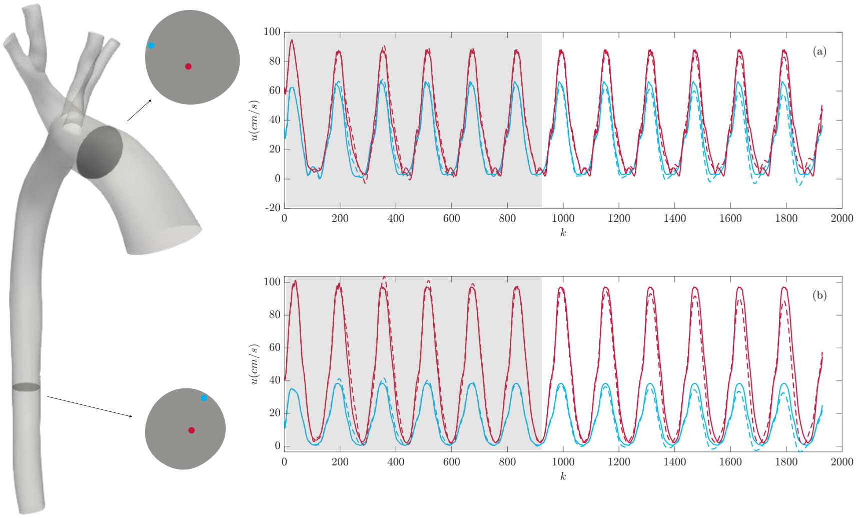

**Figure 3.7:** Flow velocity $u$ at $k$th time step, provided by the HFM (continuous line) and estimated by the ROM (dashed line) in the representation (shaded) and extrapolation (not shaded) regimes. The ROM is trained for Case 3 in Table 3.1 and truncated at $r = r_{90}$. Panels (a) and (b) show data for the two aorta's cross-sections shown on the left. The red series refer to the points located near the center of the aorta, and the blue ones to the points near the wall.

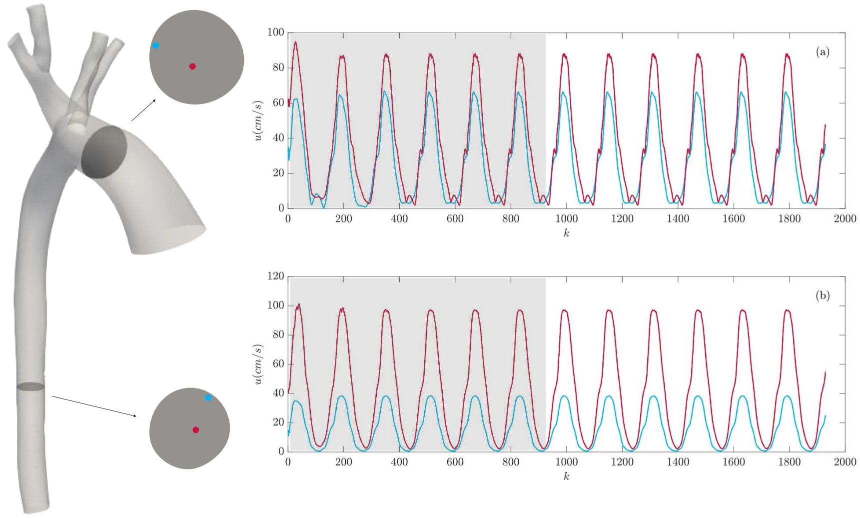

**Figure 3.8:** Flow velocity $u$ at $k$th time step, provided by the HFM (continuous line) and estimated by the ROM (dashed line) in the representation (shaded) and extrapolation (not shaded) regimes. The ROM is trained for Case 3 in Table 3.1 and truncated at $r = r^*$. Panels (a) and (b) show data for the two aorta's cross-sections shown on the left. The red series refer to the points located near the center of the aorta, and the blue ones to the points near the wall.

## 3.4 Final Remarks

We analyzed the performance of the extended version of DMD (xDMD) proposed by Lu and Tartakovsky (2021) on the task of ROM construction to approximate the high-fidelity simulations of 3D blood flow in a patient-specific aorta. Our results show that xDMD can identify dominant spatiotemporal structures in the high-fidelity dataset and provide an accurate approximation of high-fidelity simulations. The SVD truncation rank choice proved to be crucial in ROM generation to balance accuracy, generalizability, and computational saving. We explored the ROMs' ability in representation, interpolation, and extrapolation regimes, and related the performance to the choice of the truncation rank, defining optimal truncation criteria. We found that a low-rank truncation, which preserves almost all the cumulative energy in the data, retaining only the singular values corresponding to the most relevant spatiotemporal structures, avoids overfitting and yields high accuracy and error stability. Optimal truncation guarantees the successful use of DMD as a data compression technique, but also to predict values of the quantity of interest at space-time points wherein the output of the HFM is not available. Given an appropriate truncation, the xDMD-based ROMs also demonstrate a remarkable robustness to the number of space-time training data. Finally, the accuracy of xDMD was also verified locally, when used to predict time series at selected points in the flow domain. Overall, this study suggests that the use of xDMD, when optimal truncation criteria are applied, is beneficial for time-dependent data compression and computational saving when used in place of onerous HFMs.

# 4 | DMD for Global-scale Time Series

In several diverse scientific fields, the improved capability to collect observations of real systems evolution needs to be paired and supported by efficient methods to successfully impact complex problem modeling. In this context, data-driven approaches are emerging as key tools for capturing and translating information stored in high-dimensional datasets into valuable evidence to feed models and interpret phenomena.

Here, we present Dynamic Mode Decomposition (DMD) as a new viable approach to extract reduced-order models (ROMs) from high-dimensional datasets of observations. Differently from the case study described in Chapter 3, high-fidelity data are collected directly as measurements/observations of the real system behavior, and no high-fidelity model (HFM) is available.

In particular, we explore the capabilities of both the standard (sDMD) and extended (xDMD) versions of the method to learn the dynamics of global water distribution from high-dimensional satellite observations. The DMD-based ROMs are derived from the latest release of total water storage anomaly (TWSA) data provided by the Gravity Recovery and Climate Experiment (GRACE) and its successor, the GRACE Follow-On (GRACE-FO) satellite missions. Their performances are tested in terms of reproduction of training snapshots and extrapolation of future states of the system.

The study demonstrates that the idea of using snapshots of a system evolving in time to derive a linear ROM of a phenomenon perfectly fits the case of satellite data interpretation. This novel application proves that DMD can detect patterns, extract ROMs, and predict climate-related time series based on previous satellite observations. This method could enable analyses of data with improved spatial and temporal resolution currently available from satellite technologies, leading to significant advancements in modeling and understanding the water cycle. The impacts of climatic and anthropogenic pressures on water availability worldwide could be identified, analyzed, and quantified, deriving important evidence for the development of mitigation and adaptation strategies.

The chapter is organized as follows: Section 4.1 describes the opportunities deriving from the application of data-driven techniques, in particular DMD, to satellite observations; in Section 4.2 we present the GRACE and GRACE-FO missions and specify the details of our case study; in Section 4.3 results are presented and discussed; Section 4.4 closes the chapter with a set of final remarks.

The study presented in this chapter is related to the paper V. Ciriello, G. Libero, D.M. Tartakovsky, *Reconstruction of GRACE satellites data via Dynamic Mode Decomposition* expected to be submitted soon to Advances in Water Resources.

## 4.1 Data-Driven Techniques and Satellite Data

The sky above us is teeming with satellites, launched by international agencies, such as the well-known NASA and ESA, but also by emerging private companies, to serve different scopes, from communications to military aims, from astronomic observations to Earth monitoring. In the early 1970s, the launch of the first in a series of Landsat satellites developed by NASA and operated by USGS paved the way for modern land surface satellite observations. Since then, dozens of satellite missions have been developed by the major Space Agencies, which have also initiated strong collaborations to join efforts, resources, and expertise to strengthen our global understanding of Earth's evolution. These Earth-observing satellites can measure with improved spatial coverage and temporal resolution variables describing the physical and chemical processes of land, ocean, atmosphere, and cryosphere. These observations allow for a better understanding of how the Earth's systems evolve and support the development of strategies for humans' adaptation to changes.

An important field that could benefit from the use of the vast amount of satellite data collected is the study of all the phenomena related to water distribution on the Planet. The terrestrial water cycle is a notoriously complex and sensitive balance of thousands of different phenomena, increasingly forced and perturbed by climatic and anthropogenic pressures which have significant impacts on water availability worldwide (IPCC, 2023a). To identify and design appropriate strategies for adaptation and mitigation, both monitoring and modeling activities are crucial to get a deeper understanding of global hydrological processes and support the sustainable exploitation of water resources. In this context, satellite technologies can offer a valuable alternative to traditional onerous monitoring campaigns (Sheffield et al., 2018). While local modeling of the water cycle can be based on detailed in-situ measurements, the high costs and times required for the collection of these observations make their use impracticable for global analysis. Satellites, instead, with their wide and timely-regular coverage, provide more homogeneous and cheaper datasets, particularly suitable for global assessments.

Translating satellites' observations into valuable information to feed both physics-based and data-driven modeling approaches, though, requires appropriate tools and efforts. In particular, the large availability of observations and their increasing level of detail pose new challenges in their use and interpretation. Data-driven methods, able to discover constitutive relations of the observed complex real-world problems directly from data, represent a promising tool in this context. Thanks to their efficiency and flexibility, they can unmask the physics hidden in these datasets without compromising speed and accuracy (Tartakovsky et al., 2020; Asher et al., 2015).

A commonly employed technique is represented by Neural Networks. Several literature examples discuss the efficiency and accuracy of all sorts of Neural Networks (He and Tartakovsky, 2021; Lu et al., 2020; Qin et al., 2019; Pan et al., 2019). However, they are "black box" models, their structure remains uninterpretable and cannot provide any insights into the functions being approximated and the physics governing the system cannot be discovered through them.

Differently, Dynamic Mode Decomposition (DMD) derives directly from satellite observations interpretable reduced-order models (ROMs), which can both efficiently reproduce the system state and learn its primary features in space and time. As discussed in Section 2.2, the connection between the DMD method and the Koopman theory, explored in Mezić (2013) and Rowley et al. (2009), guarantees the possibility of physically interpreting the structure of the model to characterize the system's behavior. Major advantages can therefore derive from the application of DMD when traditional physics-based models are too complex to design and employ, but a large amount of observations of the spatial and temporal evolution of the system are available. This is the case with vast satellite data archives. DMD-based models can be used not only as faster and cheaper ROMs, but their structure can also be interpreted to discover patterns and physical features in the phenomena. Despite the increasing diffusion of DMD in very different fields (Ohh and Spedding, 2022; Pan et al., 2021; Liu et al., 2021; Brunton et al., 2016; Mann and Kutz, 2016; Erichson et al., 2016; Proctor and Eckhoff, 2015), to our knowledge only a few recent studies discussed the application of this technique to satellite data (Chi-Durán and Buffett, 2023; Fulton and Hegerl, 2021; Hogg et al., 2020).

In this chapter, we demonstrate how efficient and accurate ROMs of complex real-world phenomena can be derived using DMD directly from satellite data. Specifically, we focus on data concerning water distribution on Earth. The DMD method could be an innovative technique to thoroughly investigate the information contained in the precious satellite mission products and therefore to gain a deeper understanding of how the hydrological cycle works.

## 4.2 Data and Application

### 4.2.1 GRACE and GRACE-FO Missions

The Gravity Recovery and Climate Experiment (GRACE) satellite mission (Tapley et al., 2004), which was in orbit from April 2002 to October 2017, and its successor, the GRACE Follow-On (GRACE-FO) (Landerer et al., 2020), launched in 2018 and still working, take advantage of gravity field measurements, to derive information about mass distribution changes on Earth. In almost 20 years of operation, GRACE missions provided pioneering and detailed measurements of Earth's gravity field changes, revolutionizing research about Earth's water reservoirs over land, ice, and oceans, as well as earthquakes and crustal deformations. Each mission counts a pair of orbiting satellites continuously tracking how their relative positions in space are affected by Earth's gravity field variations. When the satellites encounter a change in the distribution of Earth's mass, the distance between them changes, as represented in Figure 4.1.

By precisely and continuously tracking this variation, regional and temporal gravity changes can be measured with high precision. These changes reflect mass distribution variations on the planet, primarily due to changes in water content as it moves through the water cycle. Hence, GRACE missions allow the monitoring of global water distribution evolution in time, considering ice sheets and glaciers melting, surface and underground water storage depletion, and changes in sea level and ocean currents (Figure 4.2).

All the measurements from the satellites are sent to three data processing centers, where they are translated into monthly matrices of centimeters of equivalent water thickness anomalies with respect to a baseline mean (Jan 2004 - Dec 2009) (Watkins et al., 2015). Global monthly maps of total water storage anomaly (TWSA) are available from April 2002 to the present, with some short-term gaps and a major interruption due to the transition between the missions. The native resolution is 3 degrees in both latitude and longitude, but a 0.5-degree scaling factor map is provided to compensate for the signal attenuation during sampling and post-processing (Wiese et al., 2016). This also acts as a mask, removing oceans and areas around the poles, where specific solutions should be considered. Both the collection of TWSA monthly maps and the scale factor map are freely downloadable from the NASA archives and from the specific data storage systems of the three processing centers, where all the guidelines for GRACE data interpretation are also available. The importance of this exclusive set of global water distribution observations is proved by the publication of several studies where GRACE data are analyzed to quantify the impact of climatic and anthropogenic changes on water distribution at regional and global scales (Rodell and Reager, 2023; Thomas and Famiglietti, 2019; Rodell et al., 2018; Castle et al., 2014; Joodaki et al., 2014; Döll et al., 2014; Famiglietti et al., 2011; Leblanc et al., 2009).

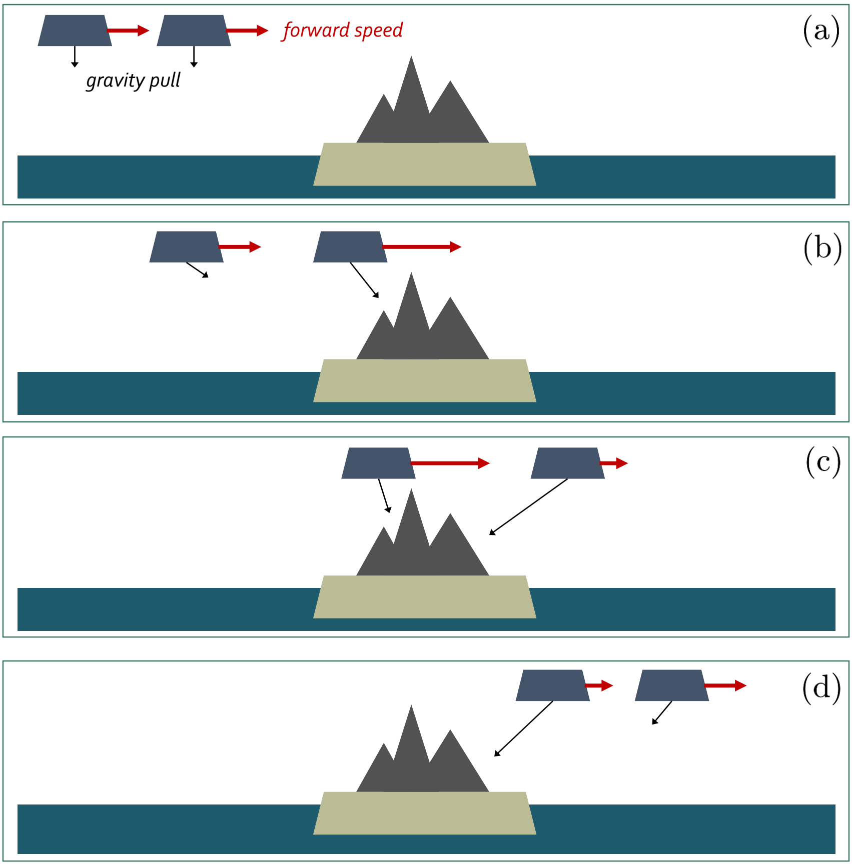

**Figure 4.1:** Simplified example of how GRACE and GRACE-FO missions measure gravity changes: (a) both spacecrafts fly over the ocean at a constant distance; (b) Spacecraft 1 encounters land and is accelerated pulled by land's higher gravity force; (c) Spacecraft 2 encounters land too, while Spacecraft 1 is pulled back slightly by the higher gravity of land before reaching the ocean again; (d) the spacecrafts return to their original inter-distance.

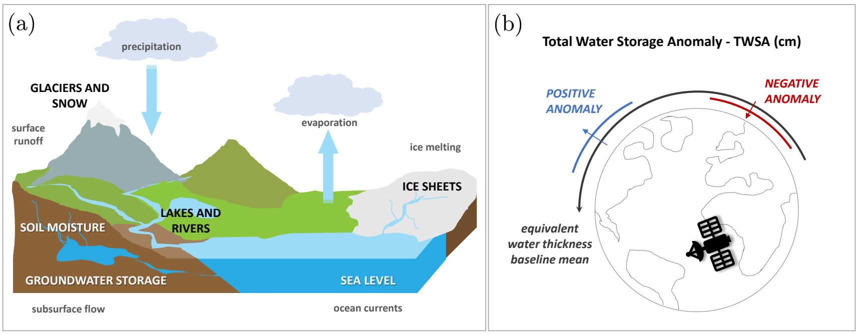

**Figure 4.2:** GRACE missions (a) monitor variations in water content as it moves through the water cycle and (b) translate them into cm/year of anomalies of equivalent water thickness relative to the baseline mean 2004-2009.

### 4.2.2 Definition of dataset for the analysis

For this study, the most recent GRACE product (RL06.1Mv03) was downloaded from the Jet Propulsion Laboratory (JPL) archive (NASA/JPL, 2023), together with the scaling factor file and a list of the months of activity of the two missions. Note that here we focus only on the land areas; dedicated products can be downloaded from the same archive if interested in ocean studies. Differently from simulated datasets, the temporal continuity of this series of real measurements is interrupted by some gaps, as can be seen in Panel (b) of Figure 4.3. The dataset collects 215 monthly snapshots of TWSA, captured in the time interval of $T = 248$ months between April 2002 ($t = 1$) and November 2022 ($t = 248$). Each snapshot is a map, which counts $N = 58908$ grid elements placed on land areas. All the other elements of the mesh located on the oceans or around the Poles are automatically set as void when the scale factor is applied and therefore not considered in the model generation. Panel (a) of Figure 4.3 represents the snapshot corresponding to November 2022.

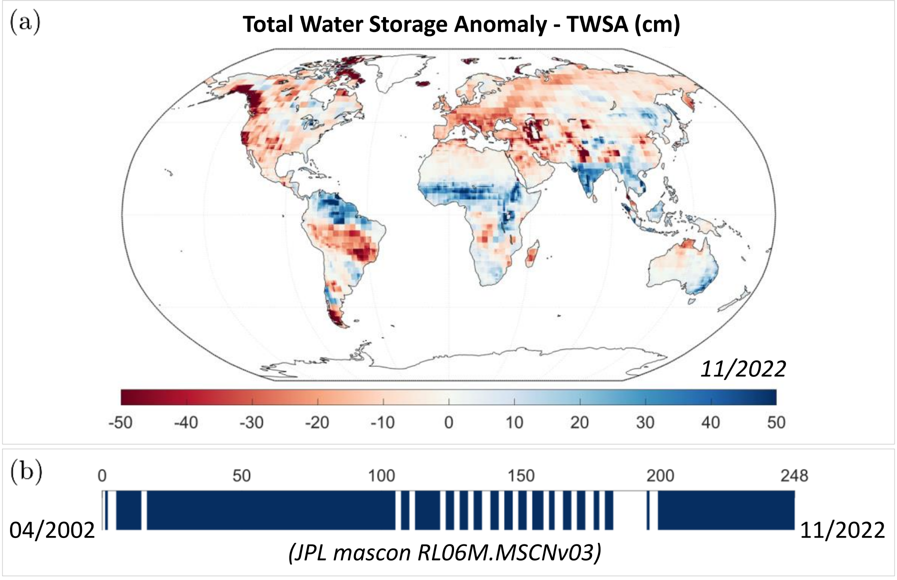

**Figure 4.3:** (a) Example of TWSA map for $t = 248$ and (b) months of activity of GRACE and GRACE-FO missions.

While the majority of the literature concerning GRACE and GRACE-FO is based on the construction of empirical relationships between TWSA and related hydrologic variables for the interpolation of missing observations (Li et al., 2020; Sun et al., 2020), here we aim at learning DMD-based ROMs able to both reconstruct and extrapolate global water distribution time series, solely depending on previous GRACE observations. We apply both the standard algorithm of DMD (sDMD) and the extended version (xDMD), described in Chapter 2, and we analyze their performances in terms of reconstruction of training data and extrapolation of snapshots of the future state of the system. The study also includes a preliminary analysis of truncation criteria for an efficient and accurate model reduction.

## 4.3 Results and Discussion

### 4.3.1 Reconstruction of Time Series

As a first step, both the standard (sDMD) and the extended (xDMD) version of the DMD algorithm are employed to generate ROMs based on the entire dataset of available snapshots. To the aim of our analysis, we need pairs of snapshots equally shifted in time to build the matrices $\mathbf{X}$ and $\mathbf{X}'$ for DMD training (see Section 2.2.1 for matrix construction). Hence, we extract, from the original dataset of 215 snapshots, $M = 196$ couples of data shifted by $\Delta t = 1$month and train the ROMs on them. The ability of both methods to reproduce their same training set is tested and quantified as representation error.

As introduced in Section 2.2.4, a key role in ROM generation is played by the selection of the Singular Value Decomposition (SVD) truncation rank, which should guarantee an optimal balance of accuracy and generalizability of the model. Here, as we did in the application presented in Chapter 3, we explore the accuracy of a sequence of ROMs obtained truncating the SVD at different ranks and we test the truncation criteria defined in Section 2.2.4.

**Table 4.1:** Data sets and ranks considered for (1) representation and (2) extrapolation tests.

| | Train | Test | $r_{90}$ | $r^*$ | `rank`$(\mathbf{X})$ |
|---|---|---|---|---|---|
| (1) | $M = 196$ in $t = 1:248$ | $M = 196$ in $t = 1:248$ | 79 | 122 | 195 |
| (2) | $M_{tr} = 148$ in $t = 1:200$ | $M_{ts} = 48$ in $t = 200:248$ | 72 | 124 | 147 |

In particular, we compare $r = r_{90}$, equal to the minimum number of singular values accounting for 90% of the cumulative energy in the SVD of $\mathbf{X}$, and $r = r^*$, identified setting $\zeta = 10^{-3}$ in Equation 2.20, with $r =$ `rank`$(\mathbf{X})$, computed with the standard `MATLAB` function and resulting in no truncation.

The curve of the cumulative energy in Panel (a) of Figure 4.4 shows how the informative content carried by the training data is decomposed and distributed among the elements of the SVD. As expected from its definition, the cumulative energy associated with $r =$ `rank`$(\mathbf{X})$ is equal to 1; 10% of the information is cut instead when $r_{90}$ is considered; while truncating at $r^*$ keeps around 95% of the energy. All truncation rank values are listed in Table 4.1.

The representation error, computed with Equation 2.23 and averaged over all the time steps, is evaluated for different truncation ranks and plotted in Panel (b) of Figure 4.4 for sDMD-based ROMs and in Panel (c) for xDMD-based ROMs. Both the curves are steeper at the beginning, where maintaining one more single element in the SVD truncation corresponds to a big gain in information and therefore in model accuracy. The error continues its decay slower in the middle so that truncating at $r_{90}$ or $r^*$ yields very similar performances. The end of the curve, instead, marks the difference between the standard and the extended version of the algorithm: while the xDMD-based ROM without truncation ($r =$ `rank`$(\mathbf{X})$) reaches an error of the order of almost $10^{-30}$, in the case of sDMD the performance of the model with no truncation is around $10^{-3.5}$.

To deeper investigate the performances of both the DMD versions in the representation regime, in Figure 4.5 we compare how the errors associated with the sDMD-based (Panel (a)) and xDMD-

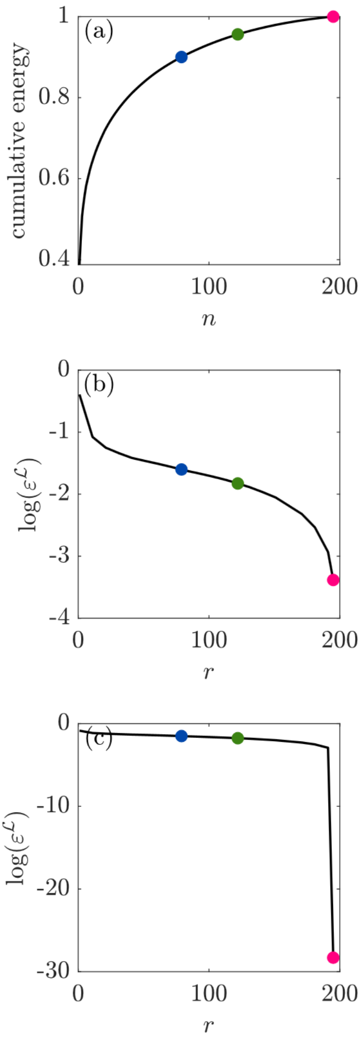

**Figure 4.4:** (a) Cumulative energy associated with the SVD of $\mathbf{X}$ plotted as function of the singular values number $n$. Representation error of (b) the sDMD-based ROM and (c) the xDMD-based ROM (averaged over the time instants) as a function of the truncation rank $r$ of the SVD. In all panels, the blue, green, and red dots correspond to $r = r_{90}$, $r = r^*$, and $r = \texttt{rank}(\mathbf{X})$, respectively.

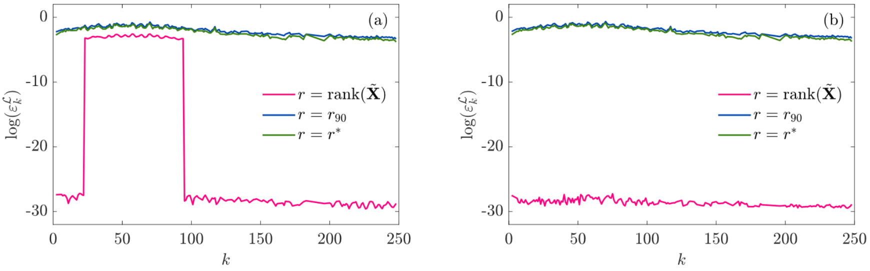

**Figure 4.5:** Representation errors of (a) the sDMD-based and (b) the xDMD-based ROMs. In each plot different lines correspond to the ROMs with different truncation ranks: $r = r_{90}$, $r = r^*$, and $r = \mathtt{rank}(\mathbf{X})$.

based (Panel (b)) ROMs, truncated at $r_{90}$, $r^*$ and $r = \mathtt{rank}(\mathbf{X})$, evolve in time. As already announced by Figure 4.4, for both sDMD and xDMD, truncating at $r_{90}$ or $r^*$ yields very close performances, with a little advantage for $r^*$. More noticeable and interesting is the behavior of the error when $r = \mathtt{rank}(\mathbf{X})$: while xDMD performance oscillates near $10^{-30}$ for the entire time interval, sDMD shows a localized increase for $t = 23 \div 94$. This time interval corresponds to months between January 2004 and December 2009, the same taken as the baseline of GRACE missions data. Considering that all GRACE data are provided as anomalies with respect to that reference interval—which means that the baseline mean is subtracted from all the measurements—the subset of snapshots in 2004-2009 has zero mean, while the rest of the dataset has not. This peculiarity of the GRACE dataset is better interpreted by the xDMD, which can count on the bias term and the residual learning (Equation 2.16) to better model this behavior. However, this advantage is evident only when no truncation is applied to the SVD, otherwise, sDMD and xDMD have very similar performance.

In Figure 4.6 the representation performances of both the formulations of DMD are analyzed from a spatial point of view. On

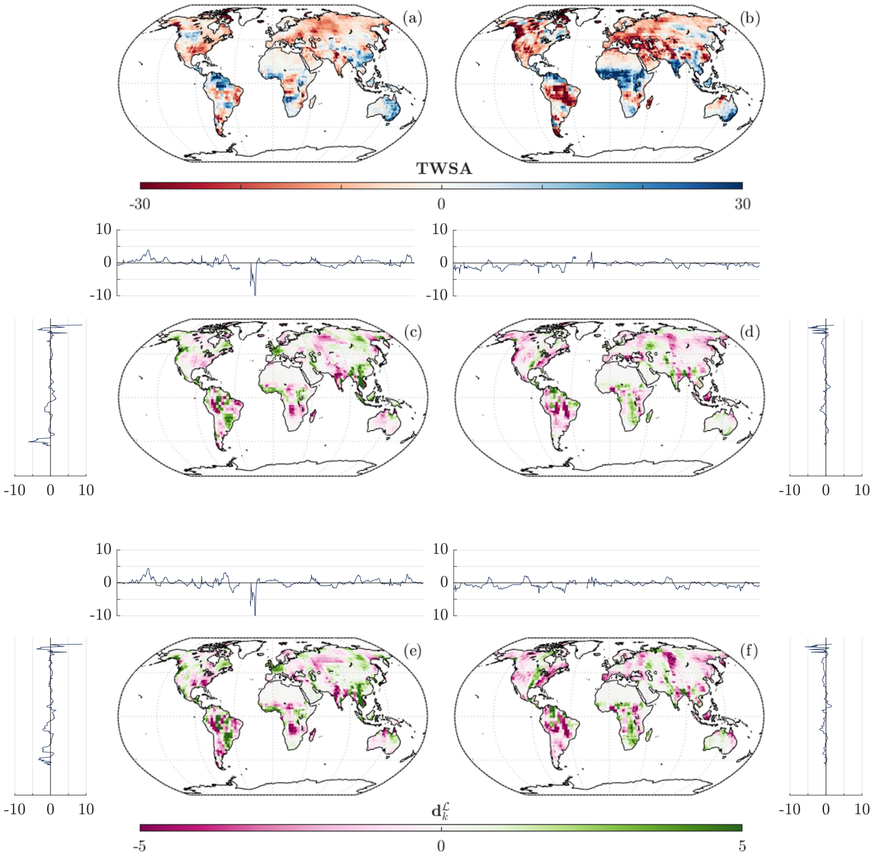

**Figure 4.6:** Maps of TWSA provided by GRACE at times (a) $t = 124$ and (b) $t = 248$ and respective differences of the reconstructions provided by (c-d) the sDMD-based ROM and (e-f) the xDMD-based ROM, both truncated at $r_{90}$. In panels (c-f) the curves represent the mean of the differences by latitude (vertical) and longitude (horizontal).

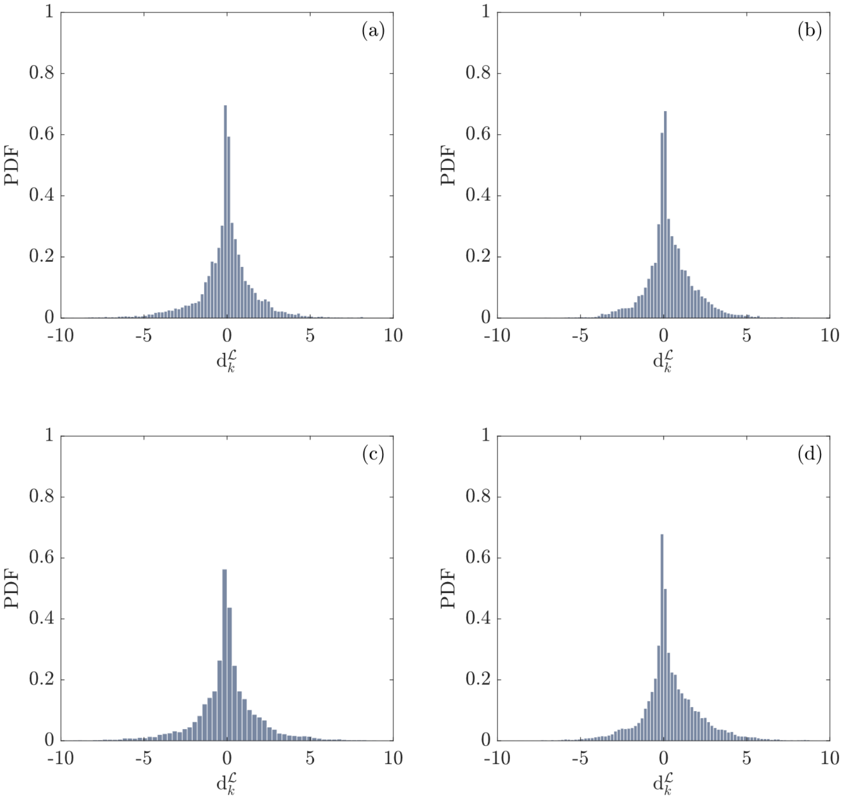

**Figure 4.7:** Probability distribution of the differences, mapped in Figure 4.6, between the reference maps of GRACE for time $t = 124$ and $t = 248$ and the respective reconstructions provided by (a-b) the sDMD-based ROM and (c-d) the xDMD-based ROM, both truncated at $r_{90}$.

the first row (Panels (a) and (b)), two maps, serving as references, show the original values of TWSA provided by GRACE for $t = 124$ and $t = 248$. The differences $\mathbf{d}_k^{\mathcal{L}}$ between them and the respective reconstructions provided by DMD algorithms are computed through Equation 2.22 and plotted in the remaining Panels. Panels (c) and (d) show how sDMD reconstructions for $t = 124$ and $t = 248$ differ from the respective original snapshots in Panels (a) and (b), while Panels (e) and (f) refer to the xDMD approximations for the same times. All the ROMs are truncated at $r_{90}$. In the xDMD case, we notice locally higher differences - for example in the South of Africa, in Siberia, or on the East Coast of North America - but the global patterns and order of magnitude are similar to those observed in the sDMD case. Moreover, the level of accuracy remains unchanged from time $t = 124$ to $t = 248$, as already evinced from Figure 4.5.

Each of the 4 maps in Panels (c)-(f) is accompanied by two curves: in the vertical graph is the mean by row of the differences plotted in the map; in the horizontal plot, instead, is the mean by column. These two curves serve to identify possible inhomogeneities in ROMs' behavior in space. Higher differences are reported for higher latitudes, where the presence of perennial ice and topographic adjustments affect the accuracy of original data and as a consequence the performance of the ROMs too. Changing the longitude, instead, doesn't seem to have a significant impact on the error, except for a negative pick in Panels (c) and (e), located in a predominantly oceanic meridian. The small number of land pixels in this longitude range are all located in Iceland and hence probably affected by the same difficulties described above.

The difference between the original TWSA by GRACE and the DMD reconstructions is computed locally so that at each time $t_k$ corresponds a vector of differences $\mathbf{d}_k^{\mathcal{L}}$ containing $N = 58908$ estimated values $\mathrm{d}_k^{\mathcal{L}}$, one for each pixel located on a land area in the map. The probability distribution of these difference values is described by the histograms of Figure 4.7. Each plot symmetrically refers to the corresponding map in Panels (c)-(f) of Figure

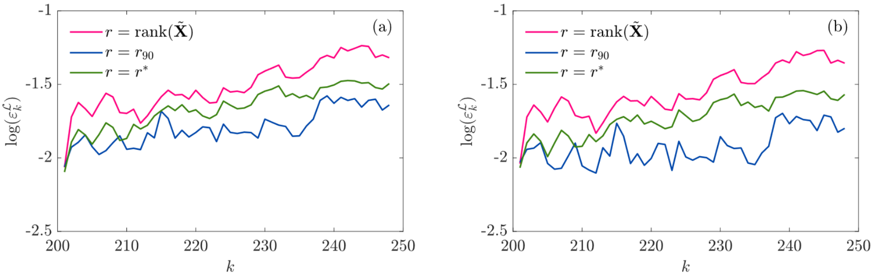

**Figure 4.8:** Extrapolation errors of (a) the sDMD-based and (b) the xDMD-based ROMs. In each plot different lines correspond to the ROMs with different truncation ranks: $r = r_{90}$, $r = r^*$, and $r = \texttt{rank}(\mathbf{X})$.

4.6. The distributions are approximately Gaussian. The majority of the cells are included in the classes near $\mathrm{d}_k^{\mathcal{L}} = 0$, and the probability quickly decays to zero for values of $|\mathrm{d}_k^{\mathcal{L}}| > 5$.

### 4.3.2 Extrapolation of Time Series

A typical application of ROMs is the prediction of the values of the quantity of interest at space-time points where measured data or high-fidelity simulations are unavailable. To test the performance of our ROMs in the extrapolation regime, the original GRACE dataset is split into a training set, counting $M_{tr} = 148$ pairs of snapshots, extracted from $t = 1$ to $t = 200$, and a test set made with the remaining $M_{ts} = 48$ pairs, from $t = 201$ to $t = 248$. The first group of data is employed to train both sDMD and xDMD models, which are then used to predict the following $M_{ts}$ states of the system. The performance is assessed by comparing the DMD predictions with the test set elements. The same truncation criteria applied in the previous section are employed here: all the corresponding values are reported in Table 4.1.

Panel (a) of Figure 4.8 shows the evolution in time of the extrapolation error of the sDMD-based ROM truncated at $r_{90}$, $r^*$

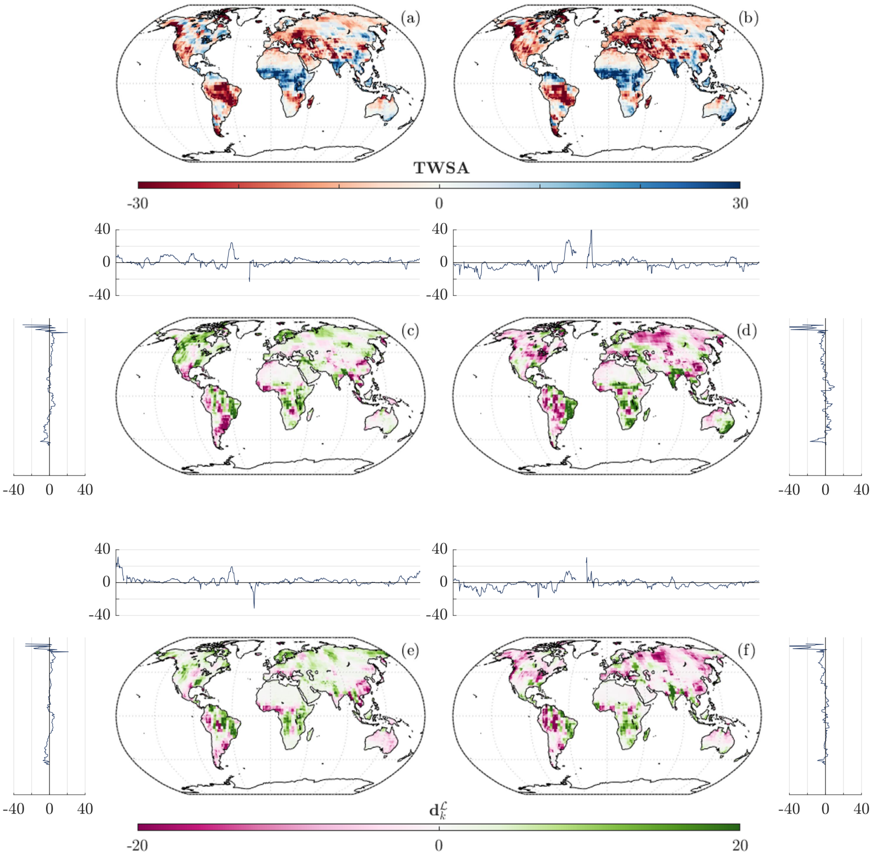

**Figure 4.9:** Maps of TWSA provided by GRACE at times (a) $t = 224$ and (b) $t = 248$ and respective differences of the predictions provided by (c-d) the sDMD-based ROM and (e-f) the xDMD-based ROM, both truncated at $r_{90}$. In panels (c-f) the curves represent the mean of the differences by latitude (vertical) and longitude (horizontal).

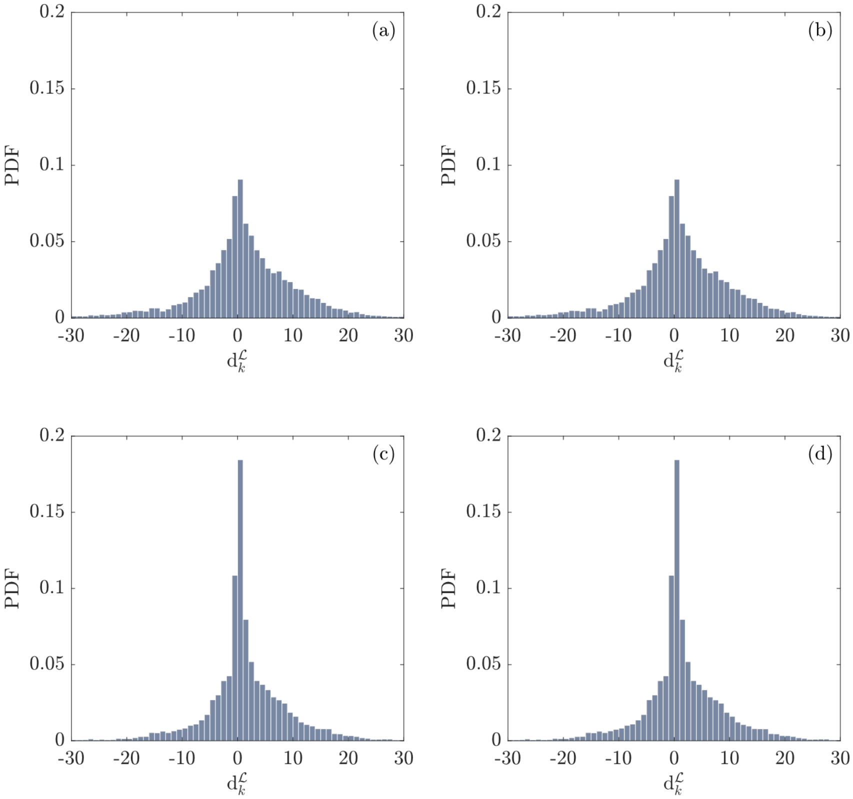

**Figure 4.10:** Probability distribution of the differences, mapped in Figure 4.9, between the reference maps of GRACE for time $t = 224$ and $t = 248$ and the respective predictions provided by (a-b) the sDMD-based ROM and (c-d) the xDMD-based ROM, both truncated at $r_{90}$.

and $r = \texttt{rank}(\mathbf{X})$; Panel (b) shows the same for xDMD. For both the DMD formulations, truncating at $r_{90}$ guarantees the best performance, while the selection of $r = \texttt{rank}(\mathbf{X})$, which provided the lowest representation error (Figure 4.5), becomes the worst option in the extrapolation regime. In the following, we will always truncate at $r_{90}$. As expected, all the ROMs have slightly higher errors when used to extrapolate future states of the system, compared to the use in the representation regime (Figure 4.5), where the model was asked to reproduce data already seen during the training phase. A slowly increasing trend of the error can also be noticed as we extend the number of predicted time steps. The difference between sDMD and xDMD is quite small to be noticed when $r = r^*$ or $r = \texttt{rank}(\mathbf{X})$, it becomes more visible for $r = r_{90}$.

Figure 4.9 focuses on the distribution of the ROMs' performances on the global map. The structure is the same as Figure 4.6, but here the maps refer to $t = 224$ and $t = 248$, as the focus is specifically on the extrapolated part of the series ($t = 201 \div 248$). The TWSA maps provided by GRACE for $t = 224$ and $t = 248$ are plotted on the first row, in Panels (a) and (b). The differences between them and the corresponding maps predicted by sDMD are on the second row (Panels (c) and (d)), while the plots of the third row (Panels (e) and (f)) refer to the xDMD case. Compared to the representation regime (Figure 4.6), here the differences between original data and DMD predictions are bigger and tend to increase moving from $t = 224$ to $t = 248$. The xDMD-based ROM has a slightly better performance than sDMD, more evident for $t = 248$. The curves plotted over and next to each map confirm major difficulties of interpretation and therefore prediction for higher latitudes, as it was in Figure 4.6.

The histograms in Figure 4.10, symmetrically paired to the maps of Panels (c)-(f) in Figure 4.9, describe the probability distributions of the difference values in each map. The shape of the distribution does not change from $t = 224$ (first column) to $t = 248$ (second column) but is affected by the choice of the DMD version. In the xDMD case (Panels (c) and (d)) values of difference near

$\mathrm{d}_k^{\mathcal{L}} = 0$ are more likely to happen, according to the more narrow and high shape of the distribution.

The maps in Figure 4.6 and 4.9 already proved that DMD is able to reproduce and even predict the global patterns of TWSA. Figure 4.11 offers a more detailed evaluation of the local performance. For 6 random points on the Planet, the time series of TWSA provided by GRACE are compared to the reconstructions ($k \leq 200$) and predictions ($k > 200$) computed by the xDMD-based ROM. Reconstruction is successfully performed in all situations. Differences are more evident after $k = 200$, especially for points (e) and (f), even there, though, the ROMs can reproduce a plausible behavior of the variable.

## 4.4 Final Remarks

We tested the viability of DMD as an efficient data-driven technique to interpret satellite data series. In particular, we analyzed the performances of the standard algorithm (sDMD) and the extended implementation (xDMD) on the task of reconstruction and extrapolation of data from GRACE missions. Our results confirm that a careful selection of the SVD truncation rank is a key factor in the generation of accurate DMD-based ROMs. A low-rank truncation, which preserves the majority of the total cumulative energy in the data, avoids overfitting and yields high accuracy and error stability. Given an optimal truncation choice, the reconstruction of GRACE observations is accurately performed by both the sDMD and xDMD algorithms. Some major difficulties are reported only at higher latitudes, where GRACE data are known to be less reliable. As for extrapolation of future TWSA on the globe, the robustness of xDMD formulation makes it preferable compared to sDMD, providing more stable performance. Finally, we verified the local accuracy of xDMD when used to reconstruct and predict time series at selected points on the global map.

Overall, our work suggests that the use of DMD is beneficial for the interpretation of satellite data series. As expected, the

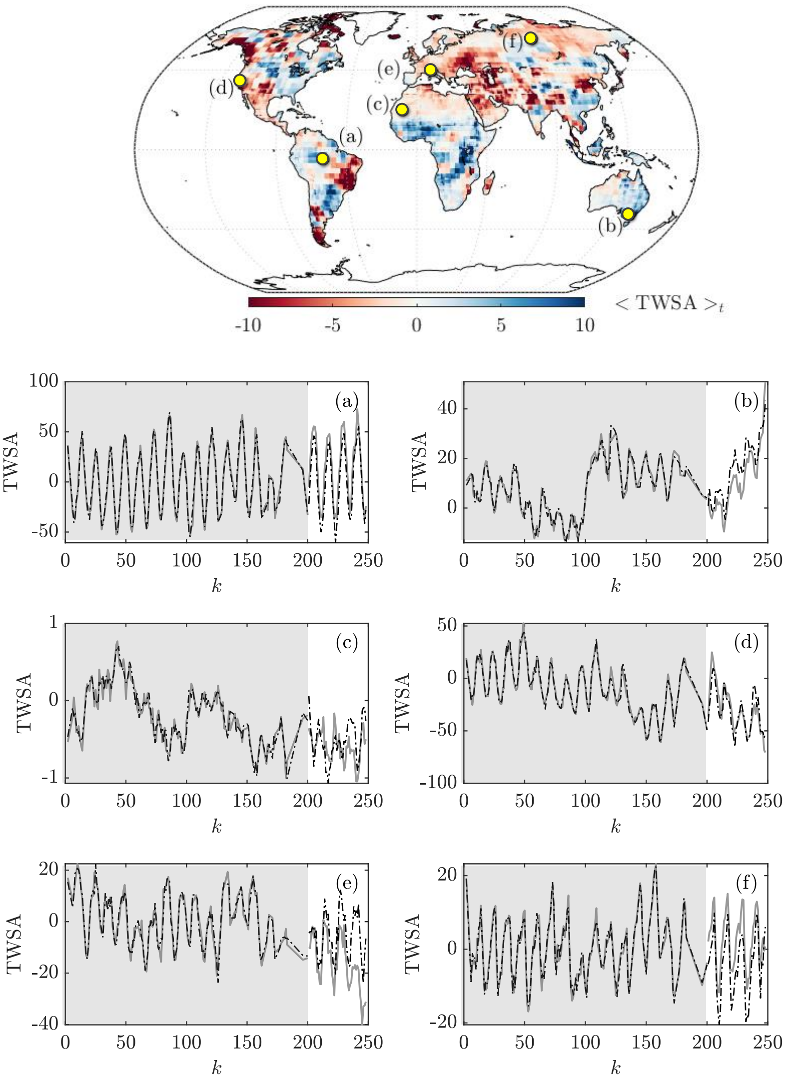

**Figure 4.11:** (a-f) Time series of TWSA for the six points on the map, provided by GRACE (continuous grey line) and estimated by the xDMD-based ROM truncated at $r_{90}$ (dashed line) in the representation (shaded) and extrapolation (not shaded) regimes. The reference map represents the mean over time of TWSA by GRACE.

performances of the ROMs are strictly related to the reliability of the training data, hence the typical inhomogeneity in measurements reflects on the final accuracy of the model. For this reason, the magnitude of the error is higher when we derive ROMs from observations than simulations (Chapter 3) and some localized difficulties appear. However, given the appropriate truncation, computationally efficient and accurate ROMs can be learned directly from satellite observations and replace HFMs in system modeling.

The applicability of DMD to satellite data also paves the way for future studies regarding the structure of the DMD-based ROM and its connections to the systems' leading order dynamics. More benefits are expected from this further step: the complex physics that stands behind the data collected by the satellites could be unveiled by analyzing the spatial patterns and temporal trends contained in the DMD-based model.

# 5 DMD-Informed Polynomial Chaos Expansion

As uncertainty usually characterizes real-world phenomena, to strengthen our ability to interpret complex systems we are generally interested in exploring the space of variability of the parameters and evaluating the effects on the system's response. This task requires large sets of simulations, resulting in prohibitive computational costs if performed through high-fidelity models (HFMs).

Even if Dynamic Mode Decomposition (DMD) is a powerful tool that allows for low-cost simulations of dynamical systems, this method can generate reduced-order models (ROMs) that strictly represent the system behavior for given values of the parameters. Therefore, even if DMD may reduce the computational cost of dynamic systems' modeling, it is not suitable to explore the system response in the space of variability of parameters.

On the other hand, Polynomial Chaos Expansion (PCE) is a widely employed technique to address the effect of parameter variability on the model response but its use in case of unsteady processes (and models) is hindered by the simulations required to calibrate the PCE-based surrogate at each space-time location of interest.

In this chapter, we suggest the combination of DMD with PCE into a robust model-reduction framework suitable to address parameter variability issues when modeling systems with time-variant responses. Specifically, we propose a DMD-informed PCE, where DMD is used to interpolate high-fidelity data, thus contributing to the dataset required to calibrate the PCE. We investigate the accuracy of the DMD-informed PCE by testing the method against a scenario of a non-linear 2D multiphase flow in a heterogeneous media. We test the approach for different replacement rates and explore the use of the standard (sDMD) and the extended (xDMD) versions of the DMD method for the interpolation task.

The study proves the effectiveness of the novel framework, in particular when the PCE is informed by xDMD. As the xDMD-ROMs have lower and more stable interpolation errors, they guarantee better performances of the xDMD-informed PCE, which does not suffer from the same degradation visible for the sDMD-informed PCE when the replacement rate increases.

The chapter is organized as follows: Section 5.1 introduces the theme of uncertainty quantification and the modeling requirements of these analyses; Section 5.2 is devoted to the presentation of the proposed methodological framework; Section 5.3 describes the case study selected to exemplify our approach and the correspondent implementation of the methodological framework; in Section 5.4 we discuss the results focusing on the different development step of the framework; while Section 5.5 provides final remarks.

The study presented in this chapter is related to the paper G. Libero, D.M. Tartakovsky, V. Ciriello, *Dynamic Mode Decomposition enhances Polynomial Chaos for order reduction of dynamic processes* under review in Advances in Water Resources.

## 5.1 Model Reduction for Uncertainty Quantification

Our ability to model complex systems is usually limited by the uncertainty. For this reason, it is crucial for the development of robust and reliable models to explore the space of variability of the parameters and evaluate the effects on the system's response, such as in the case of uncertainty quantification or scenario development. A typical example is the interpretation of hydrological processes and, specifically, of flow and transport phenomena. These typically time-dependent processes (Benettin et al., 2022; Di Dato et al., 2018; Botter et al., 2010; Dentz et al., 2004, 2000; Tartakovsky and Neuman, 1998) are strongly influenced by parametric uncertainty and wide literature concerns uncertainty quantification in this field (Kang et al., 2022; Zhan et al., 2022; Ciriello and de Barros, 2020; Guo et al., 2020).

To obtain accurate predictions, HFMs are usually required. However, their prohibitive computational cost limits the application when many simulations are needed, such as in uncertainty quantification or scenario analysis (Ciriello et al., 2012, 2017). Indeed, despite rapid advances in software and hardware development, the growth in computational power is going hand in hand with that in model complexity so that our ability to perform high-fidelity physics-based simulations remains practically unchanged.

In this context, ROMs are efficient mathematical representations that capture the key underlying dynamics of a complex system directly from a small collection of HFM's output and/or observational data and reproduce at a low cost the system behavior (Kang et al., 2022; Zhan et al., 2022; He and Tartakovsky, 2021; Tartakovsky et al., 2020; Ciriello et al., 2019; Oladyshkin et al., 2012). The basic idea behind the use of suitable model reduction techniques for uncertainty quantification is to provide a simple mathematical relationship to relate the key governing input features and the model response in output. Thus, providing a viable tool to perform a large number of simulations in the parameter space while preserving the high-fidelity representation of the phenomena (Focaccia et al., 2021; Ciriello et al., 2017, 2012; Oladyshkin et al., 2012).

DMD is a promising data-driven technique to generate ROMs of complex dynamical systems (Schmid et al., 2010; Tu et al., 2014). However, this approach, based on the identification of spatiotemporal structures that are dominant in the data through singular value decomposition (Kutz et al., 2016), is only able to reconstruct the underlying processes for a specific point in the parameter space. In other terms, a DMD-based ROM strictly represents the system described by its training data and cannot be generalized to differently parameterized situations. Therefore, even if DMD is a powerful tool to reduce the computational cost of dynamic systems' modeling, it is not suitable to explore the system response in the space of variability of parameters.

A suitable and widely employed technique to address parameters variability issues is the PCE. This meta-modeling approach can accelerate uncertainty quantification and other onerous analyses otherwise computationally prohibitive if performed directly on an HFM, as demonstrated by several literature examples concerning flow and transport problems (Ciriello and de Barros, 2020; Oladyshkin et al., 2012). The method provides an approximation of the response computed with the HFM in the parameter space in the form of a series expansion truncated to a finite order (Sudret, 2008; Ghanem and Spanos, 1991; Wiener, 1938). The result is a surrogate polynomial model made of a basis of multivariate polynomials (orthonormal with respect to the joint pdf of the parameters), multiplied by deterministic coefficients that embed the space-time dependence of the response (Sudret, 2008; Xiu and Karniadakis, 2002).

Its effectiveness, as for all ROMs, depends on the amount of high-fidelity simulations required for the calibration phase. As described in Section 2.3.1, this burden can be reduced by limiting the training of the surrogate's coefficients to a small number of optimal regression points (namely *collocation points*) provided by the Probabilistic Collocation Method (PCM). The PCM is employed to optimize the procedure reducing the HFM runs to some optimal combinations of values identified to span the entire space of variability of the model's parameters (Sudret, 2008; Webster et al., 1996). However, the number of regression points (i.e. HFM simulations needed to train the PCE) rapidly increases with the parameter space dimension, with the truncation order, and when the process is space-time dependent. In this case, we need to compute the PCE coefficients at each space-time location of interest (Ciriello and de Barros, 2020). As a result, the computational cost required by the calibration increases and hence efficiency drastically diminishes limiting the applicability of the method to problems involving complex dynamical systems.

Here, we suggest the combination of PCE and DMD in a robust model-reduction framework for models with time-variant response. The approach takes advantage of the capability of PCE to approximate the system's response in the space of variability of the parameters (thus enabling to perform stochastic analysis and scenario development) and the attitude of DMD to interpolate high-fidelity data in time. Specifically, we propose a DMD-informed PCE (hereinafter DMD-PCE) where DMD has the role of informing PCE, replacing a large number of the simulations required from the HFM with reconstructions generated at a lower computation cost.

## 5.2 Integrated Model Reduction Framework

To calibrate the surrogate polynomial model described by Equation 2.26, the regression given by Equation 2.25 has to be solved at each space-time location where we want to compute the PCE approximation. When we need to approximate a dynamical system, the number of space-time locations increases rapidly, together with the computational burden associated with the generation of the training high-fidelity data.

To reduce the computational cost of this step, which is the only

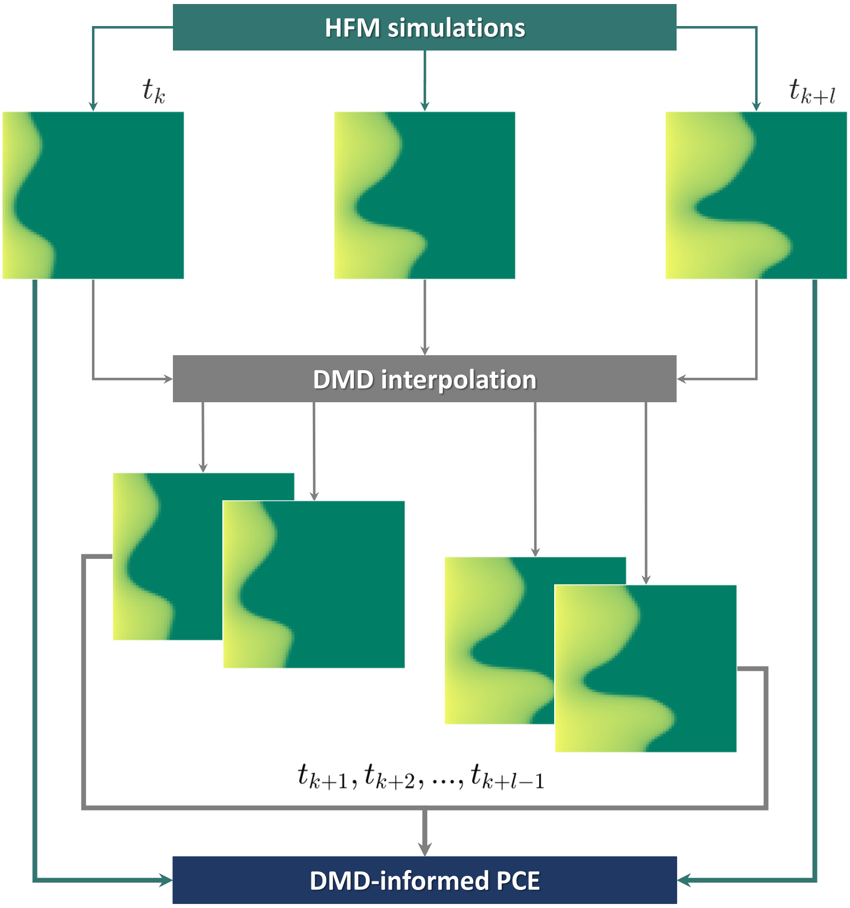

**Figure 5.1:** Methodological framework of the DMD-informed PCE approach.

relevant cost related to the PCE framework, we suggest solving the regression on a mixed dataset of high and low-fidelity data. High-fidelity snapshots are generated by the HFM only for the $(1-\eta)\%$ of the time instants, while DMD is employed to interpolate data for the others $\eta\%$ time steps (Figure 5.1). Note that if $\eta = 0$ the PCE coefficients are computed based only on the HFM, which corresponds to the traditional PCE approach.

The accuracy of the method depends on the accuracy of the DMD interpolation, which depends on the replacement rate $\eta\%$ in turn, and finally on the capabilities of the PCE technique. A first assessment of the performance consists of a comparison between the PCE coefficients calibrated on the mixed database, for a given $\eta$, and the coefficients trained on all high-fidelity data ($\eta = 0$). Secondly, we can directly compare the PCE approximation of the response for a given $\eta$, i.e. $\hat{\mathbf{x}}_\eta$, and the approximation in case of $\eta = 0$, i.e. $\hat{\mathbf{x}}_{\eta=0}$. Finally, we can compare both the approximations to the reference high-fidelity dataset, i.e. $\mathbf{x}$.

## 5.3 Data and Application

### 5.3.1 Physically-Based HFM

Modeling multi-phase flow is a challenging task that plays a notable role in many phenomena, ranging from contaminant migration to geothermal flow, and carbon sequestration (Mo et al., 2019). The need to solve strongly non-linear governing PDEs results in a high, often prohibitive, computational cost. This limits any detailed study requiring a large number of simulations, such as uncertainty quantification analysis (Song and Tartakovsky, 2021).

Here, we test our methodology on a 2D case study of multiphase flow in a heterogeneous media. Two incompressible and immiscible fluids, with viscosities $\mu_1$ and $\mu_2$, flow in a heterogeneous, incompressible, and isotropic porous medium, characterized by porosity $\phi$ and intrinsic permeability $\mathbf{k}(\mathbf{s})$ (Taverniers et al., 2020; Song and Tartakovsky, 2021). Combining the mass conservation equation

$$\phi\frac{\partial S_l}{\partial t} + \nabla \cdot \mathbf{v}_l + q_l = 0, \mathbf{s} \equiv (s_1; s_2)^\top \in D; t \in [0; T]; \tag{5.1}$$

and the generalized Darcy's law

$$\mathbf{v}_l = -\mathbf{k}\frac{\mathrm{k}_{rl}}{\mu_l}\nabla P_l \tag{5.2}$$

we can describe the propagation of each phase ($l$=1,2). Here, $S_l(\mathbf{s}, t)$ is the phase saturation, constrained by $S_1 + S_2 = 1$, while $q_l$ is the source/sink term. According to Equation 5.2, the macroscopic velocity $\mathbf{v}_l(\mathbf{s}, t)$ depends on the relative permeability of the phase, which in turn varies with the corresponding saturation, $\mathrm{k}_{rl} = \mathrm{k}_{rl}(S_l)$, in accordance with the Brooks-Corey constitutive model (Corey, 1954). The pressure is assumed to be equal within the two phases $P = P_1 = P_2 \equiv P(\mathbf{s}, t)$, hence the capillarity forces are ignored (Taverniers et al., 2020).

Similarly to Taverniers et al. (2020), we study the flow on a square simulation domain $D$ of size 150 m x 150 m, with impermeable bottom $\Gamma_b$ and top $\Gamma_t$ boundaries, and Dirichlet conditions imposed along the left $\Gamma_l$ and right $\Gamma_r$ boundaries:

$$\begin{gathered}\frac{\partial P}{\partial s_2} = 0,\ \mathbf{s} \in \Gamma_b \cup \Gamma_t; \quad P = 10,\ \mathbf{s} \in \Gamma_r; \\ P = 10 + \Delta p\ \wedge\ S_1 = 1.0,\ \mathbf{s} \in \Gamma_l.\end{gathered} \tag{5.3}$$

Initial conditions are:

$$P(\mathbf{s}, 0) = 10.1,\ S_1(\mathbf{s}, 0) = 0,\ \mathbf{s} \in D. \tag{5.4}$$

Pressure P is expressed in MPa.

The flow is simulated for two different intrinsic permeability scenarios (SC1, SC2), to account for the role played by heterogeneity. The two permeability maps are generated as second-order stationary random field, such that $Y(\mathbf{s}) = \ln \mathbf{k}$ is multivariate Gaussian with mean $\mu_Y(\mathbf{s}) = 0$, variance $\sigma_Y^2 = 2.0$ and correlation length $\lambda_Y = 19m$ in the first case and $\mu_Y(\mathbf{s}) = 0$, $\sigma_Y^2 = 3.0$ and $\lambda_Y = 10m$ in the second case. The field $Y(\mathbf{s})$ is approximated by means of a truncated Karhunen–Loève expansion, counting 31 terms, chosen to capture 95% of the energy of $Y$ (Taverniers et al., 2020). Panels (a) of Figure 5.2 and Figure 5.3 show the two permeability fields SC1 and SC2, respectively.

Equations 5.1 and 5.2 are discretized using a finite volume scheme in space and an implicit Euler scheme in time. The resulting highly non-linear system of equations is solved iteratively at each time step using the Newton-Raphson method, with modified Appleyard saturation update damping (Appleyard et al., 1981) to improve convergence. To ensure convergence of both flow ad transport solutions three convergence criteria are specified: normalized residual norm, maximum saturation update, and maximum relative pressure update, with respective tolerances set to $\varepsilon_1 = 10^{-6}$, $\varepsilon_2 = 10^{-2}$, and $\varepsilon_3 = 10^{-3}$ (Taverniers et al., 2020; Song and Tartakovsky, 2021).

### 5.3.2 Application to the Case Study

Among all the input parameters of the HFM (Taverniers et al., 2020), we focus on possible uncertainty related to the porosity $\phi$, and the external pressure difference at the boundaries $\Delta p$. We assume the two input variables to be independent random variables, described by two normal distributions: $\phi \simeq \mathcal{N}(m_\phi = 0.25; \sigma_\phi = 0.05)$, $\Delta p \simeq \mathcal{N}(m_{\Delta p} = 0.20\text{MPa}; \sigma_{\Delta p} = 0.04\text{MPa})$. As a consequence, the Hermite Polynomials should be used to build the PCE surrogate polynomial model in Equation 2.24 (Xiu and Karniadakis, 2002). We choose a second-order PCE approximation, i.e. $q = 2$, and then $P = 6$. Under these conditions, the PCM identifies $N_{CP} = 6$ optimal collocation points in the random parameter space, i.e. six combinations of values for the two uncertain inputs, $\phi$ and $\Delta p$.

For each collocation point, we perform high-fidelity simulations for both the permeability scenarios SC1 and SC2. Each high-

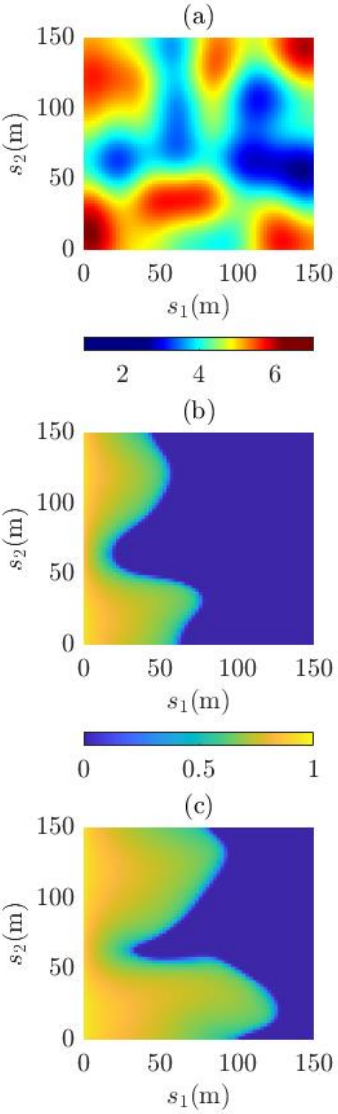

**Figure 5.2:** (a) the log-perm field $Y(\mathbf{s}) = \ln(\mathbf{k})$ (with $\mathbf{k}$ expressed in mDarcy) for the first scenario (SC1) and the corresponding saturation maps simulated by the HFM at time steps (b) $k = 910$ and (c) $k = 1820$.

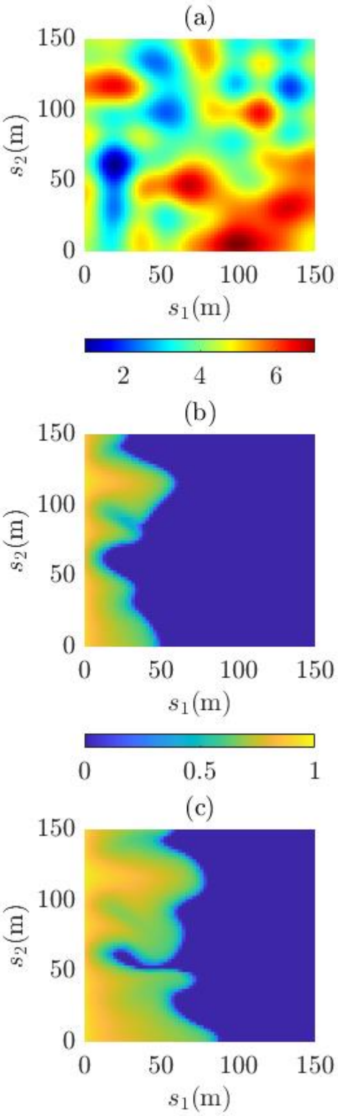

**Figure 5.3:** (a) the log-perm field $Y(\mathbf{s}) = \ln(\mathbf{k})$ (with $\mathbf{k}$ expressed in mDarcy) for the second scenario (SC2) and the corresponding saturation maps simulated by the HFM at time steps (b) $k = 910$ and (c) $k = 1820$.

fidelity simulation involves the computation of the saturation of the first fluid, $S_1$, on a 64x64 grid for 1828 time instants, from $t = 0$ to $t = 1827$ days, with a time step $\Delta t = 1$ day. The first 8 snapshots of each series are not considered for the analysis, because the initial transient behavior of the phenomena could affect their accuracy, hence, the snapshots available for the training of DMD-based ROMs are 1820. Panels (b) and (c) in Figure 5.2 and Figure 5.3 show the saturation maps simulated by the HFM for $k = 910$ ($t = 917$ days) and $k = 1820$ ($t = 1827$ days), for the two permeability scenarios SC1 and SC2. Vectorizing each snapshot, we obtain, for each combination provided by the PCM, a matrix counting $n = 64 * 64 = 4096$ rows and 1820 columns.

When high-fidelity simulations are performed for all six collocation points, we can use them to calibrate the polynomial surrogate model at all the space-time locations of interest. The result is a PCE completely trained on high-fidelity data (hereinafter this will be indicated as HF-PCE). Alternatively, for each collocation point, we propose to use a limited number of high-fidelity simulations to calibrate a DMD-based ROM able to approximate the evolution of the response in time, and then use it to interpolate and fill the missing high-fidelity data. The DMD-informed PCE (hereinafter also indicated as DMD-PCE) is then compared to the HF-PCE in terms of accuracy.

In this study, different replacement rates are considered, specifically: $\eta = (0.5, 0.75, 0.9)$. For each value of $\eta$, a different DMD approximation is computed testing both sDMD and xDMD and considering the three different truncation criteria described in Section 2.2.4 $r$=($\mathtt{rank}(\mathbf{X}), r^*, r_{90}$). Note that $r_{90}$ corresponds to the minimum number of singular values accounting for 90% of the cumulative energy in the SVD of $\mathbf{X}$; $r^*$ is identified setting $\zeta = 10^{-5}$ in Equation 2.20; while $r = \mathtt{rank}(\mathbf{X})$ is computed with the dedicated `MATLAB` function. A summary of the cases considered in this analysis is reported in Table 5.1. The framework of the proposed new approach and its application to the case study is described in Figure 5.4.

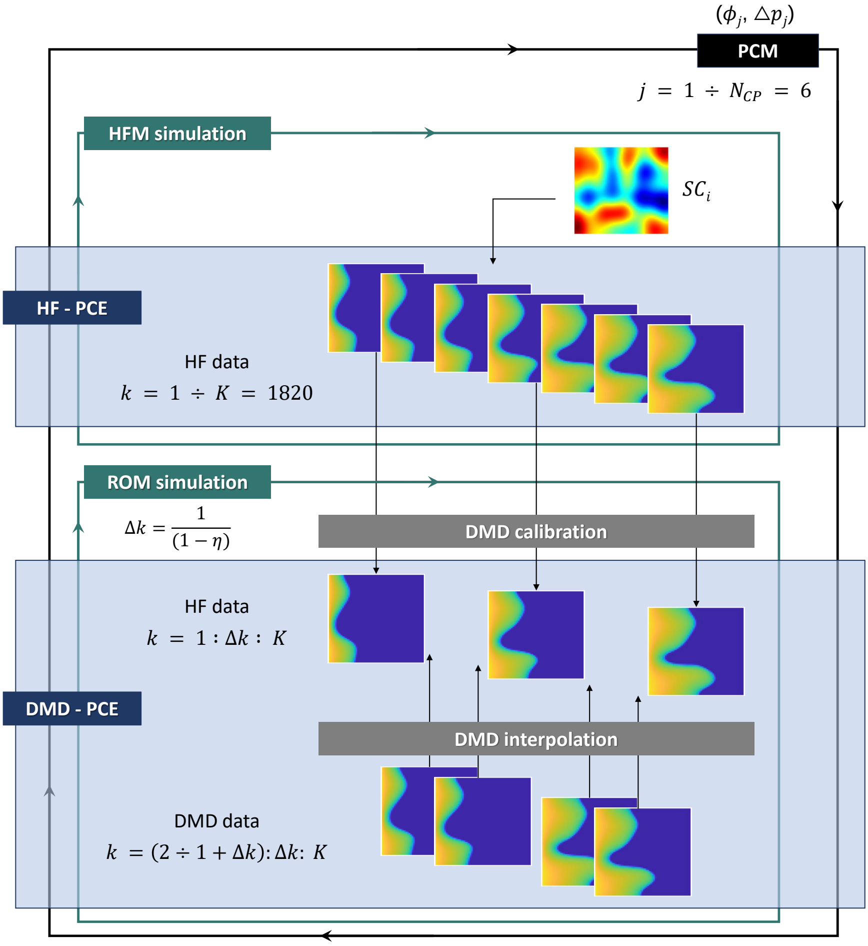

**Figure 5.4:** Methodological framework of the DMD-informed PCE approach applied to the case study. Given a permeability scenario $SC_i$, for each collocation point identified by the PCM we perform an HFM simulation of 1820 snapshots. We employ $(1-\eta)\%$ of the high-fidelity snapshots to train DMD and use the resulting ROM to interpolate the remaining time steps. We repeat the procedure for $N_{CP} = 6$ collocations points. All 6 series of high-fidelity snapshots are employed to train the HF-PCE; while the mixed datasets with DMD-interpolated snapshots are sent to the DMD-PCE.

**Table 5.1:** Datasets and truncation ranks corresponding to the three different replacement rates $\eta$, for the two permeability scenarios (SC1, SC2).

| | SC | Train (representation) | Test (interpolation) | $r_{90}$ | $r^*$ | $\texttt{rank}(\mathbf{X})$ |
|---|---|---|---|---|---|---|
| Case 1: | (1) | $k = 1 : 2 : 1820$ | $k = 2 : 2 : 1820$ | 11-16 | 75-129 | 890-909 |
| $\eta = 0.5$ | (2) | $k = 1 : 2 : 1820$ | $k = 2 : 2 : 1820$ | 8-13 | 61-105 | 765-908 |
| Case 2: | (1) | $k = 1 : 4 : 1820$ | $k = (2 : 4) : 4 : 1820$ | 10-16 | 75-129 | 454 |
| $\eta = 0.75$ | (2) | $k = 1 : 4 : 1820$ | $k = (2 : 4) : 4 : 1820$ | 8-13 | 61-104 | 454 |
| Case 3: | (1) | $k = 1 : 10 : 1820$ | $k = (2 : 10) : 10 : 1820$ | 10-16 | 74-128 | 181 |
| $\eta = 0.9$ | (2) | $k = 1 : 4 : 1820$ | $k = (2 : 4) : 4 : 1820$ | 8-13 | 60-104 | 181 |

## 5.4 Results and Discussion

### 5.4.1 DMD Calibration

We start by assessing the accuracy of both the standard (sDMD) and the extended (xDMD) algorithm in reproducing the same set of high-fidelity data employed for their calibration (hereinafter representation error). We test the performances for different dimensions of the training set (see Table 5.1).

Specifically, for each of the two permeability scenarios and the six collocation points, 3 different subsets, counting 50%, 25%, and 10% of high-fidelity data provided by the numerical model described in Section 5.3.1, are used for training. Based on these subsets, we derive DMD-based ROMs for 3 different truncation criteria to explore the accuracy of the two algorithms (see Table 5.1).

Figure 5.5 shows the results for the permeability scenario SC1, while Figure 5.6 refers to SC2. In both the Figures, columns correspond to Cases 1-3 in Table 5.1 (i.e. different dimensions of the training set), while the 6 different lines in each plot correspond to the 6 different combinations of values (i.e. collocation points) provided by the PCM to span the entire space of variability of the parameters. The three plots on the first row represent the singular values associated with the Singular Value Decomposition (SVD) of the matrix $\mathbf{X}$ of each of the 6 combinations for the three cases. The representation error, averaged over all the time instants, is plotted as a function of the truncation rank $r$ of the SVD in Panels (d)-(f) for the sDMD-based ROMs and in Panels (g)-(i) for the xDMD-based ROMs. The ranks corresponding to different truncation criteria are also reported: pointed in red is $r = \texttt{rank}(\mathbf{X})$, in green $r = r^*$, and in blue $r = r_{90}$.

We note that the range of values associated with $r_{90}$ is sensibly the same for the three cases, for both SC1 and SC2. The same holds for $r^*$, even if the range of values associated with the 6 realizations is slightly wider (see also Table 5.1). On the other hand, $\texttt{rank}(\mathbf{X})$ changes from case to case, being equal (Cases 2-3) or very close (Case 1) to the number of training snapshots. In addition, it can be noticed that the majority of the information is carried by a limited number of singular values which is sensibly captured by $r^*$ and does not change increasing the dimension of the training set. This is confirmed by the plots on the second and third row: the representation error becomes unstable after $r^*$, especially for SC2, suggesting all the exceeding features (correspondent to $r > r^*$) can be interpreted as noise or negligible details. For this reason, in all the following computing steps, we truncate the DMD models at $r = r^*$. Finally, comparing the performance of the sDMD and xDMD algorithms in Figures 5.5 and 5.6, we can verify that xDMD returns a better approximation in both the permeability scenarios.

### 5.4.2 DMD Interpolation

We consider the sDMD and xDMD models calibrated in the previous Section on 50%, 25%, and 10% of the entire set of high-fidelity data, and truncated at $r = r^*$, and we employ them to interpolate the model response. We reconstruct the saturation maps for the missing time steps, i.e. thus corresponding to a replacement rate equal to $\eta = 0.5, 0.75, 0.9$, respectively (see Table 5.1 Cases 1-3).

Interpolation performance is explored in Figure 5.7 and Figure 5.8 for permeability scenarios SC1 and SC2, respectively. In both the Figures, columns correspond to Cases 1-3 in Table 5.1,

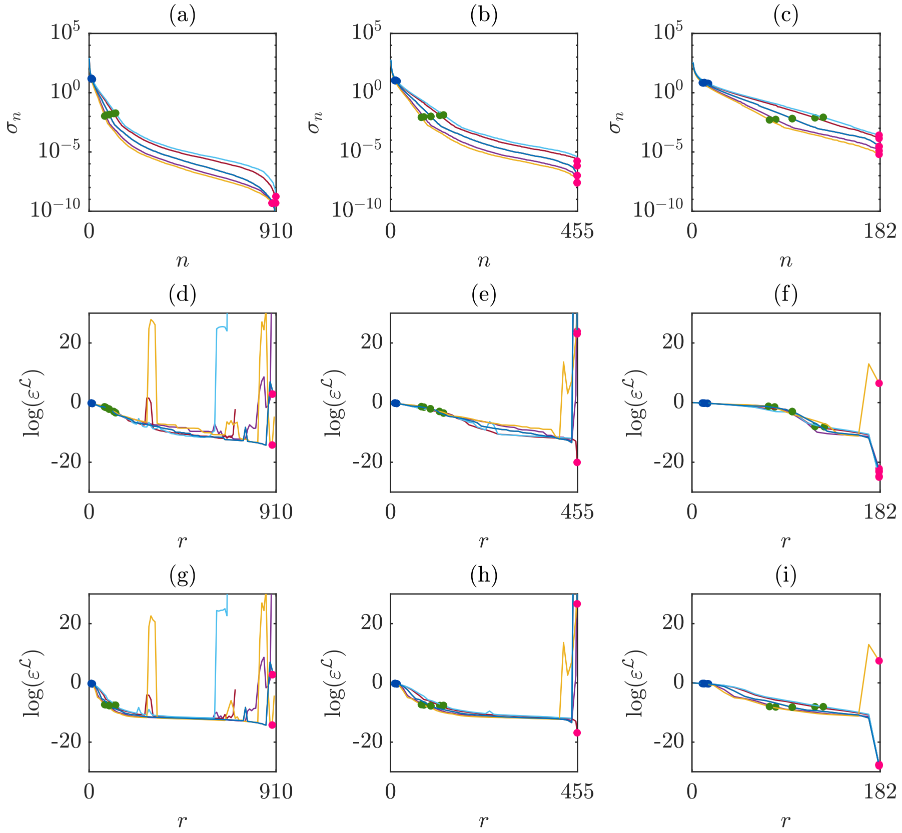

**Figure 5.5:** For the first permeability scenario (SC1): (a)-(c) represent the singular values associated with the SVD of $\mathbf{X}$ for Cases 1-3 in Table 5.1, respectively. For the same cases (d)-(f) and (g)-(h) show the representation error (averaged over the time instants) of the sDMD and xDMD as a function of the truncation rank $r$ of the SVD. Different lines in each panel correspond to the simulations for the 6 collocation points. In each panel, dots correspond to `rank`$(\mathbf{X})$ (in red), $r^*$ (in green), $r_{90}$ (in blue) (the numerical values are in Table 5.1).

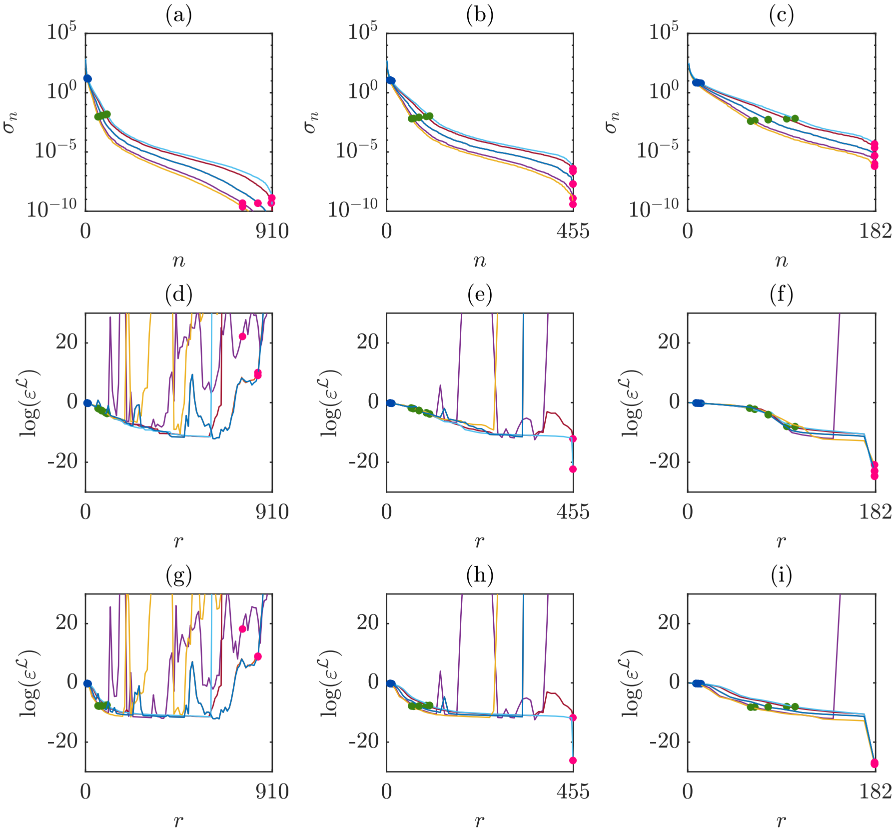

**Figure 5.6:** For the second permeability scenario (SC2): (a)-(c) represent the singular values associated with the SVD of $\mathbf{X}$ for Cases 1-3 in Table 5.1, respectively. For the same cases (d)-(f) and (g)-(h) show the representation error (averaged over the time instants) of the sDMD and xDMD as a function of the truncation rank $r$ of the SVD. Different lines in each panel correspond to the simulations for the 6 collocation points. In each panel, dots correspond to $\texttt{rank}(\mathbf{X})$ (in red), $r^*$ (in green), $r_{90}$ (in blue) (the numerical values are in Table 5.1).

while the 6 differently colored lines in each panel correspond to the 6 collocation points. Specifically, Panels (a)-(c) show the interpolation error, computed through Equation 2.23, of the sDMD, while the results for xDMD are shown in Panels (d)-(f). Focusing on the first row (Panels (a)-(c)), we observe that the logarithm of the error provided by the sDMD is in the range $(-5, -1)$ for all the cases, but an increasing instability affects the performance when the dimension of the training set decreases. Conversely, for xDMD (Panels (d)-(f)) the logarithm of error is sensibly lower, within $(-9, -3)$, and the variability between the 6 series is less remarkable. The error decreases with time for all the cases, and instabilities are smaller than for sDMD, denoting higher robustness in the behavior of the xDMD algorithm. This holds for both SC1 and SC2. On the third row of both the Figures (Panels (g)-(i)), we focus on the difference between the high-fidelity saturation map and the correspondent map interpolated by xDMD at $k = m - 1$. The difference is maximum for Case 3, though in the range $(-5 \cdot 10^{-5}, 5 \cdot 10^{-5})$.

The interpolation errors discussed in this Section have a great influence on the accuracy of the overall method, as they propagate in the DMD-PCE computation as discussed below.

### 5.4.3 DMD-Informed PCE Accuracy

At this point, for each case, new databases can be formed merging the snapshots interpolated through sDMD or xDMD together with the high-fidelity data employed for the training. In Case 1 half of the dataset is made of high-fidelity data and the other half is simulated by DMD; in Case 2 75% of data is provided by the DMD surrogate; while in Case 3 the surrogate provides 90% of the entire set. As the replacement rate increases, the computational cost to produce the dataset decays.

PCE surrogate models are built according to Equation 2.26 and trained on these new datasets. Calibration of the PCE surrogate consists of the computation of a set of 6 coefficients $\mathrm{a}_j(\mathbf{s}, t)$

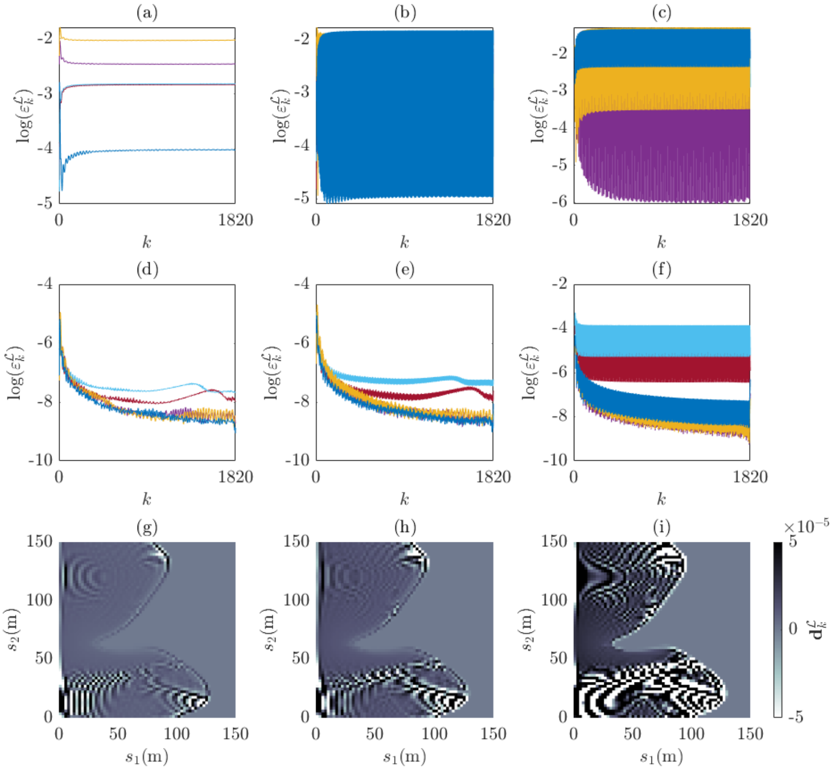

**Figure 5.7:** For the first permeability scenario (SC1): (a)-(c) represent the sDMD interpolation error in time for Cases 1-3 in Table 5.1; (d)-(f) show the same for xDMD; (g)-(h) show the difference between real data and the corresponding snapshots interpolated by xDMD at $k = 1819$ ($t = 1826$) for Cases 1-3.

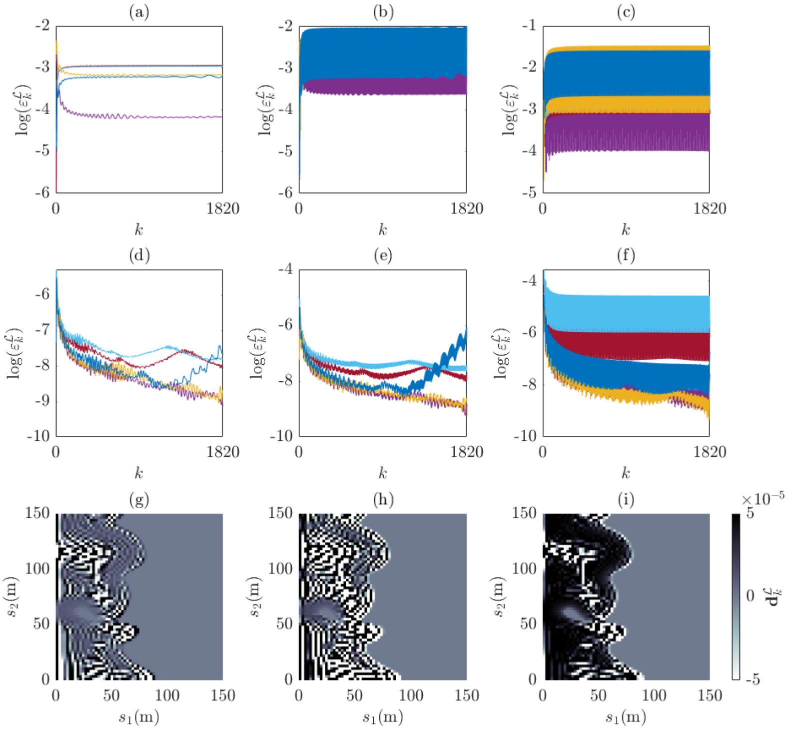

**Figure 5.8:** For the second permeability scenario (SC2): (a)-(c) represent the sDMD interpolation error in time for Cases 1-3 in Table 5.1; (d)-(f) show the same for xDMD; (g)-(h) show the difference between real data and the corresponding snapshots interpolated by xDMD at $k = 1819$ ($t = 1826$) for Cases 1-3.

for each point in space and time. We repeat this calibration for the three different replacement rates and both the permeability scenarios. The coefficients $\mathrm{a}_j^{\mathcal{L}}(\mathbf{s}, t)$ resulting from the calibration of the PCE polynomial structure on the mixed datasets (DMD-PCE) are compared to the coefficients $\mathrm{a}_j(\mathbf{s}, t)$ calibrated entirely on high-fidelity data (HF-PCE). The error between the two approaches is computed for each coefficient as $\varepsilon_k^{\mathrm{a}_j} = \frac{\|\mathrm{a}_j^{\mathcal{L}}(\mathbf{s},k) - \mathrm{a}_j(\mathbf{s},k)\|^2}{\|\mathrm{a}_j(\mathbf{s},k)\|^2}$, where $\mathrm{a}_j^{\mathcal{L}}(\mathbf{s}, k)$ is the vector of values of the coefficient $\mathrm{a}_j$ provided by the DMD-PCE at a specific time instant $k$ and for all the space locations collected in the vector $\mathbf{s}$ and $\mathrm{a}_j(\mathbf{s}, k)$ is the corresponding vector of coefficients from HF-PCE.

Boxplots in Figure 5.9 describe the distribution of the error when replacing $\eta = 0.5, 0.75, 0.9$ of high-fidelity data computed for SC1, with sDMD interpolated snapshots. Figure 5.10, instead, refers to the case where interpolation is performed through xDMD. Figure 5.11 and Figure 5.12 show the same for SC2.

As expected, the growth of the replacement rate always corresponds to an error increase. Furthermore, even if the magnitude of the error varies from one coefficient to the other, it can be noticed that xDMD is confirmed to always guarantee better performance than sDMD. Also, the differences in SC1 and SC2 have a very small impact on the coefficients' error, which varies within similar ranges.

The accuracy of the DMD-informed PCE approach is finally tested comparing the approximation of the response given by the PCE-DMD surrogates with the results given by the traditional PCE model entirely trained on high-fidelity data (HF-PCE).

In the regression plots of Figure 5.13 and Figure 5.14 results from PCE surrogates informed by sDMD and xDMD are compared with the reference prediction of the HF-PCE generated for the permeability scenario SC1. In both Figures, Panel (a) refers to replacement rate $\eta = 0.5$, Panel (b) to $\eta = 0.75$, and Panel (c) to $\eta = 0.9$. Figure 5.15 and Figure 5.16 show the same for the second permeability scenario SC2. Where sDMD is employed to inform the PCE, the accordance with the reference case declines as the

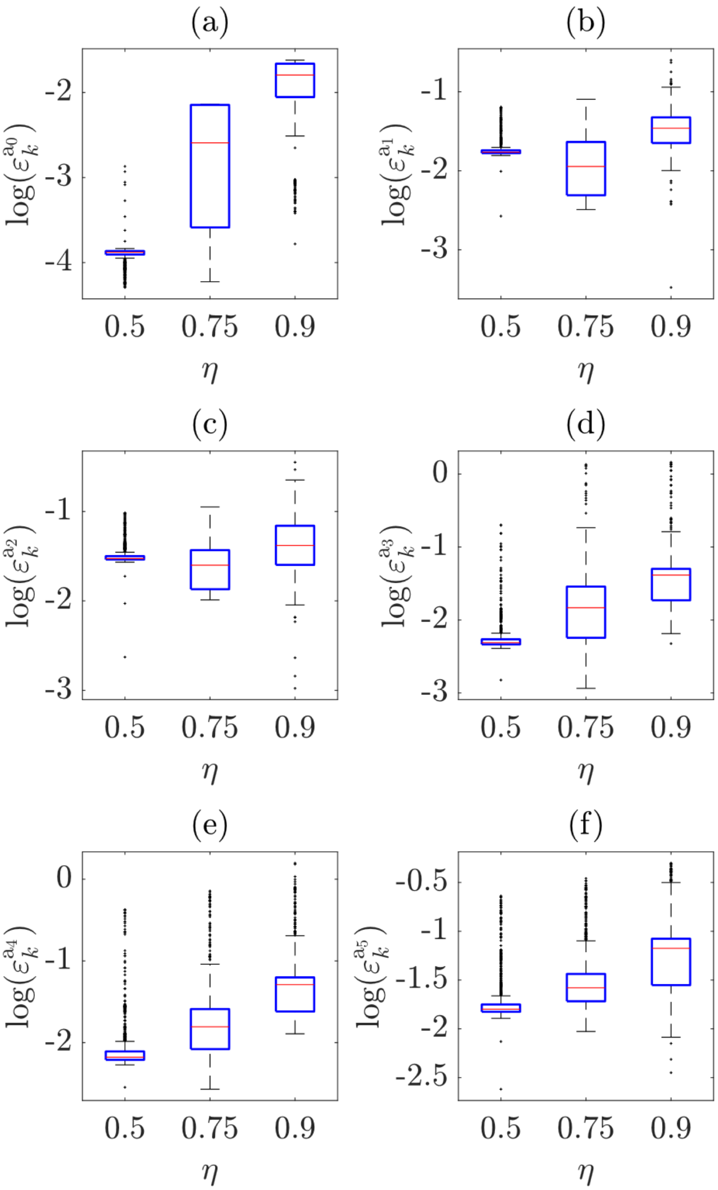

**Figure 5.9:** For the first permeability scenario (SC1): error in the estimate of PCE coefficients $\mathrm{a}_j$ when using sDMD interpolation in place of the HFM data for $\eta = 0.5, 0.75, 0.9$.

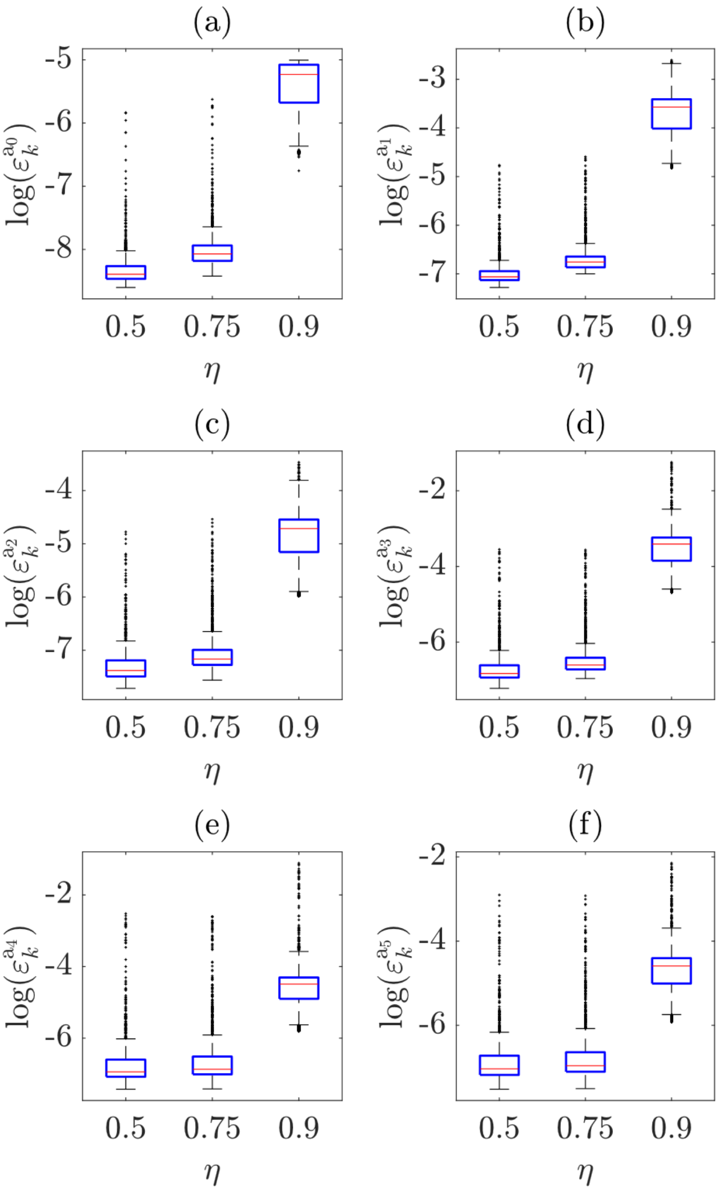

**Figure 5.10:** For the first permeability scenario (SC1): error in the estimate of PCE coefficients $a_j$ when using xDMD interpolation in place of the HFM data for $\eta = 0.5, 0.75, 0.9$.

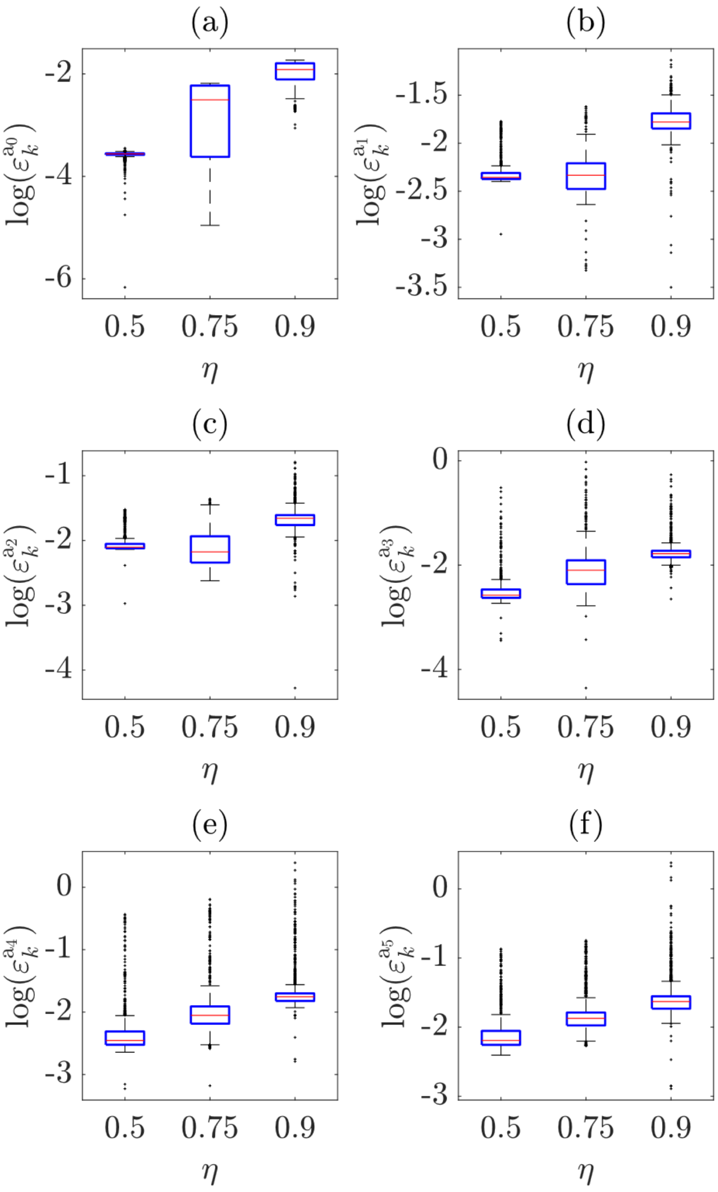

**Figure 5.11:** For the second permeability scenario (SC2): error in the estimate of PCE coefficients $\mathrm{a}_j$ when using sDMD interpolation in place of the HFM data for $\eta = 0.5, 0.75, 0.9$.

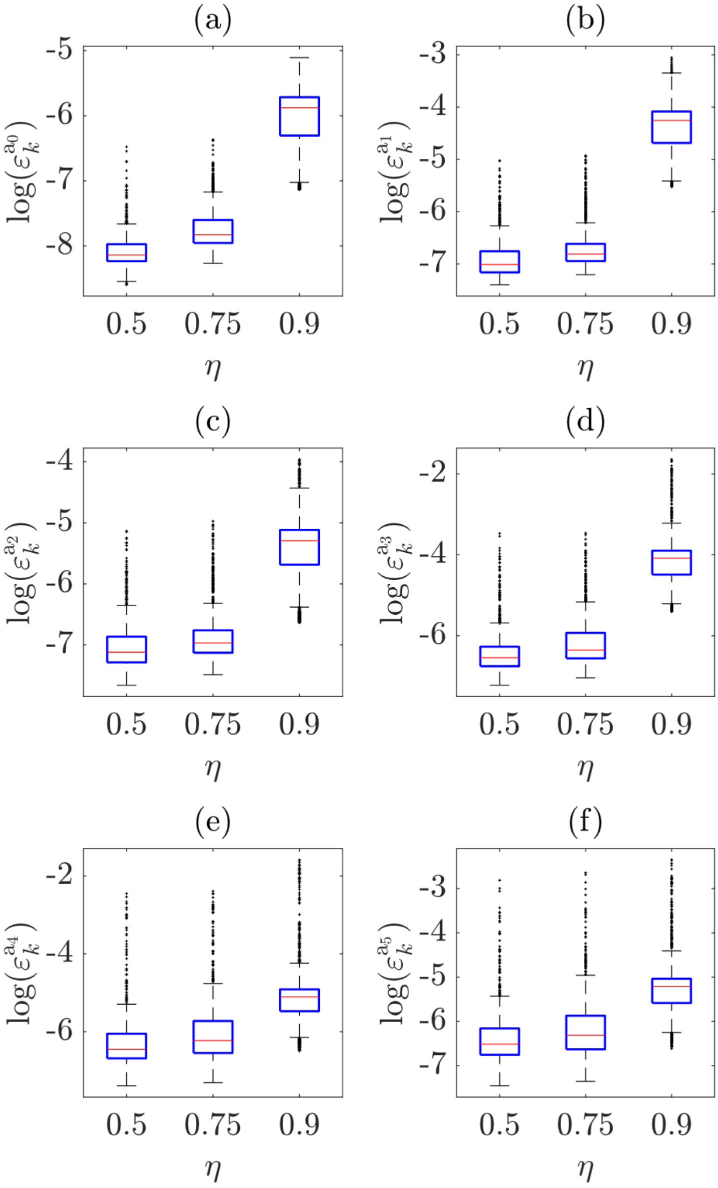

**Figure 5.12:** For the second permeability scenario (SC2): error in the estimate of PCE coefficients $\mathrm{a}_j$ when using xDMD interpolation in place of the HFM data for $\eta = 0.5, 0.75, 0.9$.

replacement rate increases. High accuracy is maintained, instead, for all the tested replacement rates when xDMD is employed.

## 5.5 Final Remarks

A new framework for robust reduction of models with time-variant response is obtained from the combination of DMD and PCE, and its accuracy is tested against a scenario of a non-linear 2D multiphase flow in a heterogeneous media. In this novel approach, the DMD method is employed to replace part of the high-fidelity simulations required to calibrate the PCE with reconstructions generated at a lower computation cost. The ability of both sDMD and xDMD to reproduce and interpolate high-fidelity data is explored for different replacement rates, and xDMD is proved to better approximate the system's states. As in the applications discussed in the previous chapters, we also investigated the truncation rank selection and identified $r = r^*$ as the optimal truncation criteria. The resulting sDMD and xDMD-based ROMs are then employed to generate, at a low computational cost, part of the simulations required for the PCE calibration. A first assessment of the accuracy of the approach is performed by comparing the PCE coefficients calibrated on the mixed database (DMD-PCE) and the coefficients trained on all high-fidelity data (HF-PCE). The comparison between the approximations of the system's response provided by the DMD-PCE and the HF-PCE further demonstrates the effectiveness of the novel framework. In particular, we prove that when the PCE is informed by xDMD, the lower and more stable interpolation errors guarantee better and more robust performances of the method. The demonstrated computational benefit and accuracy of the proposed method support the diffusion of this framework to ease onerous studies such as uncertainty quantification analysis.

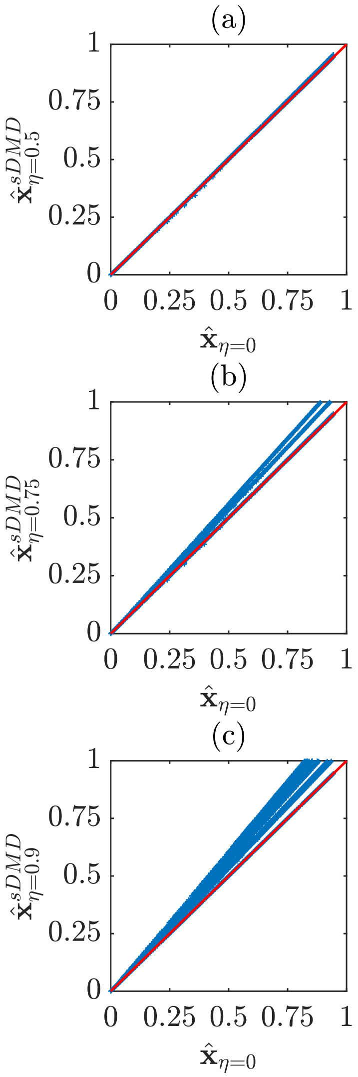

**Figure 5.13:** For the first permeability scenario (SC1): (a)-(c) show the sDMD-informed PCE prediction associated with $\eta = 0.5, 0.75, 0.9$ respectively vs the PCE prediction obtained for $\eta = 0$.

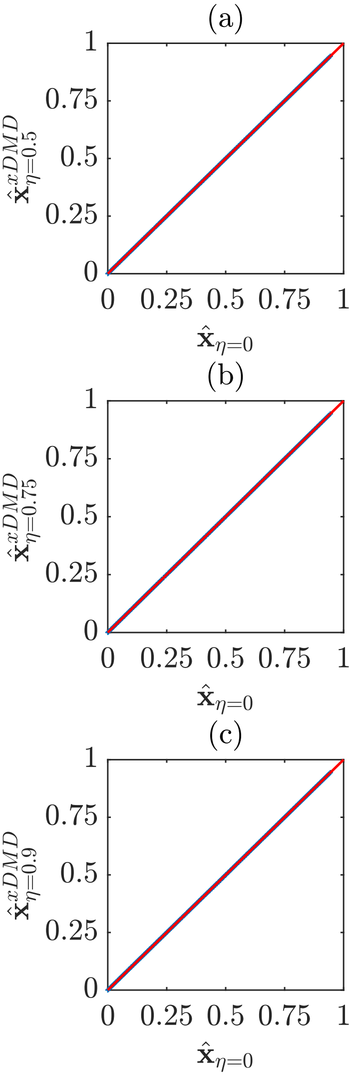

**Figure 5.14:** For the first permeability scenario (SC1): (a)-(c) show the xDMD-informed PCE prediction associated with $\eta = 0.5, 0.75, 0.9$ respectively vs the PCE prediction obtained for $\eta = 0$.

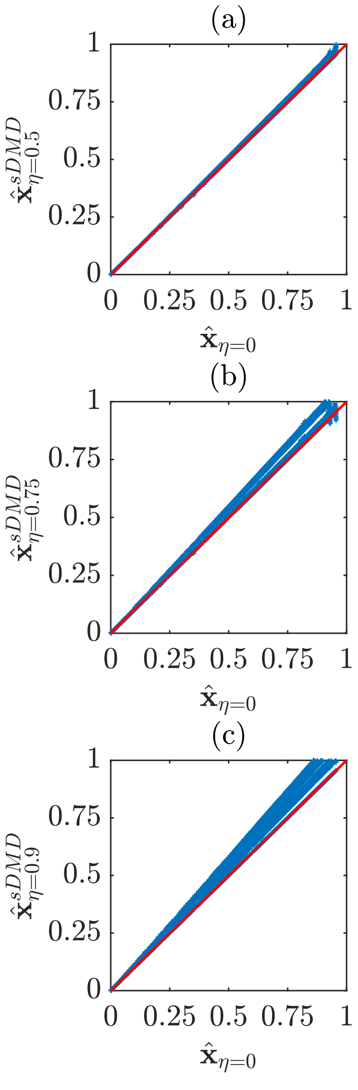

**Figure 5.15:** For the second permeability scenario (SC2): (a)-(c) show the sDMD-informed PCE prediction associated with $\eta = 0.5, 0.75, 0.9$ respectively vs the PCE prediction obtained for $\eta = 0$.

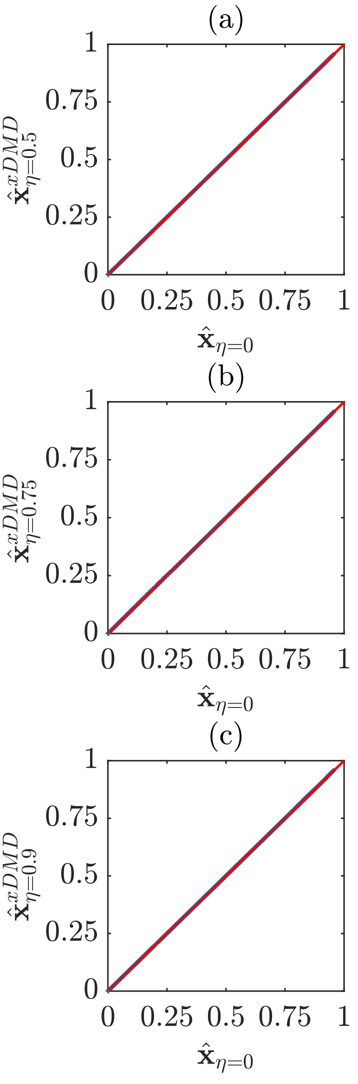

**Figure 5.16:** For the second permeability scenario (SC2): (a)-(c) show the xDMD-informed PCE prediction associated with $\eta = 0.5, 0.75, 0.9$ respectively vs the PCE prediction obtained for $\eta = 0$.

# 6 | Conclusions

In this thesis, the opportunities offered by the application of Dynamic Mode Decomposition (DMD) in the analysis of complex dynamic real-world systems are explored. Specifically, the capabilities of the DMD method are tested and investigated with respect to complex flow phenomena decoding.

The potential of this technique resides in its flexibility. As a purely data-driven and equation-free method, it can be equally employed in diverse fields, ranging from environmental science to medicine. The only requirement is the availability of a training set of snapshots of the system's evolution in time. As demonstrated by the applications proposed in this work, this dataset can collect either direct observations of reality or simulations performed through a physics-based model.

The method can derive from high-fidelity data computationally efficient reduced-order models (ROMs) with a twofold scope: reconstruct available snapshots, and interpolate or extrapolate the system state where data are missing. The ability to reconstruct high-fidelity data at a low computational rate justifies the successful use of DMD as a technique for data compression. On the other hand, the possibility of using the reduced-order model to predict the state of the system where no high-fidelity data are available can drastically reduce the computational burden of complex dynamical system analysis.

A key role in the accomplishment of both these tasks is played by the selection of an appropriate truncation. Indeed, in all the applications reported in this thesis, the truncation rank selection proved crucial to guarantee a suitable level of efficiency and accuracy in representation, interpolation, and extrapolation regimes.

One of the limits of the DMD method resides in the non-generalizability of the resulting ROMs, which are strictly valid for the system described by the training data. Hence, DMD by itself is not suitable for exploring the system response in the space of variability of parameters. However, it can be used to alleviate the computational cost related to uncertainty quantification analysis or other onerous studies performed with methods that require many simulations, especially when dealing with complex high-dimensional systems. In this context, an integrated framework for the reduction of models with time-variant response, which combines DMD with the Polynomial Chaos Expansion (PCE) technique, is proposed and tested in this thesis. In the DMD-informed PCE, part of the high-fidelity simulations required to calibrate the polynomial model is replaced by low-cost DMD reconstructions, reducing the overall computational demand of the method.

The applications presented in this thesis demonstrate that the DMD method if optimally implemented and eventually combined with PCE, can represent a powerful tool to ease diagnosis and scenario analysis when complex flow processes are involved.

The DMD-based approximations can replace the existing onerous high-fidelity models, allowing for faster but meaningful and robust predictions of flow and transport phenomena in very diverse fields. The possibility of performing reliable and cheap scenario analysis could positively impact the remediation of contaminated sites or the extraction of underground resources; while in the medical field, this translates to fast diagnosis of diseases or dysfunctions and efficient comparison of different options for intervention.

Even greater advantages can derive from the use of DMD when traditional physics-based models are not available, and the system evolution is only described by time-series of direct observations. In this case, the DMD-based models not only provide a valid approximation of the system behavior but also contain in their structure physically meaningful dynamics characterizing systems and processes. A typical example is the interpretation of satellite datasets: DMD could represent a key tool to decode these high-dimensional archives and take advantage of the embedded information to deeply understand global phenomena related to climate or the distribution of resources.

Future research could further investigate this aspect and prove that physical interpretability is another key feature in the characterization of DMD as a powerful tool to support and speed up humans' response to systems evolution.

# A Appendix: PCE to Assess Cardiovascular Function

Motivated by the efficiency of the PCE technique and the importance of diagnosis in the cardiovascular field, we tested the use of the PCE-based surrogate models to provide metrics based on blood flow for the assessment of the left ventricular function (Collia et al., 2022). The objective of the study is to explore the existence of relationships and dependencies between the clinical metrics commonly used to assess cardiac function and advanced flow-related quantities.

To this aim, we combine high-fidelity numerical simulations of blood flow in the left ventricle with PCE surrogates. The HFM is given by a direct numerical simulation method of intraventricular fluid dynamics that allows simulation and analysis of cardiac flow in individual patients. Numerical simulations are based on the immersed boundary method, often employed to analyze fluid dynamics in the left ventricle. Surrogate models are calibrated on the high-fidelity simulations to approximate four QoIs describing blood transit in the left ventricle: (i) the direct volume $V_d$, which is the volume of blood that enters the ventricle and is ejected during the same heartbeat; (ii) the residual volume $V_r$, which is the amount of blood that is present in the ventricle at the beginning of diastole and is not ejected during systole; (iii) the ventricular kinetic energy $KE_{Epeak}$; (iv) the vorticity $\overline{\omega}_{Epeak}$. The approximations of these quantities are derived in the space of variability of four input parameters. The selected parameters are the ejection fraction ($EF$); the ratio between the early (E-wave) and atrial

(A-wave) peaks of the mitral inflow velocity ($E/A$); the left ventricle shape ratio ($LV_r$); and the ratio between the diameter of the mitral valve annulus and that of the entire left ventricle valvular plane ($MVLV_r$). These parameters, selected among the standard diagnostic clinical metrics, embed relevant information on the left ventricle's geometry and function and help discriminate between healthy and diseased patients. Optimal combinations of values of these four input parameters are given by the PCM and applied to a parameterized geometry on which the HFM is run. The resulting high-fidelity dataset is employed for calibration and validation of the surrogate models (Figure A.1).

Once the PCE surrogates are available, we compare their predictions to the results of the HFM for several patient-specific cases under both healthy and diseased conditions to assess the robustness of the method (Figure A.2).

The surrogates also allow us to perform global sensitivity analysis at a negligible computational cost. The sensitivity indices of Sobol are computed as analytical post-processing of the PCE coefficients for each governing parameter and employed to assess the relative impact of input variability on the quantities of interest (Figure A.3).

This study demonstrates that the application of the PCE technique can provide us with surrogates that are directly employable to speed up clinical analysis, but also allow us to uncover relationships between common global flow transit quantities and standard clinical parameters. Such surrogate models require an initial computational effort to generate the high-fidelity simulations for the calibration, then they present a notable simplicity and can be employed in real-time during the clinical analysis. Merging surrogate models and HFMs, hence, results in a promising winning approach to describe important metrics behavior in the cardiology field.

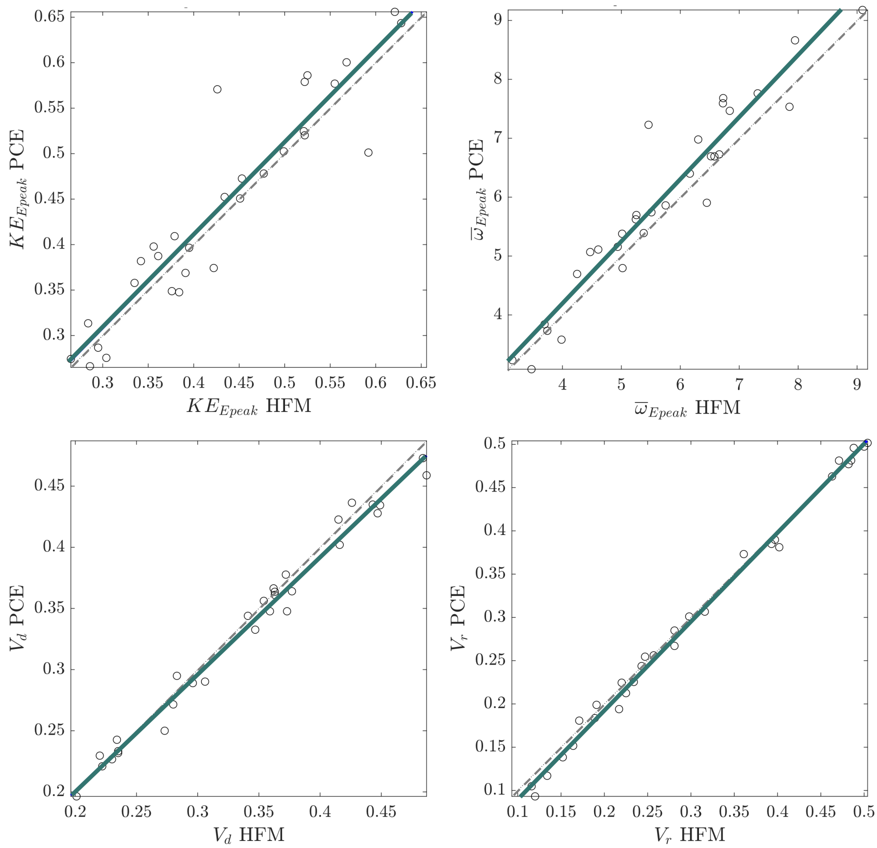

**Figure A.1:** Comparison between HFM and PCE predictions for the 4 quantities of interest ($KE_{Epeak}$, $\overline{\omega}_{Epeak}$, $V_d$, $V_r$) on a set of 30 synthetic cases. Solid green lines represent the linear regressions, while dashed grey lines are the bisectors.

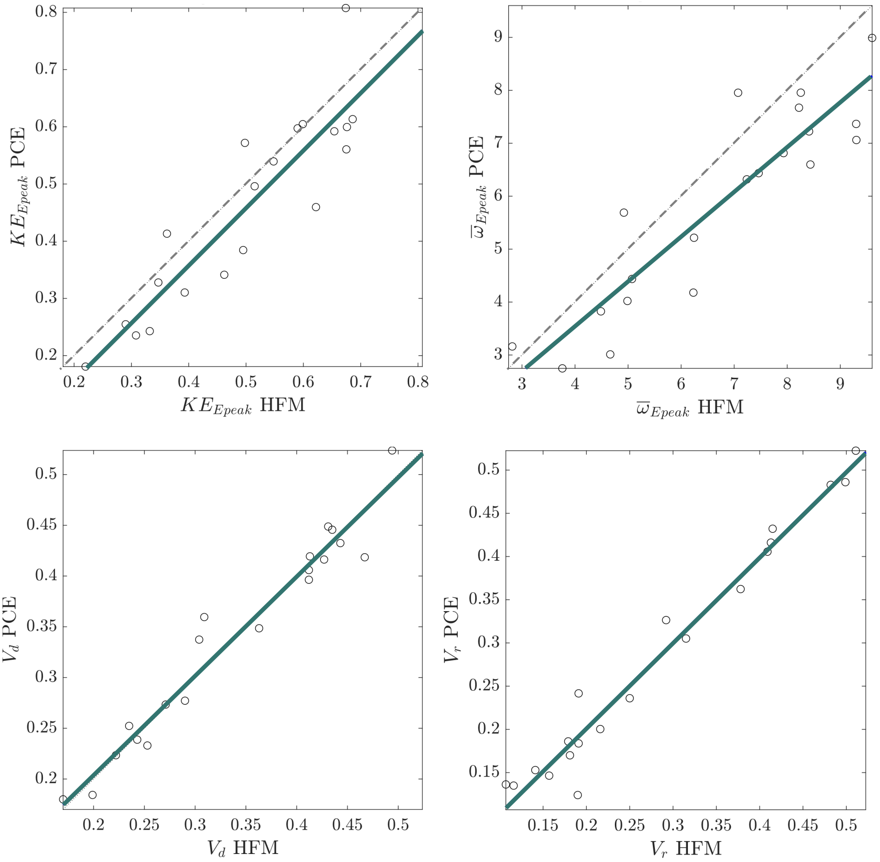

**Figure A.2:** Comparison between HFM and PCE predictions for the 4 quantities of interest ($KE_{Epeak}$, $\overline{\omega}_{Epeak}$, $V_d$, $V_r$) on a set of 20 real, patient-specific, cases. Solid green lines represent the linear regressions, while dashed grey lines are the bisectors.

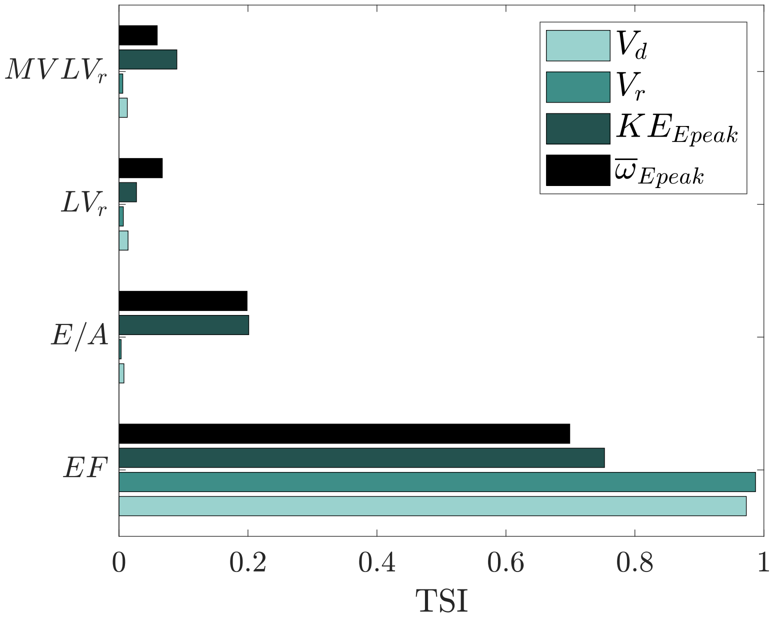

**Figure A.3:** Total sensitivity indices (TSI) of Sobol computed for the input parameters ($EF$, $E/A$, $LV_r$, and $MVLV_r$) with respect to each quantity of interest ($V_d$, $V_r$, $KE_{Epeak}$, $\overline{\omega}_{Epeak}$).

# Bibliography

J. R. Appleyard, I. M. Cheshire, and R. K. Pollard. Special techniques for fully implicit simulators. *Proc. European Symp. on Enhanced Oil Recovery, Bournemouth, UK*, pages 395–408, 1981.

M. J. Asher, B. F. W. Croke, A. J. Jakeman, and L. J. M. Peeters. A review of surrogate models and their application to groundwater modeling. *Water Resources Research*, 51(8):5957–5973, 2015. doi: 10.1002/2015wr016967.

S. Bagheri. Koopman-mode decomposition of the cylinder wake. *Journal of Fluid Mechanics*, 726:596–623, 2013. doi: 10.1017/jfm.2013.249.

G. D. Baldassarre, F. Martinez, Z. Kalantari, and A. Viglione. Drought and flood in the anthropocene: Feedback mechanisms in reservoir operation. *Earth System Dynamics*, 8(1):225–233, 2017. doi: 10.5194/esd-8-225-2017.

G. Bao, Y. Bazilevs, J.-H. Chung, P. Decuzzi, H. D. Espinosa, M. Ferrari, H. Gao, S. S. Hossain, T. J. R. Hughes, R. D. Kamm, W. K. Liu, A. Marsden, and B. Schrefler. USNCTAM perspectives on mechanics in medicine. *Journal of The Royal Society Interface*, 11(97):20140301, 2014. doi: 10.1098/rsif.2014.0301.

P. Benettin, N. B. Rodriguez, M. Sprenger, M. Kim, J. Klaus, C. J. Harman, Y. van der Velde, M. Hrachowitz, G. Botter, K. J. McGuire, J. W. Kirchner, A. Rinaldo, and J. J. McDonnell.

Transit time estimation in catchments: Recent developments and future directions. *Water Resources Research*, 58(11), 2022. doi: 10.1029/2022wr033096.

G. Botter, E. Bertuzzo, and A. Rinaldo. Transport in the hydrologic response: Travel time distributions, soil moisture dynamics, and the old water paradox. *Water Resources Research*, 46 (3), 2010. doi: 10.1029/2009wr008371.

B. W. Brunton, L. A. Johnson, J. G. Ojemann, and J. N. Kutz. Extracting spatial–temporal coherent patterns in large-scale neural recordings using dynamic mode decomposition. *Journal of Neuroscience Methods*, 258:1–15, 2016. doi: 10.1016/j.jneumeth.2015.10.010.

S. L. Brunton and J. N. Kutz. *Data-driven Science And Engineering*. Cambridge University Press, 2019. doi: 10.1017/9781108380690.

J. L. Callaham, S. L. Brunton, and J.-C. Loiseau. On the role of nonlinear correlations in reduced-order modelling. *Journal of Fluid Mechanics*, 938, 2022. doi: 10.1017/jfm.2021.994.

F. Cappelli, V. Costantini, and D. Consoli. The trap of climate change-induced “natural” disasters and inequality. *Global Environmental Change*, 70:102329, 2021. doi: 10.1016/j.gloenvcha.2021.102329.

S. L. Castle, B. F. Thomas, J. T. Reager, M. Rodell, S. C. Swenson, and J. S. Famiglietti. Groundwater depletion during drought threatens future water security of the colorado river basin. *Geophysical Research Letters*, 41(16):5904–5911, 2014. doi: 10.1002/2014gl061055.

T. B. Cetrulo, R. C. Marques, T. F. Malheiros, and N. M. Cetrulo. Monitoring inequality in water access: Challenges for the 2030 agenda for sustainable development. *Science of The Total Environment*, 727:138746, 2020. doi: 10.1016/j.scitotenv.2020.138746.

Z. Chen and D. Xiu. On generalized residual network for deep learning of unknown dynamical systems. *Journal of Computational Physics*, 438:110362, 2021. doi: 10.1016/j.jcp.2021.110362.

R. Chi-Durán and B. A. Buffett. Extracting spatial–temporal coherent patterns in geomagnetic secular variation using dynamic mode decomposition. *Geophysical Research Letters*, 50(5), 2023. doi: 10.1029/2022gl101288.

V. Ciriello and F. P. J. de Barros. Characterizing the influence of multiple uncertainties on predictions of contaminant discharge in groundwater within a Lagrangian stochastic formulation. *Water Resources Research*, 56(10), 2020. doi: 10.1029/2020wr027867.

V. Ciriello, V. D. Federico, M. Riva, F. Cadini, J. D. Sanctis, E. Zio, and A. Guadagnini. Polynomial chaos expansion for global sensitivity analysis applied to a model of radionuclide migration in a randomly heterogeneous aquifer. *Stochastic Environmental Research and Risk Assessment*, 27(4):945–954, 2012. doi: 10.1007/s00477-012-0616-7.

V. Ciriello, I. Lauriola, S. Bonvicini, V. Cozzani, V. Di Federico, and D. M. Tartakovsky. Impact of hydrogeological uncertainty on estimation of environmental risks posed by hydrocarbon transportation networks. *Water Resources Research*, 53 (11):8686–8697, 2017. doi: 10.1002/2017wr021368.

V. Ciriello, I. Lauriola, and D. M. Tartakovsky. Distribution-based global sensitivity analysis in hydrology. *Water Resources Research*, 55(11):8708–8720, 2019. doi: 10.1029/2019wr025844.

D. Collia, G. Libero, G. Pedrizzetti, and V. Ciriello. Surrogate models provide new insights on metrics based on blood flow for the assessment of left ventricular function. *Scientific Reports*, 12(1), 2022. doi: 10.1038/s41598-022-12560-3.

A. T. Corey. The interrelation between gas and oil relative permeabilities. *Producers monthly*, pages 38–41, 1954.

W. J. Cosgrove and D. P. Loucks. Water management: Current and future challenges and research directions. *Water Resources Research*, 51(6):4823–4839, 2015. doi: 10.1002/2014wr016869.

M. Dentz, H. Kinzelbach, S. Attinger, and W. Kinzelbach. Temporal behavior of a solute cloud in a heterogeneous porous medium: 1. point-like injection. *Water Resources Research*, 36(12):3591–3604, 2000. doi: 10.1029/2000wr900162.

M. Dentz, A. Cortis, H. Scher, and B. Berkowitz. Time behavior of solute transport in heterogeneous media: Transition from anomalous to normal transport. *Advances in Water Resources*, 27(2):155–173, 2004. doi: 10.1016/j.advwatres.2003.11.002.

M. Di Dato, F. P. J. de Barros, A. Fiori, and A. Bellin. Improving the efficiency of 3-d hydrogeological mixers: Dilution enhancement via coupled engineering-induced transient flows and spatial heterogeneity. *Water Resources Research*, 54(3):2095–2111, 2018. doi: 10.1002/2017WR022116.

P. D'Odorico, F. Laio, A. Porporato, L. Ridolfi, A. Rinaldo, and I. Rodriguez-Iturbe. Ecohydrology of terrestrial ecosystems. *BioScience*, 60(11):898–907, 2010. doi: 10.1525/bio.2010.60.11.6.

S. Dutta, M. W. Farthing, E. Perracchione, G. Savant, and M. Putti. A greedy non-intrusive reduced order model for shallow water equations. *Journal of Computational Physics*, 439: 110378, 2021. doi: 10.1016/j.jcp.2021.110378.

P. Döll, H. M. Schmied, C. Schuh, F. T. Portmann, and A. Eicker. Global-scale assessment of groundwater depletion and related groundwater abstractions: Combining hydrological modeling with information from well observations and GRACE satellites. *Water Resources Research*, 50(7):5698–5720, 2014. doi: 10.1002/2014wr015595.

N. B. Erichson, S. L. Brunton, and J. N. Kutz. Compressed dynamic mode decomposition for background modeling. *Journal of Real-Time Image Processing*, 16(5):1479–1492, 2016. doi: 10.1007/s11554-016-0655-2.

J. S. Famiglietti, M. Lo, S. L. Ho, J. Bethune, K. J. Anderson, T. H. Syed, S. C. Swenson, C. R. de Linage, and M. Rodell. Satellites measure recent rates of groundwater depletion in california's central valley. *Geophysical Research Letters*, 38(3), 2011. doi: 10.1029/2010gl046442.

S. Fankhauser, S. M. Smith, M. Allen, K. Axelsson, T. Hale, C. Hepburn, J. M. Kendall, R. Khosla, J. Lezaun, E. Mitchell-Larson, M. Obersteiner, L. Rajamani, R. Rickaby, N. Seddon, and T. Wetzer. The meaning of net zero and how to get it right. *Nature Climate Change*, 12(1):15–21, 2021. doi: 10.1038/s41558-021-01245-w.

S. Focaccia, G. Panini, P. Pedrazzoli, and V. Ciriello. A meta-modeling approach for hydrological forecasting under uncertainty: Application to groundwater nitrate response to climate change. *Journal of Hydrology*, 603:127173, 2021. doi: 10.1016/j.jhydrol.2021.127173.

D. J. Fulton and G. C. Hegerl. Testing methods of pattern extraction for climate data using synthetic modes. *Journal of Climate*, 34(18):7645–7660, 2021. doi: 10.1175/jcli-d-20-0871.1.

H.-M. Füssel. Adaptation planning for climate change: Concepts, assessment approaches, and key lessons. *Sustainability Science*, 2(2):265–275, 2007. doi: 10.1007/s11625-007-0032-y.

R. G. Ghanem and P. D. Spanos. *Stochastic Finite Elements: A Spectral Approach.* Springer New York, 1991. doi: 10.1007/978-1-4612-3094-6.

P. Gourbesville. Challenges for integrated water resources management. *Physics and Chemistry of the Earth, Parts A/B/C*, 33(5):284–289, 2008. doi: 10.1016/j.pce.2008.02.002.

Z. Guo, C. V. Henri, G. E. Fogg, Y. Zhang, and C. Zheng. Adaptive multirate mass transfer (ammt) model: A new approach to upscale regional-scale transport under transient flow conditions. *Water Resources Research*, 56(2):e2019WR026000, 2020. doi: 10.1029/2019WR026000.

S. Hallegatte. Strategies to adapt to an uncertain climate change. *Global Environmental Change*, 19(2):240–247, 2009. doi: 10.1016/j.gloenvcha.2008.12.003.

Q. He and A. M. Tartakovsky. Physics-informed neural network method for forward and backward advection-dispersion equations. *Water Resources Research*, 57, 2021. doi: 10.1029/2020wr029479.

J. D. Herman, J. D. Quinn, S. Steinschneider, M. Giuliani, and S. Fletcher. Climate adaptation as a control problem: Review and perspectives on dynamic water resources planning under uncertainty. *Water Resources Research*, 56(2), 2020. doi: 10.1029/2019wr025502.

R. Hilborn, C. J. Walters, and D. Ludwig. Sustainable exploitation of renewable resources. *Annual Review of Ecology and Systematics*, 26(1):45–67, 1995. doi: 10.1146/annurev.es.26.110195.000401.

J. Hogg, M. Fonoberova, and I. Mezić. Exponentially decaying modes and long-term prediction of sea ice concentration using koopman mode decomposition. *Scientific Reports*, 10(1), 2020. doi: 10.1038/s41598-020-73211-z.

G. Howard, R. Calow, A. Macdonald, and J. Bartram. Climate change and water and sanitation: Likely impacts and emerging trends for action. *Annual Review of Environment and Resources*, 41(1):253–276, 2016. doi: 10.1146/annurev-environ-110615-085856.

IPCC. *Climate Change 2014 – Impacts, Adaptation, And Vulnerability*. Cambridge University Press, 2014a. doi: 10.1017/cbo9781107415379.

IPCC. *Climate Change 2014 – Mitigation Of Climate Change*. Cambridge University Press, 2014b.

IPCC. *Climate Change 2022 – Impacts, Adaptation And Vulnerability*. Cambridge University Press, 2023a. doi: 10.1017/9781009325844.

IPCC. *Climate Change 2023 – Synthesis Report*. Intergovernmental Panel on Climate Change, 2023b. doi: 10.59327/ipcc/ar6-9789291691647.

G. Joodaki, J. Wahr, and S. Swenson. Estimating the human contribution to groundwater depletion in the middle east, from GRACE data, land surface models, and well observations. *Water Resources Research*, 50(3):2679–2692, 2014. doi: 10.1002/2013wr014633.

X. Kang, A. Kokkinaki, X. Shi, H. Yoon, J. Lee, P. K. Kitanidis, and J. Wu. Integration of deep learning-based inversion and upscaled mass-transfer model for dnapl mass-discharge estimation and uncertainty assessment. *Water Resources Research*, 58(10): e2022WR033277, 2022. doi: 10.1029/2022WR033277.

H. R. Karbasian and B. C. Vermeire. Application of physics-constrained data-driven reduced-order models to shape optimization. *Journal of Fluid Mechanics*, 934, 2022. doi: 10.1017/jfm.2021.1051.

B. O. Koopman. Hamiltonian systems and transformation in hilbert space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931. doi: 10.1073/pnas.17.5.315.

D. K. Kreamer. The past, present, and future of water conflict and international security. *Journal of Contemporary Water Research & Education*, 149(1):87–95, 2012. doi: 10.1111/j.1936-704x.2012.03130.x.

J. N. Kutz, S. L. Brunton, B. W. Brunton, and J. L. Proctor. *Dynamic Mode Decomposition*. Society for Industrial and Applied Mathematics, 2016. doi: 10.1137/1.9781611974508.

F. W. Landerer, F. M. Flechtner, H. Save, F. H. Webb, T. Bandikova, W. I. Bertiger, S. V. Bettadpur, S. H. Byun, C. Dahle, H. Dobslaw, E. Fahnestock, N. Harvey, Z. Kang, G. L. H. Kruizinga, B. D. Loomis, C. McCullough, M. Murböck, P. Nagel, M. Paik, N. Pie, S. Poole, D. Strekalov, M. E. Tamisiea, F. Wang, M. M. Watkins, H.-Y. Wen, D. N. Wiese, and D.-N. Yuan. Extending the global mass change data record: GRACE follow-on instrument and science data performance. *Geophysical Research Letters*, 47(12), 2020. doi: 10.1029/2020gl088306.

G. H. Leavesley. Modeling the effects of climate change on water resources - a review. *Climatic Change*, 28(1-2):159–177, 1994. doi: 10.1007/bf01094105.

M. J. Leblanc, P. Tregoning, G. Ramillien, S. O. Tweed, and A. Fakes. Basin-scale, integrated observations of the early 21st century multiyear drought in southeast australia. *Water Resources Research*, 45(4), 2009. doi: 10.1029/2008wr007333.

F. Li, J. Kusche, R. Rietbroek, Z. Wang, E. Forootan, K. Schulze, and C. Lück. Comparison of data-driven techniques to reconstruct (1992–2002) and predict (2017–2018) GRACE-like gridded total water storage changes using climate inputs. *Water Resources Research*, 56(5), 2020. doi: 10.1029/2019wr026551.

L. Lin, H. Yang, and X. Xu. Effects of water pollution on human health and disease heterogeneity: A review. *Frontiers in Environmental Science*, 10, 2022. doi: 10.3389/fenvs.2022.880246.

Y. Liu, J. Long, Q. Wu, B. Huang, and G. Wang. Data-driven modal decomposition of transient cavitating flow. *Physics of Fluids*, 33(11):113316, 2021. doi: 10.1063/5.0073266.

H. Lu and D. M. Tartakovsky. Lagrangian dynamic mode decomposition for construction of reduced-order models of advection-dominated phenomena. *Journal of Computational Physics*, 407: 109229, 2020. doi: 10.1016/j.jcp.2020.109229.

H. Lu and D. M. Tartakovsky. Extended dynamic mode decomposition for inhomogeneous problems. *Journal of Computational Physics*, 444:110550, 2021. doi: 10.1016/j.jcp.2021.110550.

H. Lu, D. Ermakova, H. M. Wainwright, L. Zheng, and D. M. Tartakovsky. Data-informed emulators for multi-physics simulations, 2020.

Y. Lu, S. Song, R. Wang, Z. Liu, J. Meng, A. J. Sweetman, A. Jenkins, R. C. Ferrier, H. Li, W. Luo, and T. Wang. Impacts of soil and water pollution on food safety and health risks in china. *Environment International*, 77:5–15, 2015. doi: 10.1016/j.envint.2014.12.010.

J. Mann and J. N. Kutz. Dynamic mode decomposition for financial trading strategies. *Quantitative Finance*, 16(11):1643–1655, 2016. doi: 10.1080/14697688.2016.1170194.

I. Mezić. Analysis of fluid flows via spectral properties of the Koopman operator. *Annual Review of Fluid Mechanics*, 45(1): 357–378, 2013. doi: 10.1146/annurev-fluid-011212-140652.

S. Mo, Y. Zhu, N. Zabaras, X. Shi, and J. Wu. Deep convolutional encoder-decoder networks for uncertainty quantification of dynamic multiphase flow in heterogeneous media. *Water Resources Research*, 55(1):703–728, 2019. ISSN 1944-7973. doi: 10.1029/2018wr023528.

C. Mora, T. McKenzie, I. M. Gaw, J. M. Dean, H. von Hammerstein, T. A. Knudson, R. O. Setter, C. Z. Smith, K. M. Webster, J. A. Patz, and E. C. Franklin. Over half of known human pathogenic diseases can be aggravated by climate change. *Nature Climate Change*, 12(9):869–875, 2022. doi: 10.1038/s41558-022-01426-1.

H. Mu, X. Li, Y. Wen, J. Huang, P. Du, W. Su, S. Miao, and M. Geng. A global record of annual terrestrial human footprint dataset from 2000 to 2018. *Scientific Data*, 9(1), 2022. doi: 10.1038/s41597-022-01284-8.

K. D. W. Nandalal and S. P. Simonovic. Resolving conflicts in water sharing: A systemic approach. *Water Resources Research*, 39(12), 2003. doi: 10.1029/2003wr002172.

NASA/JPL. Jpl grace and grace-fo mascon ocean, ice, and hydrology equivalent water height coastal resolution improvement (cri) filtered release 06.1 version 03, 2023.

C.-Y. Ohh and G. R. Spedding. Wake identification of stratified flows using dynamic mode decomposition. *Physical Review Fluids*, 7(2):024801, 2022. doi: 10.1103/physrevfluids.7.024801.

S. Oladyshkin, F. P. J. de Barros, and W. Nowak. Global sensitivity analysis: A flexible and efficient framework with an example from stochastic hydrogeology. *Advances in Water Resources*, 37: 10–22, 2012. doi: 10.1016/j.advwatres.2011.11.001.

B. Pan, K. Hsu, A. AghaKouchak, and S. Sorooshian. Improving precipitation estimation using convolutional neural network. *Water Resources Research*, 55(3):2301–2321, 2019. doi: 10.1029/2018wr024090.

S. Pan, N. Arnold-Medabalimi, and K. Duraisamy. Sparsity-promoting algorithms for the discovery of informative Koopman-invariant subspaces. *Journal of Fluid Mechanics*, 917, 2021. doi: 10.1017/jfm.2021.271.

J. A. Patz, D. Campbell-Lendrum, T. Holloway, and J. A. Foley. Impact of regional climate change on human health. *Nature*, 438(7066):310–317, 2005. doi: 10.1038/nature04188.

G. Pedrizzetti, G. L. Canna, O. Alfieri, and G. Tonti. The vortex—an early predictor of cardiovascular outcome? *Nature Reviews Cardiology*, 11(9):545–553, 2014. doi: 10.1038/nrcardio.2014.75.

L. Pegolotti, M. R. Pfaller, A. L. Marsden, and S. Deparis. Model order reduction of flow based on a modular geometrical approximation of blood vessels. *Computer Methods in Applied Mechanics and Engineering*, 380:113762, 2021. doi: 10.1016/j.cma.2021.113762.

L. Pegolotti, M. R. Pfaller, N. L. Rubio, K. Ding, R. B. Brufau, E. Darve, and A. L. Marsden. Learning reduced-order models for cardiovascular simulations with graph neural networks, 2023.

T. D. Phan, E. Bertone, and R. A. Stewart. Critical review of system dynamics modelling applications for water resources planning and management. *Cleaner Environmental Systems*, 2: 100031, 2021. doi: 10.1016/j.cesys.2021.100031.

J. L. Proctor and P. A. Eckhoff. Discovering dynamic patterns from infectious disease data using dynamic mode decomposition. *International Health*, 7(2):139–145, 2015. doi: 10.1093/inthealth/ihv009.

T. Qin, K. Wu, and D. Xiu. Data driven governing equations approximation using deep neural networks. *Journal of Computational Physics*, 395:620–635, 2019. doi: 10.1016/j.jcp.2019.06.042.

S. Quicken, W. P. Donders, E. M. J. van Disseldorp, K. Gashi, B. M. E. Mees, F. N. van de Vosse, R. G. P. Lopata, T. Delhaas, and W. Huberts. Application of an adaptive polynomial chaos expansion on computationally expensive three-dimensional cardiovascular models for uncertainty quantification and sensitivity analysis. *Journal of Biomechanical Engineering*, 138(12), 2016. doi: 10.1115/1.4034709.

A. Rinaldo and I. Rodriguez-Iturbe. Ecohydrology 2.0. *Rendiconti Lincei. Scienze Fisiche e Naturali*, 33(2):245–270, 2022. doi: 10.1007/s12210-022-01071-y.

M. Riva, A. Guadagnini, and A. Dell'Oca. Probabilistic assessment of seawater intrusion under multiple sources of uncertainty. *Advances in Water Resources*, 75:93–104, 2015. doi: 10.1016/j.advwatres.2014.11.002.

M. Rodell and J. T. Reager. Water cycle science enabled by the GRACE and GRACE-FO satellite missions. *Nature Water*, 1 (1):47–59, 2023. doi: 10.1038/s44221-022-00005-0.

M. Rodell, J. S. Famiglietti, D. N. Wiese, J. T. Reager, H. K. Beaudoing, F. W. Landerer, and M.-H. Lo. Emerging trends in global freshwater availability. *Nature*, 557(7707):651–659, 2018. doi: 10.1038/s41586-018-0123-1.

C. W. Rowley, I. Mezić, S. Bagheri, P. Schlatter, and D. S. Henningson. Spectral analysis of nonlinear flows. *Journal of Fluid Mechanics*, 641:115–127, 2009. doi: 10.1017/s0022112009992059.

S. Sankaran and A. L. Marsden. A stochastic collocation method for uncertainty quantification and propagation in cardiovascular simulations. *Journal of Biomechanical Engineering*, 133(3), 2011. doi: 10.1115/1.4003259.

D. E. Schiavazzi, G. Arbia, C. Baker, A. M. Hlavacek, T. Y. Hsia, A. L. Marsden, I. E. Vignon-Clementel, and T. M. of Congenital Hearts Alliance (MOCHA) Investigators. Uncertainty quantification in virtual surgery hemodynamics predictions for single ventricle palliation. *International Journal for Numerical Methods in Biomedical Engineering*, 32(3), 2015. doi: 10.1002/cnm.2737.

P. J. Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656:5–28, 2010. doi: 10.1017/s0022112010001217.

P. J. Schmid. Dynamic mode decomposition and its variants. *Annual Review of Fluid Mechanics*, 54(1):225–254, 2022. doi: 10.1146/annurev-fluid-030121-015835.

P. J. Schmid, L. Li, M. P. Juniper, and O. Pust. Applications of the dynamic mode decomposition. *Theoretical and Computational Fluid Dynamics*, 25(1-4):249–259, 2010. doi: 10.1007/s00162-010-0203-9.

J. Sheffield, E. F. Wood, M. Pan, H. Beck, G. Coccia, A. Serrat-Capdevila, and K. Verbist. Satellite remote sensing for water resources management: Potential for supporting sustainable development in data-poor regions. *Water Resources Research*, 54 (12):9724–9758, 2018. doi: 10.1029/2017wr022437.

G. B. Simpson and G. P. W. Jewitt. The development of the water-energy-food nexus as a framework for achieving resource security: A review. *Frontiers in Environmental Science*, 7, 2019. doi: 10.3389/fenvs.2019.00008.

D. H. Song and D. M. Tartakovsky. Transfer learning on multi-fidelity data. *Journal of Machine Learning for Modeling and Computing*, 2021. doi: 10.48550/ARXIV.2105.00856.

A. Steer. Resource depletion, climate change, and economic growth1. In *Towards a Better Global Economy*, pages 381–426. Oxford University Press, 2014. doi: 10.1093/acprof:oso/9780198723455.003.0006.

B. Sudret. Global sensitivity analysis using polynomial chaos expansions. *Reliability Engineering & System Safety*, 93(7):964–979, 2008. doi: 10.1016/j.ress.2007.04.002.

Z. Sun, D. Long, W. Yang, X. Li, and Y. Pan. Reconstruction of GRACE data on changes in total water storage over the global land surface and 60 basins. *Water Resources Research*, 56(4), 2020. doi: 10.1029/2019wr026250.

D. L. Swain, D. Singh, D. Touma, and N. S. Diffenbaugh. Attributing extreme events to climate change: A new frontier in a warming world. *One Earth*, 2(6):522–527, 2020. doi: 10.1016/j.oneear.2020.05.011.

B. D. Tapley, S. Bettadpur, J. C. Ries, P. F. Thompson, and M. M. Watkins. GRACE measurements of mass variability in the earth system. *Science*, 305(5683):503–505, 2004. doi: 10.1126/science.1099192.

A. M. Tartakovsky, C. O. Marrero, P. Perdikaris, G. D. Tartakovsky, and D. Barajas-Solano. Physics-informed deep neural networks for learning parameters and constitutive relationships in subsurface flow problems. *Water Resources Research*, 56(5), 2020. doi: 10.1029/2019wr026731.

D. M. Tartakovsky and S. P. Neuman. Transient flow in bounded randomly heterogeneous domains: 2. localization of conditional mean equations and temporal nonlocality effects. *Water Resources Research*, 34(1):13–20, 1998. doi: 10.1029/97wr02117.

S. Taverniers, S. B. M. Bosma, and D. M. Tartakovsky. Accelerated multilevel monte carlo with kernel-based smoothing and latinized stratification. *Water Resources Research*, 56(9), 2020. doi: 10.1029/2019wr026984.

B. F. Thomas and J. S. Famiglietti. Identifying climate-induced groundwater depletion in GRACE observations. *Scientific Reports*, 9(1), 2019. doi: 10.1038/s41598-019-40155-y.

J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz. On dynamic mode decomposition: Theory and applications. *Journal of Computational Dynamics*, 1(2):391–421, 2014. doi: 10.3934/jcd.2014.1.391.

A. Updegrove, N. M. Wilson, J. Merkow, H. Lan, A. L. Marsden, and S. C. Shadden. SimVascular: An open source pipeline for cardiovascular simulation. *Annals of Biomedical Engineering*, 45(3):525–541, 2016. doi: 10.1007/s10439-016-1762-8.

A. Vespignani. Modelling dynamical processes in complex socio-technical systems. *Nature Physics*, 8(1):32–39, 2011. doi: 10.1038/nphys2160.

F. Wang, J. D. Harindintwali, Z. Yuan, M. Wang, F. Wang, S. Li, Z. Yin, L. Huang, Y. Fu, L. Li, S. X. Chang, L. Zhang, J. Rinklebe, Z. Yuan, Q. Zhu, L. Xiang, D. C. Tsang, L. Xu, X. Jiang, J. Liu, N. Wei, M. Kästner, Y. Zou, Y. S. Ok, J. Shen, D. Peng, W. Zhang, D. Barceló, Y. Zhou, Z. Bai, B. Li, B. Zhang, K. Wei, H. Cao, Z. Tan, L. bin Zhao, X. He, J. Zheng, N. Bolan, X. Liu, C. Huang, S. Dietmann, M. Luo, N. Sun, J. Gong, Y. Gong, F. Brahushi, T. Zhang, C. Xiao, X. Li, W. Chen, N. Jiao, J. Lehmann, Y.-G. Zhu, H. Jin, A. Schäffer, J. M. Tiedje, and J. M. Chen. Technologies and perspectives for achieving carbon neutrality. *The Innovation*, 2(4):100180, 2021. doi: 10.1016/j.xinn.2021.100180.

M. M. Watkins, D. N. Wiese, D.-N. Yuan, C. Boening, and F. W. Landerer. Improved methods for observing earth's time variable mass distribution with GRACE using spherical cap mascons. *Journal of Geophysical Research: Solid Earth*, 120(4): 2648–2671, 2015. doi: 10.1002/2014jb011547.

M. D. Webster, M. A. Tatang, and G. J. McRae. Application of the probabilistic collocation method for an uncertainty analysis of a simple ocean model. In *Tech. Rep. MIT joint program on the science and policy of global change reports series No. 4.*, 1996.

N. Wiener. The homogeneous chaos. *American Journal of Mathematics*, 60(4):897–936, 1938. ISSN 00029327, 10806377.

D. N. Wiese, F. W. Landerer, and M. M. Watkins. Quantifying and reducing leakage errors in the JPL RL05m GRACE mascon solution. *Water Resources Research*, 52(9):7490–7502, 2016. doi: 10.1002/2016wr019344.

N. M. Wilson, A. K. Ortiz, and A. B. Johnson. The vascular model repository: A public resource of medical imaging data and blood flow simulation results. *Journal of Medical Devices*, 7(4), 2013. doi: 10.1115/1.4025983.

WSH. Un-water global analysis and assessment of sanitation and drinking-water (glaas) 2014 report: Investing in water and sanitation: Increasing access, reducing inequalities. Technical report, World Health Organization, UN-Water, 2014.

D. Xiu and G. E. Karniadakis. The wiener–askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, 24(2):619–644, 2002. doi: 10.1137/s1064827501387826.

C. Zhan, Z. Dai, M. R. Soltanian, and F. P. J. de Barros. Data-worth analysis for heterogeneous subsurface structure identification with a stochastic deep learning framework. *Water Resources Research*, 58(11):e2022WR033241, 2022. doi: 10.1029/2022WR033241.